Lab 4: 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.

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

$$\log(\text{FEV}) = \beta_0 + \beta_1 \text{ pack years} + \beta_2 \text{ age} + \varepsilon \quad [1]$$

$$\text{FEV} = \beta_0 + s_3(\text{pack years}) \times s_3(\text{age}) + \varepsilon \quad [2]$$

Exercise 1: State the number of degrees of freedom used by each model. Let's briefly discuss some relative advantages and disadvantages of each model.

Exercise 2: Consider the following four subgroups:

Group 1: Individuals of age 75 years and a smoking history of 20 pack years.
Group 2: Individuals of age 75 years and a smoking history of 100 pack years.
Group 3: Individuals of age 95 years and a smoking history of 20 pack years.
Group 4: Individuals of age 95 years and a smoking history of 100 pack years.

Construct a scatter plot of age against pack years; for a given model, how would you expect the width of these four prediction intervals to compare? How might you expect this hypothesized discrepancy to vary between the two models?

Exercise 3: Fill in the table below, cross-tabulating the prediction intervals for each of the four groups across the four models. We will discuss the results as a group.

	Group 1	Group 2	Group 3	Group 4
Model 1				
Model 2				

Code for Exercise 3:

```
*** CODE FOR REGRESSION MODEL #1 ***

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

* 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 == .
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 = 75 if ptid == _N - 3
replace age = 75 if ptid == _N - 2
replace age = 95 if ptid == _N - 1
replace age = 95 if ptid == _N
replace packyrs = 20 if ptid == _N - 3
replace packyrs = 100 if ptid == _N - 2
replace packyrs = 20 if ptid == _N - 1
replace packyrs = 100 if ptid == _N

* Regression Model [1]

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

* Estimate parameters of regression Model [2]
regress logfev age packyrs if flag == 0

* Generate predictions and corresponding standard deviations
predict prm if flag == 1
predict prsd if flag == 1, stdf

* Mind the degrees of freedom (724 complete observations minus 3 parameters)
local tquant invt(724 - 3, 0.975)

* Generate prediction interval for Group 1 (Model [1])
quietly: summarize prm if ptid == _N - 3
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N - 3
local fevsd1 = r(mean)
display "GROUP 1: ["exp(`fevpr1' - `tquant') * `fevsd1' ", " exp(`fevpr1' + `tquant' * `fevsd1') "]"

* Generate prediction interval for Group 2 (Model [1])
quietly: summarize prm if ptid == _N - 2
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N - 2
local fevsd1 = r(mean)
display "GROUP 2: ["exp(`fevpr1' - `tquant' * `fevsd1') ", " exp(`fevpr1' + `tquant' * `fevsd1') "]"

* Generate prediction interval for Group 3 (Model [1])
quietly: summarize prm if ptid == _N - 1
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N - 1
local fevsd1 = r(mean)
display "GROUP 3: [" exp(`fevpr1' - `tquant' * `fevsd1') ", " exp(`fevpr1' + `tquant' * `fevsd1') "]"
* Generate prediction interval for Group 4 (Model [1])
quietly: summarize prm if ptid == _N
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N
local fevsd1 = r(mean)
display "GROUP 4: ["exp(`fevpr1' - `tquant' * `fevsd1') ", " exp(`fevpr1' + `tquant' * `fevsd1') "]"
```

```
*** CODE FOR REGRESSION MODEL #2 ***

import delimited "…mri.csv", clear

* 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 == .
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 = 75 if ptid == _N - 3
replace age = 75 if ptid == _N - 2
replace age = 95 if ptid == _N - 1
replace age = 95 if ptid == _N
replace packyrs = 20 if ptid == _N - 3
replace packyrs = 100 if ptid == _N - 2
replace packyrs = 20 if ptid == _N - 1
replace packyrs = 100 if ptid == _N

* Regression Model [2]

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

* Estimate parameters of regression Model [2]
regress fev aspl1-aspl2 pspl1-pspl2 c.aspl1#c.pspl1 c.aspl1#c.pspl2 c.aspl2#c.pspl1 c.aspl2#c.pspl2 if
flag == 0

* Generate predictions and corresponding standard deviations
predict prm if flag == 1
predict prsd if flag == 1, stdf

* Mind the degrees of freedom (724 complete observations minus 9 parameters)
local tquant invt(724 - 9, 0.975)

* Generate prediction interval for Group 1 (Model [2])
quietly: summarize prm if ptid == _N - 3
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N - 3
local fevsd1 = r(mean)
display "GROUP 1: ["`fevpr1' - `tquant' * `fevsd1' ", " `fevpr1' + `tquant' * `fevsd1' "]"

* Generate prediction interval for Group 2 (Model [2])
quietly: summarize prm if ptid == _N - 2
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N - 2
local fevsd1 = r(mean)
display "GROUP 2: ["`fevpr1' - `tquant' * `fevsd1' ", " `fevpr1' + `tquant' * `fevsd1' "]"

* Generate prediction interval for Group 3 (Model [2])
quietly: summarize prm if ptid == _N - 1
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N - 1
local fevsd1 = r(mean)
display "GROUP 3: [" `fevpr1' - `tquant' * `fevsd1' ", " `fevpr1' + `tquant' * `fevsd1' "]"

* Generate prediction interval for Group 4 (Model [2])
quietly: summarize prm if ptid == _N
local fevpr1 = r(mean)
quietly: summarize prsd if ptid == _N
local fevsd1 = r(mean)
display "GROUP 4: ["`fevpr1' - `tquant' * `fevsd1' ", " `fevpr1' + `tquant' * `fevsd1' "]"
```