MRC Clinical Trials Unit

Handling interactions in Stata, especially with continuous predictors

Patrick Royston & Willi Sauerbrei

UK Stata Users' meeting, London, 13-14 September 2012

Interactions – general concepts

- General idea of a (two-way) interaction in multiple regression is **effect modification**:
  - $\eta(x_1,x_2) = f_1(x_1) + f_2(x_2) + f_3(x_1,x_2)$
- Often, $\eta(x_1,x_2) = E(Y \mid x_1,x_2)$, with obvious extension to GLM, Cox regression, etc.
- Simplest case: $\eta(x_1,x_2)$ is linear in the x's and $f_3(x_1,x_2)$ is the product of the x's:
  - $\eta(x_1,x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
- Can extend to non-linear functions of x_1 & x_2

The simplest type of interaction: Binary x binary

- E.g. in the MRC RE01 trial in kidney cancer
- 12 month % survival since randomisation
- Substantial treatment effect in patients with low white cell count
- Little or no treatment effect in those with high white cell count
- But really, white cell count is a continuous variable …

Treatment group	White cell count low (<=10)	White cell count high (>10)
MPA	34% (se 4)	24% (se 4)
Interferon	49% (se 4)	21% (se 7)

Overview

- Fitting linear interaction models in Stata
- General case: analyzing interactions between continuous covariates in observational studies
  - Focus on continuous covariates
  - Maximize power
  - People may not know how to handle them
- Special case: analyzing interactions between treatment and continuous covariates in randomized controlled trials

Fitting models with linear x linear interactions in Stata

Binary x continuous interactions

- Use `c.` prefix to indicate continuous variable
- Use the ## operator

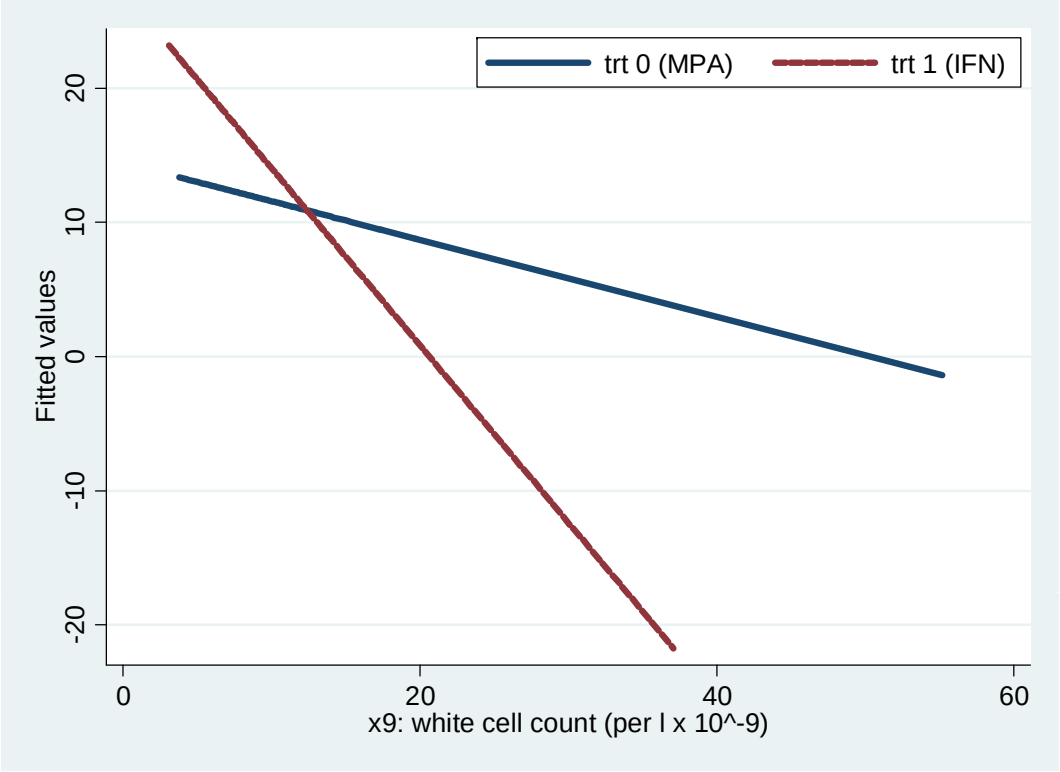
```
. regress _t trt##c.wcc

      Source |       SS       df       MS              Number of obs =     347
-------------+------------------------------           F(  3,   343) =    7.44
       Model |  5678.62935     3  1892.87645            Prob > F      =  0.0001
    Residual |  87228.7534   343  254.311234            R-squared     =  0.0611
-------------+------------------------------           Adj R-squared =  0.0529
       Total |  92907.3828   346  268.518447            Root MSE      =  15.947

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.trt |   12.81405   4.124167     3.11   0.002     4.702208    20.92589
         wcc |  -.2867831   .2741174    -1.05   0.296    -.8259457    .2523796
             |
   trt#c.wcc |
           1 |  -1.034239   .4327233    -2.39   0.017    -1.885365   -.1831142
             |
       _cons |   14.45292   2.712383     5.33   0.000     9.117919    19.78791
------------------------------------------------------------------------------
```

Binary x continuous interactions (cont.)

- The main effect of **`wcc`** is the slope in group 0
- The interaction parameter is the difference between the slopes in groups 1 & 0
- Test of **`trt#c.wcc`** provides the interaction parameter and test
- Results are nicely presented graphically
  - Predict linear predictor **`xb`**
  - Plot **`xb`** by levels of the factor variable
  - Also, 'treatment effect plot' (*coming later*)

Plotting a binary x continuous interaction

```
. regress _t trt##c.wcc
. predict fit
. twoway (line fit wcc if trt==0, sort) (line fit wcc
   if trt==1, sort lp(-)), legend(lab(1 "trt 0 (MPA)")
   lab(2 "trt 1 (IFN)") ring(0) pos(1))
```



Continuous x continuous interaction

- Just use **`c.`** prefix on each variable

```
. regress _t c.age##c.t_mt

      Source |       SS       df       MS              Number of obs =     347
-------------+------------------------------           F(  3,   343) =   10.35
       Model |  7714.26052     3  2571.42017            Prob > F      =  0.0000
    Residual |  85193.1223   343   248.37645            R-squared     =  0.0830
-------------+------------------------------           Adj R-squared =  0.0750
       Total |  92907.3828   346  268.518447            Root MSE      =   15.76

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0719063   .0876542     0.82   0.413    -.1005011    .2443137
        t_mt |   .0659781   .0128802     5.12   0.000      .040644    .0913122
             |
c.age#c.t_mt |  -.0008783   .0001861    -4.72   0.000    -.0012443   -.0005124
             |
       _cons |   8.055213   5.256114     1.53   0.126     -2.28306    18.39349
------------------------------------------------------------------------------
```

Continuous x continuous interaction

- Results are best explored graphically
- Consider in more detail next

Continuous x continuous interactions

Motivation

- Many people only consider linear by linear interactions
- Not sensible if main effect of either variable is non-linear
- Mismodelling the main effect may introduce spurious interactions
  - E.g. false assumption of linearity can create a spurious linear x linear interaction
- Or, people categorise the continuous variables
  - Many problems, including loss of power

MFPIgen

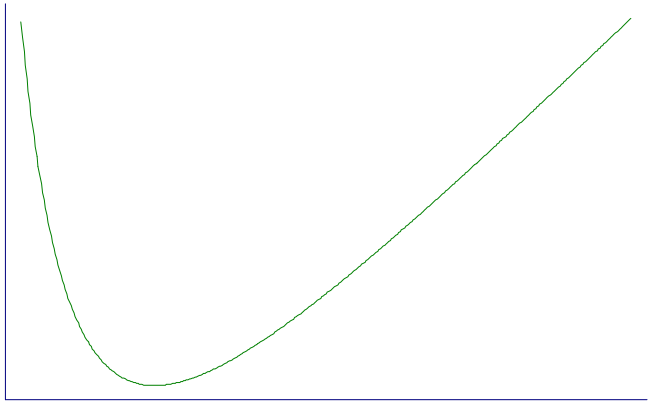
- MFP = multivariable fractional polynomials
- I = interaction
- gen = general
- Fractional polynomials (FPs) can be used to model relationships that may be non-linear
- In Stata, FPs are implemented through the standard **`fracpoly`** and **`mfp`** commands
- MFPIgen is implemented through a user-written command, **`mfpigen`**

Fractional polynomial models

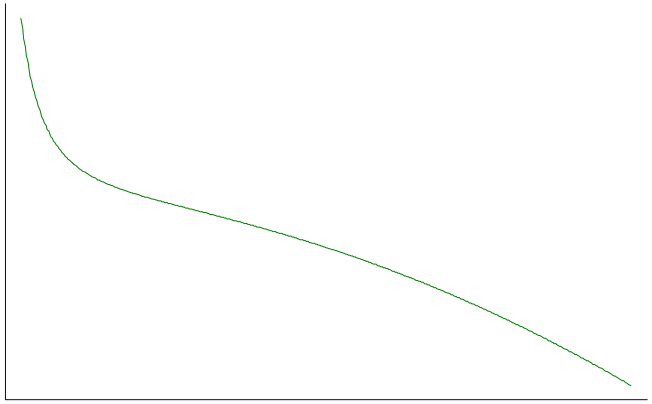
- Fractional polynomials are an extension of ordinary polynomials
- Degree 1: FP1(x) = $\beta_0+\beta_1 x^p$
- Degree 2: FP2(x) = $\beta_0+\beta_1 x^p+\beta_1 x^q$
- Powers p, q are taken from a special set S = {–2, – 1, – 0.5, 0, 0.5, 1, 2, 3}
- 8 FP1, 36 FP2 models
- Flexibility - many function shapes are available

Examples of FP2 curves - varying powers

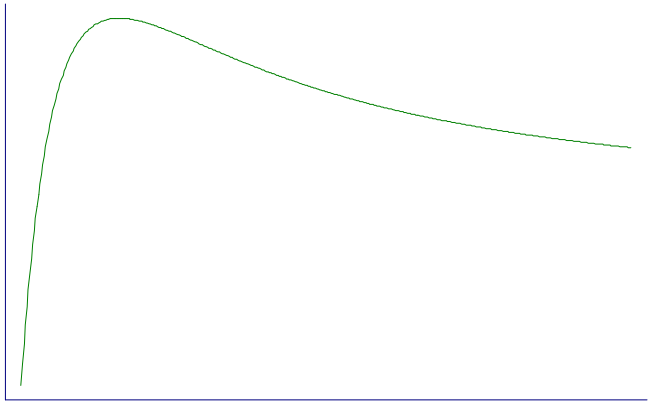
(-2, 1)



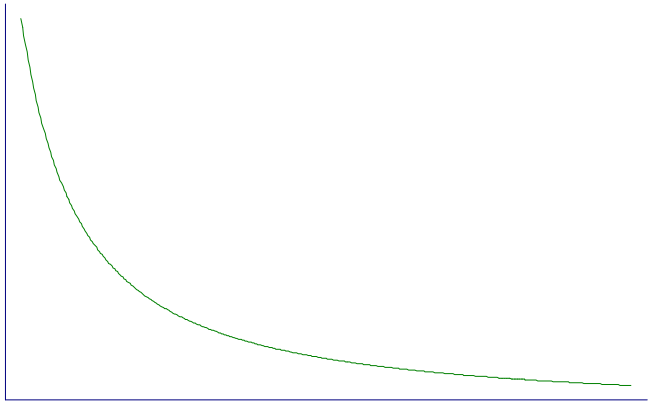
(-2, 2)



(-2, -2)



(-2, -1)



Several predictors - MFP

- With many continuous predictors, selection of best FP for each becomes more difficult $\rightarrow$
- The MFP algorithm is a standardized approach to variable and function selection
- The MFP algorithm combines backward elimination with a systematic FP function selection procedure
- Allows continuous, categorical and binary predictors

The MFPIgen approach in principle

- MFPIgen aims to identify non-linear main effects and their two-way interactions
- Suppose x_1 are x_2 continuous covariates
- Apply MFP to x_1 and x_2
  - Selects FP functions $FP_1(x_1)$ and $FP_2(x_2)$
  - (Linear functions could be selected)
- Add interaction term $FP_1(x_1) \times FP_2(x_2)$ to the chosen model
- Apply likelihood ratio test of interaction
- (Can include confounders z in the model)

Example: Whitehall 1

- Prospective cohort study of 17,260 Civil Servants in London
- Studied various standard risk factors for common causes of death
- Also studied social factors, particularly job grade
- We consider 10-year all-cause mortality as the outcome
- Logistic regression analysis

Example: Whitehall 1 (2)

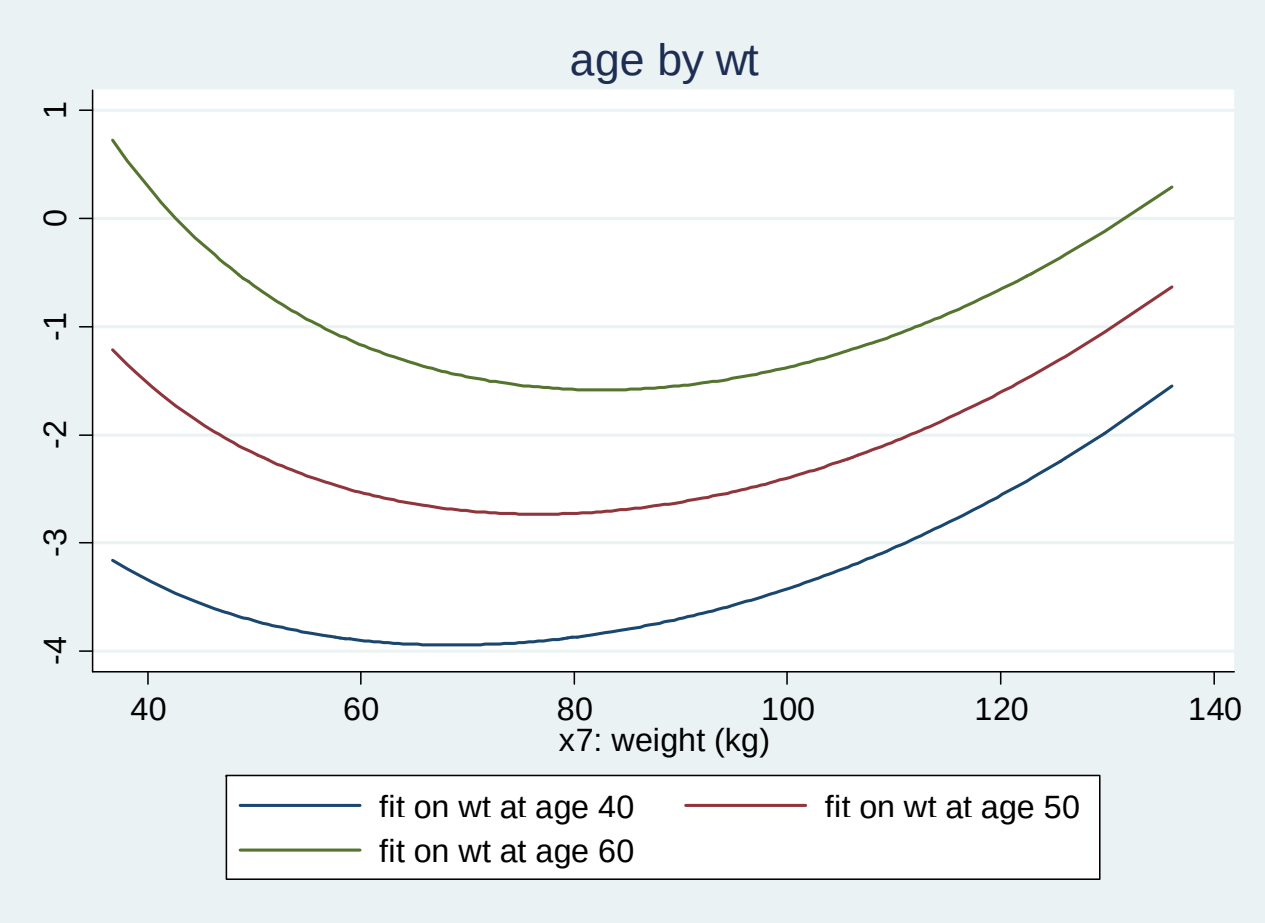
- Consider weight and age

```
. mfpigen: logit all10 age wt

MFPIGEN - interaction analysis for dependent variable all10
-------------------------------------------------------------------------------
variable 1   function 1   variable 2   function 2  dev. diff.  d.f.    P   Sel
-------------------------------------------------------------------------------
age          Linear        wt           FP2(-1 3)      5.2686     2   0.0718  0
-------------------------------------------------------------------------------
Sel = number of variables selected in MFP adjustment model
```

- Age function is linear, weight is FP2(-1, 3)
- No strong interaction (P = 0.07)

Plotting the interaction model

```
. mfpigen, fplot(40 50 60): logit all10 age wt
```



Mis-specifying the main effects function(s)

- Assume age and weight are linear
- The **`dfdefault(1)`** option imposes linearity

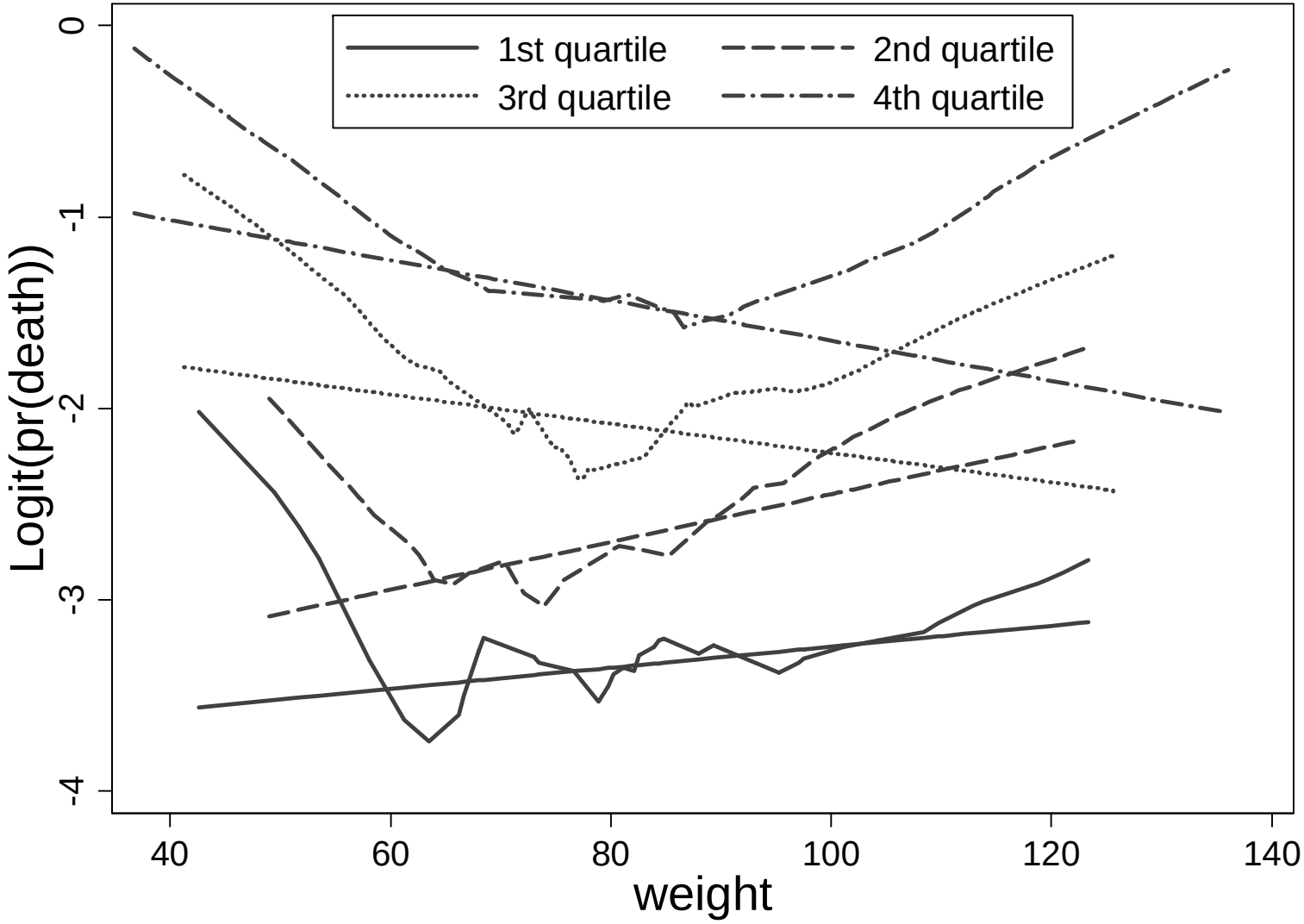
```
. mfpigen, dfdefault(1): logit all10 age wt

MFPIGEN - interaction analysis for dependent variable all10
-------------------------------------------------------------------------------
variable 1   function 1   variable 2   function 2  dev. diff.  d.f.     P   Sel
-------------------------------------------------------------------------------
age          Linear       wt           Linear         8.7375     1   0.0031  0
-------------------------------------------------------------------------------
Sel = number of variables selected in MFP adjustment model
```

- There appears to be a highly significant interaction (P = 0.003)

Checking the interaction model

- Linear age x weight interaction seems important
- Check if it's real, or the result of mismodelling
- Categorize age into (equal sized) groups
  - for example, 4 groups
- Compute running line smooth of the binary outcome on weight in each age group, transform to logits
- Plot results for each group
- Compare with the functions predicted by the interaction model

Whitehall 1: Check of age x weight linear interaction



Interpreting the plot

- Running line smooths are roughly parallel across age groups $\Rightarrow$ no (strong) interactions
- Erroneously assuming that the effect of weight is linear $\Rightarrow$ estimated slopes of weight in age-groups indicate strong interaction between age and weight
- We should have been more careful when modelling the main effect of weight

The MFPIgen approach in practice

- Consider a pair of covariates of interest
- **`mfpigen`** uses MFP to select a suitable function (FP/linear) simultaneously for each covariate
- **`mfpigen`** tests interaction between the 2 functions
  - use a low significance level, e.g. 1%
- Present the interaction model graphically
- Check the model graphically for artefacts

- **`mfpigen`** can use MFP to adjust for other covariates (confounders)
- **`mfpigen`** can analyze all pairs of covars in one run
- Can apply forward selection of interactions

Whitehall 1: 7 variables, any interactions?

```
. mfpigen, select(0.05): logit all10 cigs
sysbp age ht wt chol i.jobgrade
```

```
MFPIGEN - interaction analysis for dependent variable all10
-------------------------------------------------------------------------------
variable 1   function 1   variable 2   function 2  dev. diff.  d.f.     P    Sel
-------------------------------------------------------------------------------
cigs         FP1(.5)      sysbp        FP2(-2 -2)     0.7961     2   0.6716  5
             FP1(.5)      age          Linear         0.0028     1   0.9576  5
             FP1(.5)      ht           Linear         2.1029     1   0.1470  5
             FP1(.5)      wt           FP2(-2 3)      0.1560     2   0.9249  5
             FP1(.5)      chol         Linear         1.7712     1   0.1832  5
             FP1(.5)      i.jobgrade   Factor         4.3061     3   0.2303  5

sysbp        FP2(-2 -2)   age          Linear         3.1169     2   0.2105  5
```

(remaining output omitted)

What **`mfpigen`** is doing (Whitehall example)

- See the Stata log just given
- The **`select(0.05)`** option tests confounders for inclusion in each interaction model at the 5% significance level
- The **`Sel`** column in the output shows how many variables are actually included in each confounder model

Results: P-values for interactions

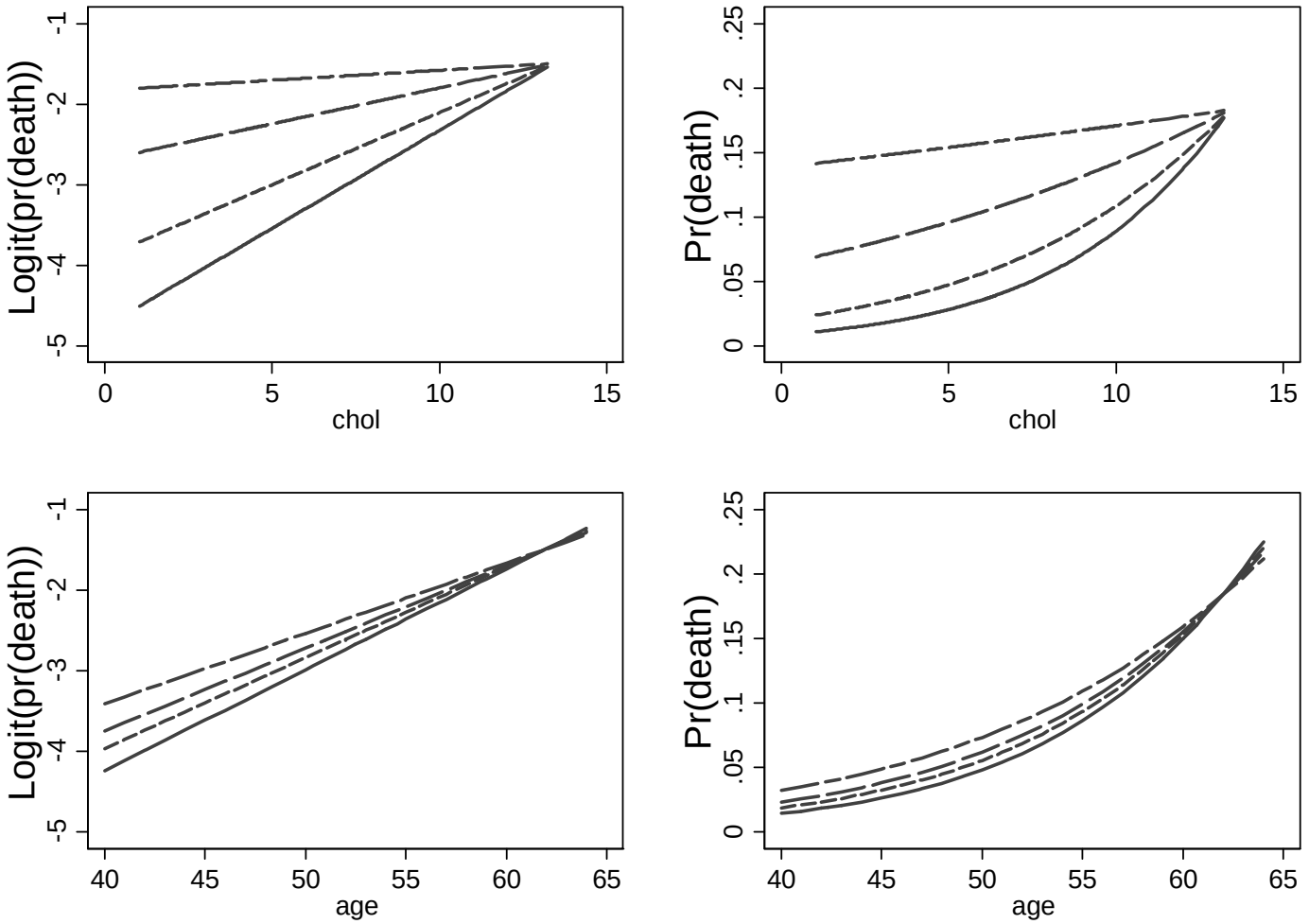
Variable	cigs*	sysbp*	age	height	weight*	chol
cigs*	–					
sysbp*	0.7	–				
age	0.9	0.2	–			
height	0.1	0.5	1.0	–		
weight*	0.9	0.5	0.1	0.4	–	
chol	0.2	0.07	0.001	0.8	0.2	–
grade	0.2	0.2	0.2	0.2	0.04	0.4

*FP transformations were selected; otherwise, linear

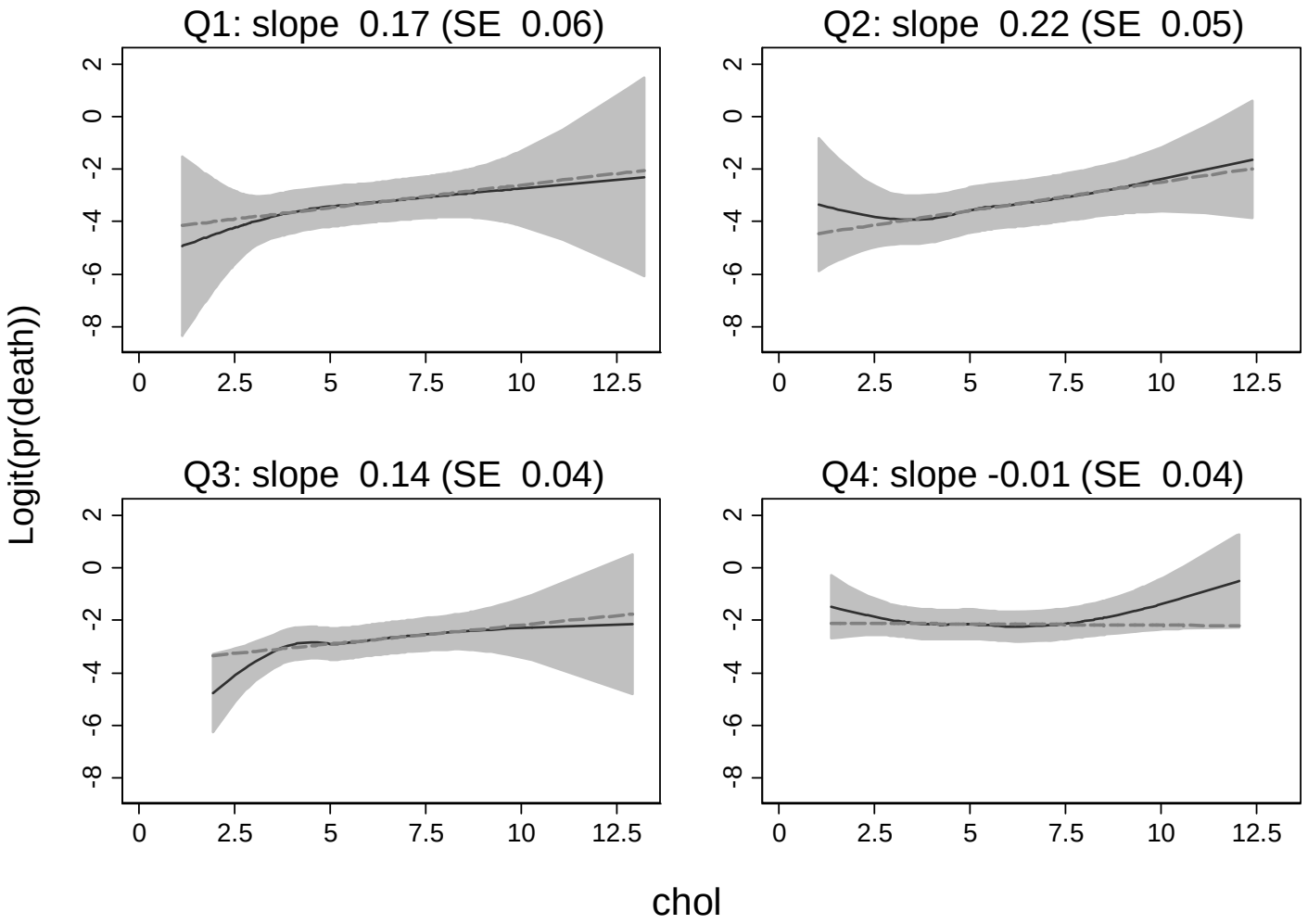
Graphical presentation of age x chol interaction

```
. fracgen cigs .5, center(mean)
. fracgen sysbp -2 -2, center(mean)
. fracgen wt -2 3, center(mean)

. mfpigen, linadj(cigs_1 sysbp_1 sysbp_2
> wt_1 wt_2 ht i.jobgrade) df(1)
> fplot(%10 35 65 90): logit all10 age chol
```

Graphical presentation of **age** x **chol** intn.



Checking the `chol` x `age` interaction model



Interactions with continuous covariates in randomized trials

MFPI method (Royston & Sauerbrei 2004)

- Consider continuous covariate *x*, binary randomized treatment variable *t*
  - Can adjust for other covariates
- Analysis follows the same principles as MFPIgen
- Get a function of *x* in each treatment group (level of *t*), based on main-effect model for *x*
- Consider just 2 groups – *t* binary
- Get an FP function with the same powers in each of the two treatment groups

MFPI in Stata

- MFPI is implemented as a user command, **`mfpi`**
- **`mfpi`** is available on SSC
- Details are given by Royston & Sauerbrei, *Stata Journal* **9**(2): 230-251 (2009)
- Program was updated in 2012 to support factor variables

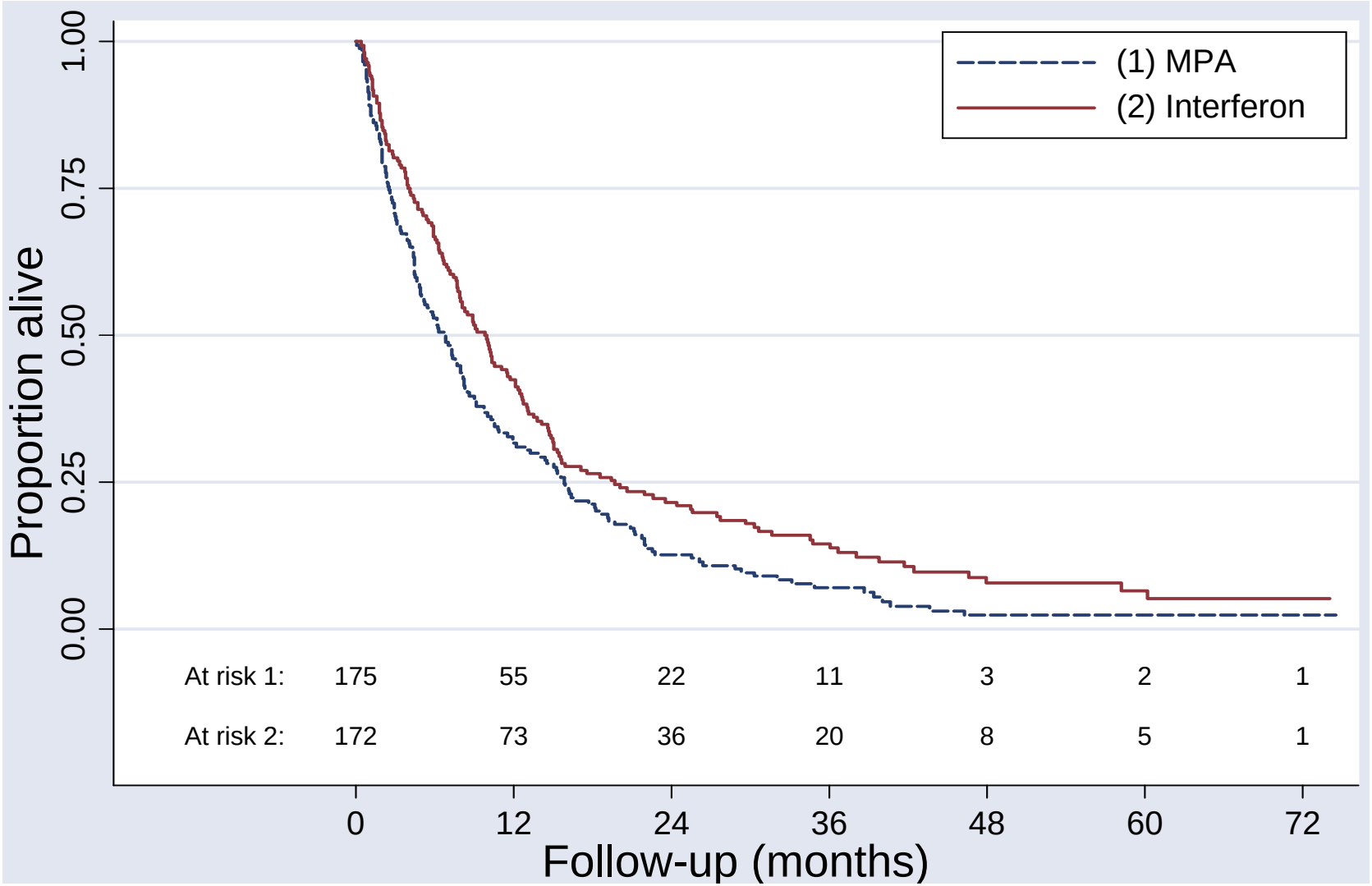
Treatment effect function

- Have estimated two functions – one per treatment group
- Plot the difference between functions against *x* to show the interaction
  - i.e. the treatment effect at different *x*
- Pointwise 95% CI shows how strongly the interaction is supported at different values of *x*
  - i.e. variation in the treatment effect with *x*

Example: MRC RE01 trial in kidney cancer

- Survival analysis (Cox regression)
- Main analysis: Interferon improves survival
- HR: 0.76 (0.62 - 0.95), P = 0.015
- Is the treatment effect similar in all patients?
- Nine possible covariates available for the investigation of treatment-covariate interactions
- Only one is significant – white cell count (**`wcc`**)

Kaplan-Meier showing treatment effect

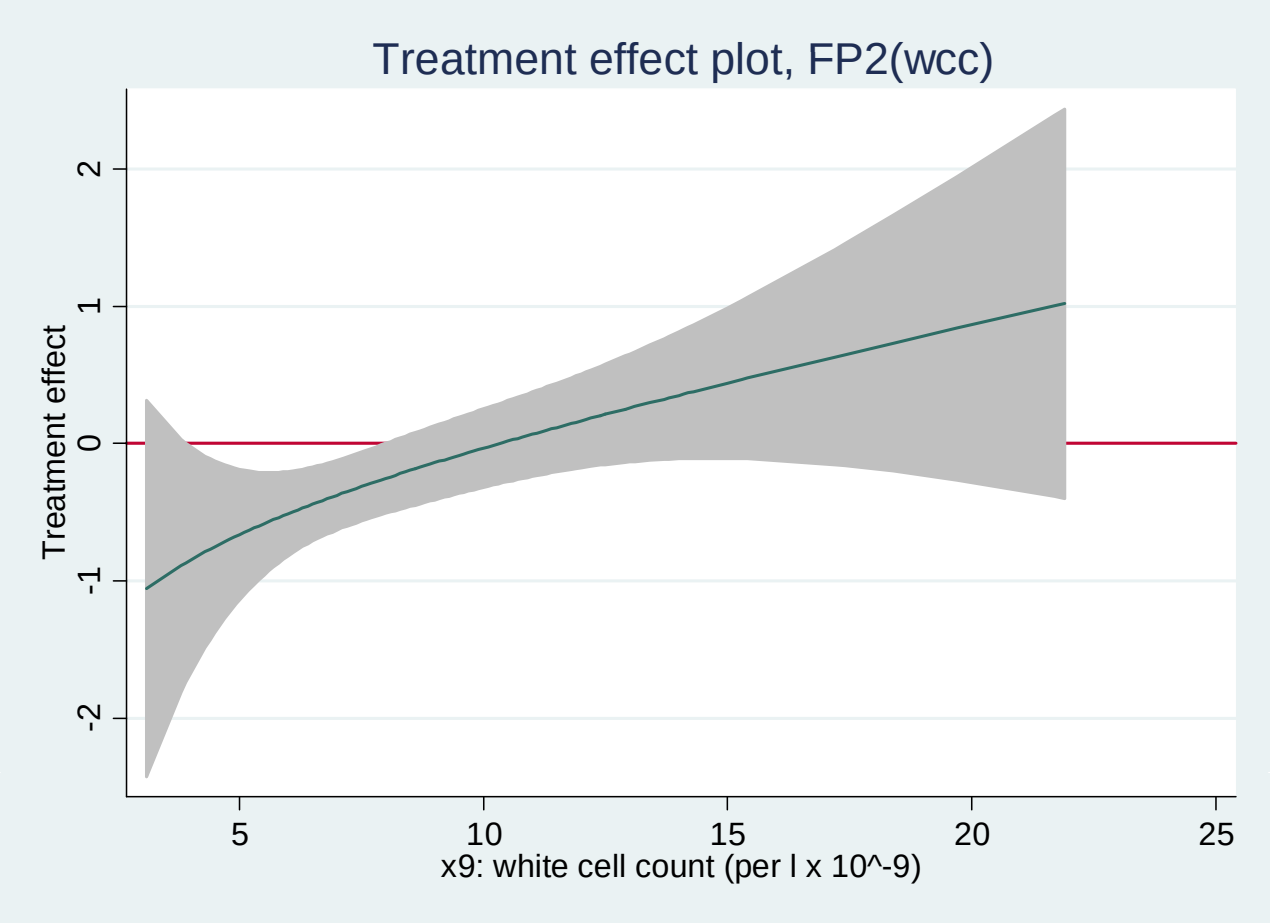


The **mfpi** command: example

- **wcc** has outliers, first truncate at 99th centile

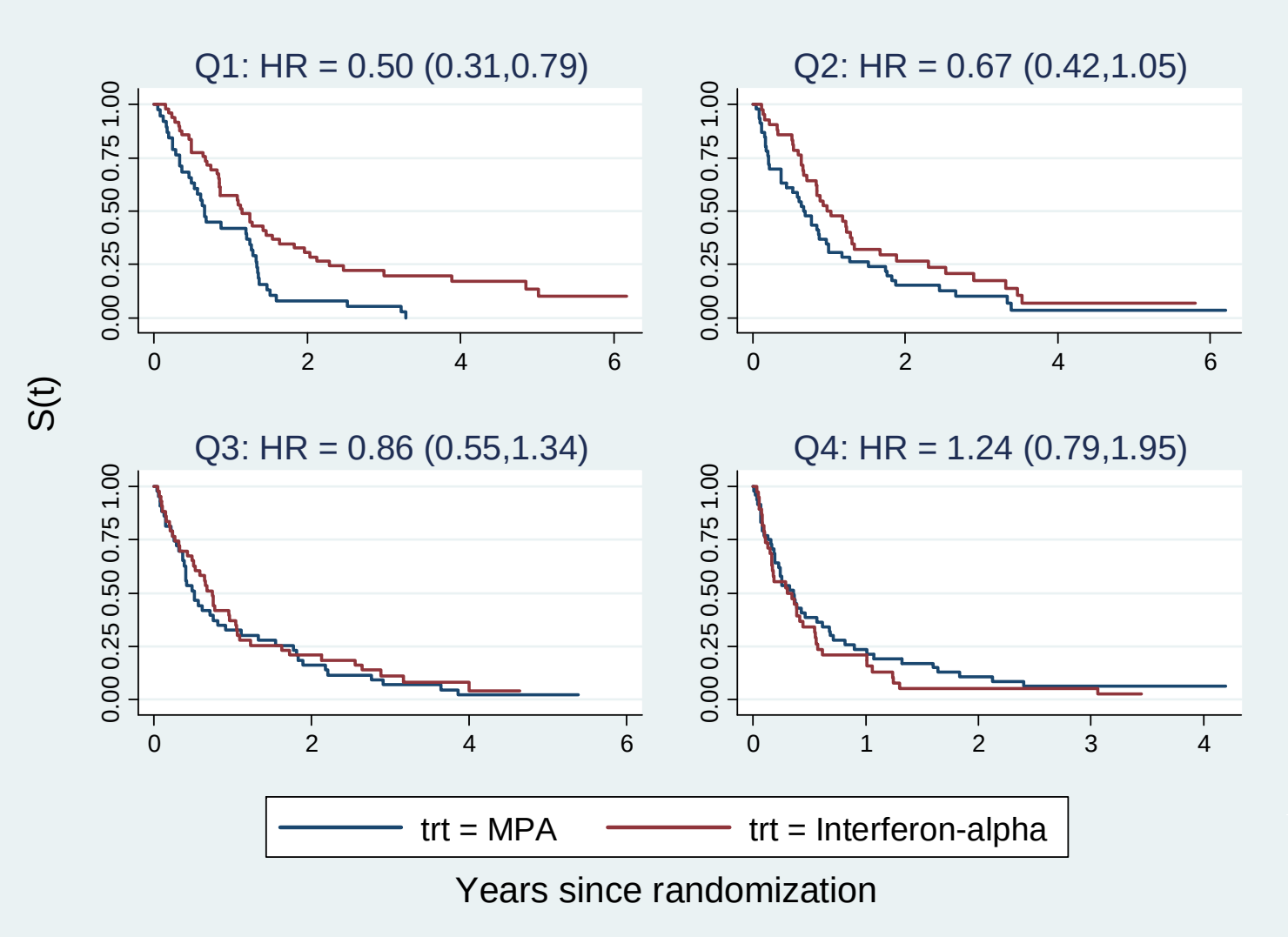
```
. mfpi, linear(wcc) fp1(wcc) fp2(wcc) with(trt)
gendiff(d): stcox
[treating trt as a factor variable, i.trt]

Interactions with i.trt (347 observations). Flex-1 model (least flexible)

-------------------------------------------------------------------------------
Var         Main         Interact     idf   Chi2     P      Deviance tdf   AIC
-------------------------------------------------------------------------------
wcc         Linear       Linear         1    8.13   0.0043  3186.561  3  3192.561
wcc         FP1(2)       FP1(2)         1    5.62   0.0178  3187.954  4  3195.954
wcc         FP2(-.5 1)   FP2(-.5 1)     2    8.19   0.0166  3185.237  7  3199.237
-------------------------------------------------------------------------------
idf = interaction degrees of freedom; tdf = total model degrees of freedom

. mfpi_plot wcc, vn(3)
[using variables created by gendiff(d)]
```

Treatment effect plot for **wcc**



About 25% of patients, those with WCC > 10 seem not to benefit from interferon

Checking the `wcc` x `trt` interaction model



Concluding remarks

- **`mfpigen`** and **`mfpi`** should help researchers detect, model and visualize interactions with continuous covariates
- Usually, we are searching for interactions, so small P-values are required
- Other methods not considered
  - STEPP – mainly graphical
  - …

Thank you.