Analysis of the salary distribution in Luxembourg

A finite mixture model approach

Jang SCHILTZ (University of Luxembourg)

joint work with
Jean-Daniel GUIGOU (University of Luxembourg),
Bruno LOVAT (University Nancy II)
& Cristian PREDA (University of Lille)

December 10, 2011

LUXEMBOURG SCHOOL OF FINANCE

Outline

Outline

1. Nagin's Finite Mixture Model
2. The Luxemburgish salary trajectories

Outline

1. Nagin's Finite Mixture Model
2. The Luxemburgish salary trajectories
3. Stability of the results

Outline

Outline

1. Nagin's Finite Mixture Model
2. The Luxemburgish salary trajectories
3. Stability of the results
4. Generalization of the basic model

LUXEMBOURG SCHOOL OF FINANCE

General description of Nagin's model

We have a collection of individual trajectories.

General description of Nagin's model

We have a collection of individual trajectories.

We try to divide the population into a number of homogenous subpopulations and to estimate a mean trajectory for each subpopulation.

General description of Nagin's model

We have a collection of individual trajectories.

We try to divide the population into a number of homogenous subpopulations and to estimate a mean trajectory for each subpopulation.

This is still an inter-individual model, but unlike other classical models such as standard growth curve models, it allows the existence of subpolulations with completely different behaviors.

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ for subject number i.

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ for subject number i.

$P(Y_i)$ denotes the probability of Y_i

LUXEMBOURG SCHOOL OF FINANCE

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ for subject number i.

$P(Y_i)$ denotes the probability of Y_i

- count data $\Rightarrow$ Poisson distribution

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ for subject number i.

$P(Y_i)$ denotes the probability of Y_i

- count data $\Rightarrow$ Poisson distribution
- binary data $\Rightarrow$ Binary logit distribution

LUXEMBOURG SCHOOL OF FINANCE

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ for subject number i.

$P(Y_i)$ denotes the probability of Y_i

- count data $\Rightarrow$ Poisson distribution
- binary data $\Rightarrow$ Binary logit distribution
- censored data $\Rightarrow$ Censored normal distribution

The Likelihood Function (1)

Consider a population of size N and a variable of interest Y.

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ for subject number i.

$P(Y_i)$ denotes the probability of Y_i

- count data $\Rightarrow$ Poisson distribution
- binary data $\Rightarrow$ Binary logit distribution
- censored data $\Rightarrow$ Censored normal distribution

Aim of the analysis: Find r groups of trajectories of a given kind (for instance polynomials of degree 4, $P(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$.

LUXEMBOURG SCHOOL OF FINANCE

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$$\Rightarrow \pi_j \text{ is the size of group } j.$$

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$\Rightarrow \pi_j$ is the size of group j.

We try to estimate a set of parameters $\Omega = \left\{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j\right\}$ which allow to maximize the probability of the measured data.

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$$\Rightarrow \pi_j \text{ is the size of group } j.$$

We try to estimate a set of parameters $\Omega = \left\{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j\right\}$ which allow to maximize the probability of the measured data.

$P^j(Y_i)$: probability of Y_i if subject i belongs to group j

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$$\Rightarrow \pi_j \text{ is the size of group } j.$$

We try to estimate a set of parameters $\Omega = \left\{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j\right\}$ which allow to maximize the probability of the measured data.

$P^j(Y_i)$: probability of Y_i if subject i belongs to group j

$$\Rightarrow P(Y_i) = \sum_{j=1}^{r} \pi_j P^j(Y_i). \qquad (1)$$

LUXEMBOURG SCHOOL OF FINANCE

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$$\Rightarrow \pi_j \text{ is the size of group } j.$$

We try to estimate a set of parameters $\Omega = \left\{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j\right\}$ which allow to maximize the probability of the measured data.

$P^j(Y_i)$: probability of Y_i if subject i belongs to group j

$$\Rightarrow P(Y_i) = \sum_{j=1}^{r} \pi_j P^j(Y_i). \quad (1)$$

Finite mixture model (Daniel S. Nagin (Carnegie Mellon University))

LUXEMBOURG SCHOOL OF FINANCE

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$$\Rightarrow \pi_j \text{ is the size of group } j.$$

We try to estimate a set of parameters $\Omega = \left\{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j\right\}$ which allow to maximize the probability of the measured data.

$P^j(Y_i)$: probability of Y_i if subject i belongs to group j

$$\Rightarrow P(Y_i) = \sum_{j=1}^{r} \pi_j P^j(Y_i). \qquad (1)$$

Finite mixture model (Daniel S. Nagin (Carnegie Mellon University))

- finite : sums across a finite number of groups

The Likelihood Function (2)

π_j : probability of a given subject to belong to group number j

$\Rightarrow \pi_j$ is the size of group j.

We try to estimate a set of parameters $\Omega = \left\{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j\right\}$ which allow to maximize the probability of the measured data.

$P^j(Y_i)$: probability of Y_i if subject i belongs to group j

$$\Rightarrow P(Y_i) = \sum_{j=1}^{r} \pi_j P^j(Y_i). \tag{1}$$

Finite mixture model (Daniel S. Nagin (Carnegie Mellon University))

- finite : sums across a finite number of groups
- mixture : population composed of a mixture of unobserved groups

The Likelihood Function (3)

Hypothesis: In a given group, conditional independence is assumed for the sequential realizations of the elements of Y_i !!!

The Likelihood Function (3)

Hypothesis: In a given group, conditional independence is assumed for the sequential realizations of the elements of Y_i !!!

$$\Rightarrow P^j(Y_i) = \prod_{t=1}^{T} p^j(y_{i_t}), \qquad (2)$$

where $p^j(y_{i_t})$ denotes the probability of y_{i_t} given membership in group j.

The Likelihood Function (3)

Hypothesis: In a given group, conditional independence is assumed for the sequential realizations of the elements of Y_i !!!

$$\Rightarrow P^j(Y_i) = \prod_{t=1}^{T} p^j(y_{i_t}), \tag{2}$$

where $p^j(y_{i_t})$ denotes the probability of y_{i_t} given membership in group j.

Likelihood of the estimator:

LUXEMBOURG SCHOOL OF FINANCE

The Likelihood Function (3)

Hypothesis: In a given group, conditional independence is assumed for the sequential realizations of the elements of Y_i !!!

$$\Rightarrow P^j(Y_i) = \prod_{t=1}^{T} p^j(y_{i_t}), \tag{2}$$

where $p^j(y_{i_t})$ denotes the probability of y_{i_t} given membership in group j.

Likelihood of the estimator:

$$L = \prod_{i=1}^{N} P(Y_i)$$

The Likelihood Function (3)

Hypothesis: In a given group, conditional independence is assumed for the sequential realizations of the elements of Y_i !!!

$$\Rightarrow P^j(Y_i) = \prod_{t=1}^{T} p^j(y_{i_t}), \tag{2}$$

where $p^j(y_{i_t})$ denotes the probability of y_{i_t} given membership in group j.

Likelihood of the estimator:

$$L = \prod_{i=1}^{N} P(Y_i) = \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} p^j(y_{i_t}). \tag{3}$$

The case of a censored normal distribution (1)

Y^*: latent variable measured by Y.

The case of a censored normal distribution (1)

Y^*: latent variable measured by Y.

$$y^*_{i_t} = \beta^j_0 + \beta^j_1 Age_{i_t} + \beta^j_2 Age^2_{i_t} + \beta^j_3 Age^3_{i_t} + \beta^j_4 Age^4_{i_t} + \varepsilon_{i_t}, \qquad (4)$$

where $\varepsilon_{i_t} \sim \mathcal{N}(0, \sigma)$, σ being a constant standard deviation.

The case of a censored normal distribution (1)

Y^*: latent variable measured by Y.

$$y^*_{i_t} = \beta^j_0 + \beta^j_1 Age_{i_t} + \beta^j_2 Age^2_{i_t} + \beta^j_3 Age^3_{i_t} + \beta^j_4 Age^4_{i_t} + \varepsilon_{i_t}, \qquad (4)$$

where $\varepsilon_{i_t} \sim \mathcal{N}(0, \sigma)$, σ being a constant standard deviation.

Hence,

$$\begin{aligned} y_{i_t} &= S_{min} \quad \text{si} \quad y^*_{i_t} < S_{min}, \\ y_{i_t} &= y^*_{i_t} \quad \text{si} \quad S_{min} \leq y^*_{i_t} \leq S_{max}, \\ y_{i_t} &= S_{max} \quad \text{si} \quad y^*_{i_t} > S_{max}, \end{aligned}$$

where S_{min} and S_{max} dennote the minimum and maximum of the censored normal distribution.

The case of a censored normal distribution (2)

Notations :

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta^j_0 + \beta^j_1 Age_{i_t} + \beta^j_2 Age^2_{i_t} + \beta^j_3 Age^3_{i_t} + \beta^j_4 Age^4_{i_t}$.

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta_0^j + \beta_1^j Age_{i_t} + \beta_2^j Age_{i_t}^2 + \beta_3^j Age_{i_t}^3 + \beta_4^j Age_{i_t}^4$.
- ϕ: density of standard centered normal law.

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta^j_0 + \beta^j_1 Age_{i_t} + \beta^j_2 Age^2_{i_t} + \beta^j_3 Age^3_{i_t} + \beta^j_4 Age^4_{i_t}$.
- ϕ: density of standard centered normal law.
- Φ: cumulative distribution function of standard centered normal law.

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta_0^j + \beta_1^j Age_{i_t} + \beta_2^j Age_{i_t}^2 + \beta_3^j Age_{i_t}^3 + \beta_4^j Age_{i_t}^4$.
- ϕ: density of standard centered normal law.
- Φ: cumulative distribution function of standard centered normal law.

Hence,

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta_0^j + \beta_1^j Age_{i_t} + \beta_2^j Age_{i_t}^2 + \beta_3^j Age_{i_t}^3 + \beta_4^j Age_{i_t}^4$.
- ϕ: density of standard centered normal law.
- Φ: cumulative distribution function of standard centered normal law.

Hence,

$$p^j(y_{i_t} = S_{min}) = \Phi\left(\frac{S_{min} - \beta^j t_{i_t}}{\sigma}\right), \tag{5}$$

LUXEMBOURG SCHOOL OF FINANCE

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta_0^j + \beta_1^j Age_{i_t} + \beta_2^j Age_{i_t}^2 + \beta_3^j Age_{i_t}^3 + \beta_4^j Age_{i_t}^4$.
- ϕ: density of standard centered normal law.
- Φ: cumulative distribution function of standard centered normal law.

Hence,

$$p^j(y_{i_t} = S_{min}) = \Phi\left(\frac{S_{min} - \beta^j t_{i_t}}{\sigma}\right), \tag{5}$$

$$p^j(y_{i_t}) = \frac{1}{\sigma}\phi\left(\frac{y_{i_t} - \beta^j t_{it}}{\sigma}\right) \quad \text{pour} \quad S_{min} \leq y_{it} \leq S_{max}, \tag{6}$$

LUXEMBOURG SCHOOL OF FINANCE

The case of a censored normal distribution (2)

Notations :

- $\beta^j t_{i_t} = \beta_0^j + \beta_1^j Age_{i_t} + \beta_2^j Age_{i_t}^2 + \beta_3^j Age_{i_t}^3 + \beta_4^j Age_{i_t}^4$.
- ϕ: density of standard centered normal law.
- Φ: cumulative distribution function of standard centered normal law.

Hence,

$$p^j(y_{i_t} = S_{min}) = \Phi\left(\frac{S_{min} - \beta^j t_{i_t}}{\sigma}\right), \quad (5)$$

$$p^j(y_{i_t}) = \frac{1}{\sigma}\phi\left(\frac{y_{i_t} - \beta^j t_{it}}{\sigma}\right) \quad \text{pour} \quad S_{min} \leq y_{it} \leq S_{max}, \quad (6)$$

$$p^j(y_{i_t} = S_{max}) = 1 - \Phi\left(\frac{S_{max} - \beta^j t_{i_t}}{\sigma}\right). \quad (7)$$

LUXEMBOURG SCHOOL OF FINANCE

The case of a censored normal distribution (3)

If all the measures are in the interval $[S_{min}, S_{max}]$, we get

The case of a censored normal distribution (3)

If all the measures are in the interval $[S_{min}, S_{max}]$, we get

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi \left(\frac{y_{i_t} - \beta^j t_{i_t}}{\sigma} \right). \tag{8}$$

The case of a censored normal distribution (3)

If all the measures are in the interval $[S_{min}, S_{max}]$, we get

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi\left(\frac{y_{i_t} - \beta^j t_{i_t}}{\sigma}\right). \tag{8}$$

It is too complicated to get closed-forms equations

The case of a censored normal distribution (3)

If all the measures are in the interval $[S_{min}, S_{max}]$, we get

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi \left(\frac{y_{i_t} - \beta^j t_{i_t}}{\sigma} \right). \quad (8)$$

It is too complicated to get closed-forms equations

$\Rightarrow$ quasi-Newton procedure maximum research routine

The case of a censored normal distribution (3)

If all the measures are in the interval $[S_{min}, S_{max}]$, we get

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi \left(\frac{y_{i_t} - \beta^j t_{i_t}}{\sigma} \right). \tag{8}$$

It is too complicated to get closed-forms equations

$\Rightarrow$ quasi-Newton procedure maximum research routine

Software:

SAS-based Proc Traj procedure
by Bobby L. Jones (Carnegie Mellon University).

LUXEMBOURG SCHOOL OF FINANCE

A computational trick

The estimations of π_j must be in $[0, 1]$.

A computational trick

The estimations of π_j must be in $[0, 1]$.

It is difficult to force this constraint in model estimation.

A computational trick

The estimations of π_j must be in $[0, 1]$.

It is difficult to force this constraint in model estimation.

Instead, we estimate the real parameters θ_j such that

A computational trick

The estimations of π_j must be in $[0, 1]$.

It is difficult to force this constraint in model estimation.

Instead, we estimate the real parameters θ_j such that

$$\pi_j = \frac{e^{\theta_j}}{\sum_{j=1}^{r} e^{\theta_j}}, \tag{9}$$

A computational trick

The estimations of π_j must be in $[0, 1]$.

It is difficult to force this constraint in model estimation.

Instead, we estimate the real parameters θ_j such that

$$\pi_j = \frac{e^{\theta_j}}{\sum_{j=1}^{r} e^{\theta_j}}, \tag{9}$$

Finally,

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \frac{e^{\theta_j}}{\sum_{j=1}^{r} e^{\theta_j}} \prod_{t=1}^{T} \phi \left(\frac{y_{i_t} - \beta^j t_{i_t}}{\sigma} \right). \tag{10}$$

LUXEMBOURG SCHOOL OF FINANCE

Muthén's model (1)

Muthén and Shedden (1999): Generalized growth curve model

Muthén's model (1)

Muthén and Shedden (1999): Generalized growth curve model

Elegant and technically demanding extension of the uncensored normal model.

Muthén's model (1)

Muthén and Shedden (1999): Generalized growth curve model

Elegant and technically demanding extension of the uncensored normal model.

Adds random effects to the parameters β^j that define a group's mean trajectory.

LUXEMBOURG SCHOOL OF FINANCE

Muthén's model (1)

Muthén and Shedden (1999): Generalized growth curve model

Elegant and technically demanding extension of the uncensored normal model.

Adds random effects to the parameters β^j that define a group's mean trajectory.

Trajectories of individual group members can vary from the group trajectory.

Muthén's model (1)

Muthén and Shedden (1999): Generalized growth curve model

Elegant and technically demanding extension of the uncensored normal model.

Adds random effects to the parameters β^j that define a group's mean trajectory.

Trajectories of individual group members can vary from the group trajectory.

Software:

Mplus package by L.K. Muthén and B.O Muthén.

Muthén's model (2)

Advantage of GGCM

Fewer groups are required to specify a satisfactory model.

Muthén's model (2)

Advantage of GGCM

Fewer groups are required to specify a satisfactory model.

Disadvantages of GGCM

1. Difficult to extend to other types of data.

LUXEMBOURG SCHOOL OF FINANCE

Muthén's model (2)

Advantage of GGCM

Fewer groups are required to specify a satisfactory model.

Disadvantages of GGCM

1. Difficult to extend to other types of data.
2. Group cross-over effects.

Muthén's model (2)

Advantage of GGCM

Fewer groups are required to specify a satisfactory model.

Disadvantages of GGCM

1. Difficult to extend to other types of data.
2. Group cross-over effects.
3. can create the illusion of non-existing groups.

Model Selection

Bayesian Information Criterion:

Model Selection

Bayesian Information Criterion:

$$\mathrm{BIC} = \log(L) - 0,5k\log(N), \tag{11}$$

where k denotes the number of parameters in the model.

Model Selection

Bayesian Information Criterion:

$$\text{BIC} = \log(L) - 0,5k\log(N), \tag{11}$$

where k denotes the number of parameters in the model.

Rule:

The bigger the BIC, the better the model!

Posterior Group-Membership Probabilities

Posterior Group-Membership Probabilities

Posterior probability of individual i's membership in group j : $P(j/Y_i)$.

Posterior Group-Membership Probabilities

Posterior probability of individual i's membership in group j : $P(j/Y_i)$.

Bayes's theorem

$$\Rightarrow P(j/Y_i) = \frac{P(Y_i/j)\hat{\pi}_j}{\sum_{j=1}^{r} P(Y_i/j)\hat{\pi}_j}. \tag{12}$$

Posterior Group-Membership Probabilities

Posterior probability of individual i's membership in group j : $P(j/Y_i)$.

Bayes's theorem

$$\Rightarrow P(j/Y_i) = \frac{P(Y_i/j)\hat{\pi}_j}{\sum_{j=1}^{r} P(Y_i/j)\hat{\pi}_j}. \quad (12)$$

Bigger groups have on average larger probability estimates.

LUXEMBOURG SCHOOL OF FINANCE

Posterior Group-Membership Probabilities

Posterior probability of individual i's membership in group j : $P(j/Y_i)$.

Bayes's theorem

$$\Rightarrow P(j/Y_i) = \frac{P(Y_i/j)\hat{\pi}_j}{\sum_{j=1}^{r} P(Y_i/j)\hat{\pi}_j}. \tag{12}$$

Bigger groups have on average larger probability estimates.

To be classified into a small group, an individual really needs to be strongly consistent with it.

LUXEMBOURG SCHOOL OF FINANCE

Use for Model diagnostics (2)

Diagnostic 1: Average Posterior Probability of Assignment

AvePP should be at least $0,7$ for all groups.

Use for Model diagnostics (2)

Diagnostic 1: Average Posterior Probability of Assignment

AvePP should be at least $0,7$ for all groups.

Diagonostic 2: Odds of Correct Classification

$$OCC_j = \frac{AvePP_j/1 - AvePP_j}{\hat{\pi}_j/1 - \hat{\pi}_j}. \tag{13}$$

LUXEMBOURG
SCHOOL OF FINANCE

Use for Model diagnostics (2)

Diagnostic 1: Average Posterior Probability of Assignment

AvePP should be at least $0,7$ for all groups.

Diagonostic 2: Odds of Correct Classification

$$OCC_j = \frac{AvePP_j/1 - AvePP_j}{\hat{\pi}_j/1 - \hat{\pi}_j}. \tag{13}$$

OCC_j should be greater than 5 for all groups.

LUXEMBOURG SCHOOL OF FINANCE

Use for Model diagnostics (2)

Diagonostic 3: Comparing $\hat{\pi}_j$ to the Proportion of the Sample Assigned to Group j

The ratio of the two should be close to 1.

Use for Model diagnostics (2)

Diagonostic 3: Comparing $\hat{\pi}_j$ to the Proportion of the Sample Assigned to Group j

The ratio of the two should be close to 1.

Diagonostic 4: Confidence Intervals for Group Membership Probabilities

The confidence intervals for group membership probabilities estimates should be narrow, i.e. standard deviation of π_j should be small.

Outline

LUXEMBOURG SCHOOL OF FINANCE

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

About 7 million salary lines corresponding to 718.054 workers.

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

About 7 million salary lines corresponding to 718.054 workers.

Some sociological variables:

- gender (male, female)

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

About 7 million salary lines corresponding to 718.054 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship (luxemburgish residents, foreign residents, foreign non residents)

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

About 7 million salary lines corresponding to 718.054 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship (luxemburgish residents, foreign residents, foreign non residents)
- working status (white collar worker, blue collar worker)

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

About 7 million salary lines corresponding to 718.054 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship (luxemburgish residents, foreign residents, foreign non residents)
- working status (white collar worker, blue collar worker)
- year of birth

The data : first dataset

Salaries of workers in the private sector in Luxembourg from 1940 to 2006.

About 7 million salary lines corresponding to 718.054 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship (luxemburgish residents, foreign residents, foreign non residents)
- working status (white collar worker, blue collar worker)
- year of birth
- age in the first year of professional activity

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship
- sector of activity

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship
- sector of activity
- year of birth

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship
- sector of activity
- year of birth
- age in the first year of professional activity

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship
- sector of activity
- year of birth
- age in the first year of professional activity
- marital status

The data : second dataset

Salaries of all workers in Luxembourg which began to work in Luxembourg between 1980 and 1990 at an age less than 30 years.

1.303.010 salary lines corresponding to 85.049 workers.

Some sociological variables:

- gender (male, female)
- nationality and residentship
- sector of activity
- year of birth
- age in the first year of professional activity
- marital status
- year of birth of children

Dataset transformations

Mathematica programming

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker
- selection of the period we are interested in

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker
- selection of the period we are interested in
- taking out the years without work up to a maximum of five years

LUXEMBOURG SCHOOL OF FINANCE

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker
- selection of the period we are interested in
- taking out the years without work up to a maximum of five years
- selection of the people who worked at least during 20 years

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker
- selection of the period we are interested in
- taking out the years without work up to a maximum of five years
- selection of the people who worked at least during 20 years

Transformations in SPSS

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker
- selection of the period we are interested in
- taking out the years without work up to a maximum of five years
- selection of the people who worked at least during 20 years

Transformations in SPSS

- the number of months problem!
- elimination of all the workers who had monthly salaries above 15.000

Dataset transformations

Mathematica programming

- 1 row per year $\rightarrow$ 1 row per worker
- selection of the period we are interested in
- taking out the years without work up to a maximum of five years
- selection of the people who worked at least during 20 years

Transformations in SPSS

- the number of months problem!
- elimination of all the workers who had monthly salaries above 15.000
- creation of the time variables necessary for the Proc Traj procedure

Proc Traj procedure

Selection of the time period for macroeconomic reasons

Proc Traj procedure

Selection of the time period for macroeconomic reasons (Crisis in the steel industry and emergence of the financial market place of Luxembourg)

Proc Traj procedure

Selection of the time period for macroeconomic reasons (Crisis in the steel industry and emergence of the financial market place of Luxembourg)

20 years of work for workers beginning their carrier between 1982 and 1987

Proc Traj procedure

Selection of the time period for macroeconomic reasons (Crisis in the steel industry and emergence of the financial market place of Luxembourg)

20 years of work for workers beginning their carrier between 1982 and 1987

Proc Traj Macro:

Proc Traj procedure

Selection of the time period for macroeconomic reasons (Crisis in the steel industry and emergence of the financial market place of Luxembourg)

20 years of work for workers beginning their carrier between 1982 and 1987

Proc Traj Macro:

```
DATA TEST;
    INPUT ID O1-O20 T1-T20;
    CARDS;
data
RUN;
```

Proc Traj procedure

Selection of the time period for macroeconomic reasons (Crisis in the steel industry and emergence of the financial market place of Luxembourg)

20 years of work for workers beginning their carrier between 1982 and 1987

Proc Traj Macro:

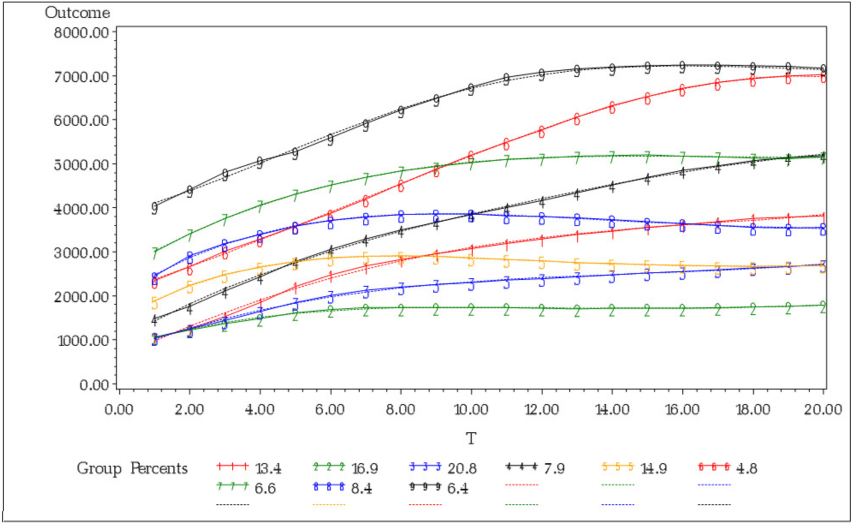
```
DATA TEST;
    INPUT ID O1-O20 T1-T20;
    CARDS;
data
RUN;

PROC TRAJ DATA=TEST OUTPLOT=OP OUTSTAT=OS OUT=OF
OUTEST=OE ITDETAIL;
    ID ID; VAR O1-O20; INDEP T1-T20;
    MODEL CNORM; MAX 8000; NGROUPS 6; ORDER 4 4 4 4 4 4;
RUN;
```

LUXEMBOURG SCHOOL OF FINANCE

Result for 9 groups (dataset 1)

Result for 9 groups (dataset 1)



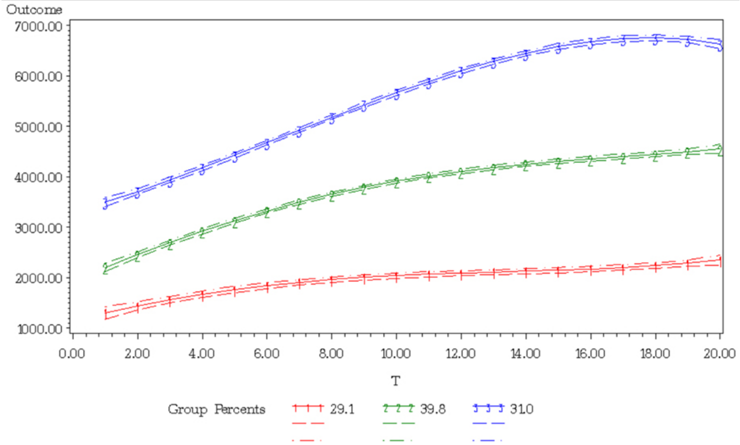
Results for 9 groups (dataset 1)

Maximum Likelihood Estimates
Model: Censored Normal (CNORM)

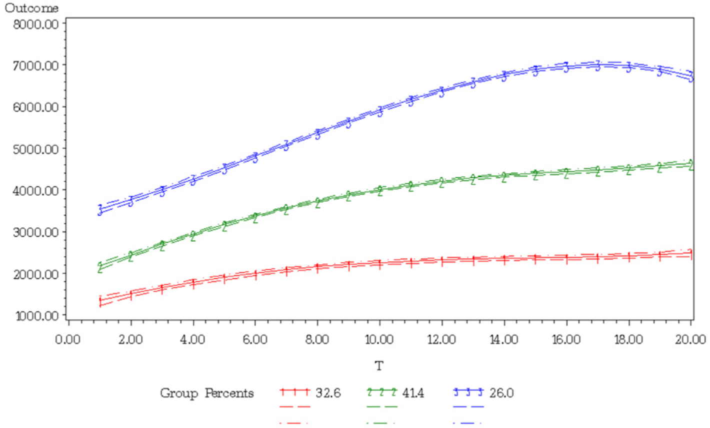
Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	589.03067	18.46813	31.894	0.0000
	Linear	387.72145	11.31617	34.263	0.0000
	Quadratic	-14.36621	2.12997	-6.745	0.0000
	Cubic	-0.01563	0.15109	-0.103	0.9176
	Quartic	0.00856	0.00358	2.395	0.0166
2	Intercept	784.79156	15.75939	49.798	0.0000
	Linear	277.63602	9.78078	28.386	0.0000
	Quadratic	-28.36731	1.83236	-15.481	0.0000
	Cubic	1.17739	0.12972	9.076	0.0000
	Quartic	-0.01635	0.00307	-5.330	0.0000
3	Intercept	709.28728	15.90545	44.594	0.0000
	Linear	318.88029	8.97949	35.512	0.0000
	Quadratic	-21.54540	1.69611	-12.703	0.0000
	Cubic	0.62010	0.12002	5.167	0.0000
	Quartic	-0.00440	0.00284	-1.554	0.1203

Outline

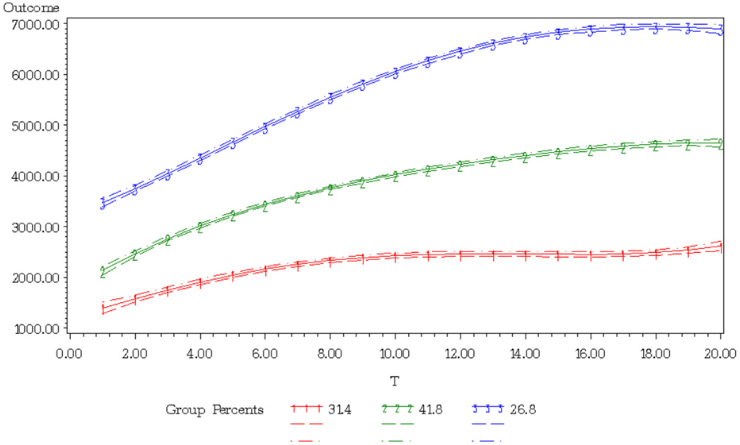
Result for 3 groups (dataset 2): workers beginning their career in 1982



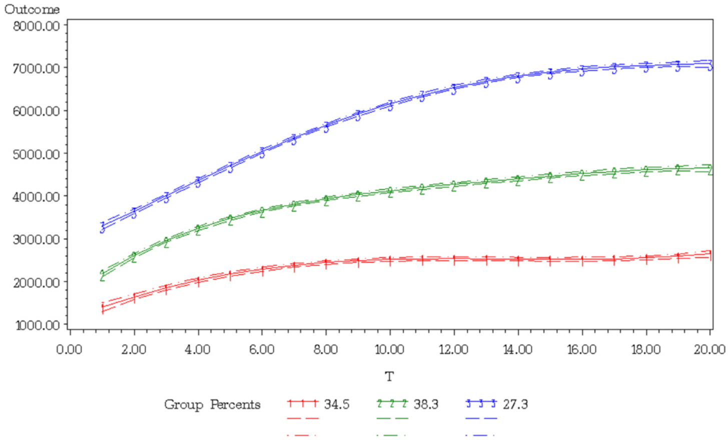
Result for 3 groups (dataset 2): workers beginning their career in 1983



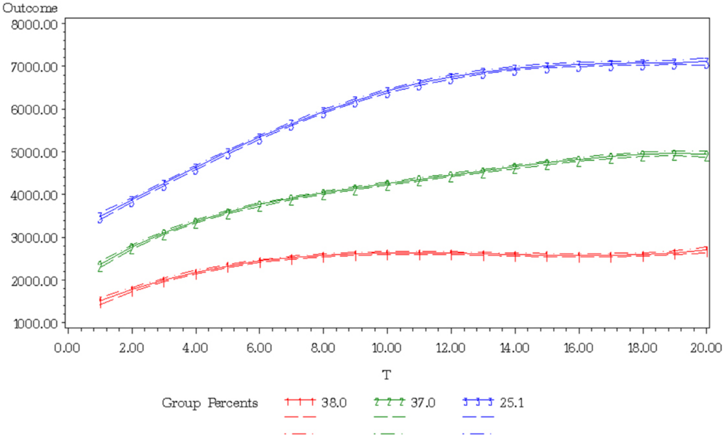
Result for 3 groups (dataset 2): workers beginning their career in 1984



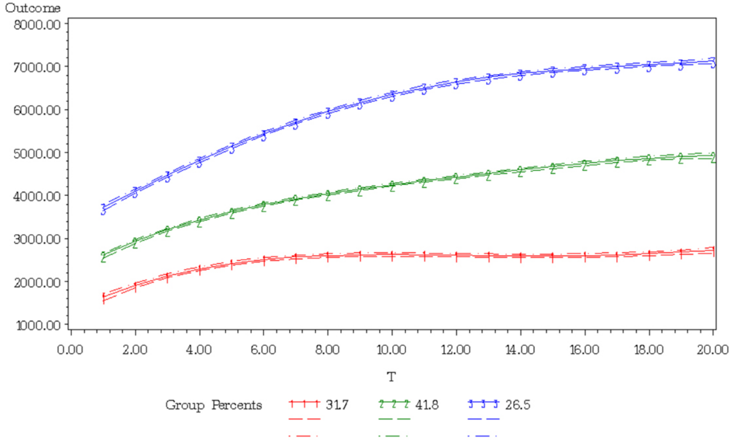
Result for 3 groups (dataset 2): workers beginning their career in 1985



Result for 3 groups (dataset 2): workers beginning their career in 1986



Result for 3 groups (dataset 2): workers beginning their career in 1987



The statistical shape analysis approach

The statistical shape analysis approach

Comparing the geometrical figure of the trajectories

The statistical shape analysis approach

Comparing the geometrical figure of the trajectories

$\longrightarrow$ statistical shape analyis:

The statistical shape analysis approach

Comparing the geometrical figure of the trajectories

$\longrightarrow$ statistical shape analyis:

Compute the mean shape of the different results.

The statistical shape analysis approach

Comparing the geometrical figure of the trajectories

$\longrightarrow$ statistical shape analyis:

Compute the mean shape of the different results.

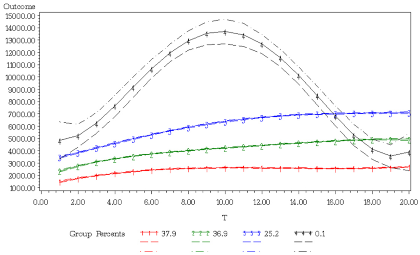
Use Ziezold's test for every set of trajectories to see if it is significantly different from the mean set of trajectories.

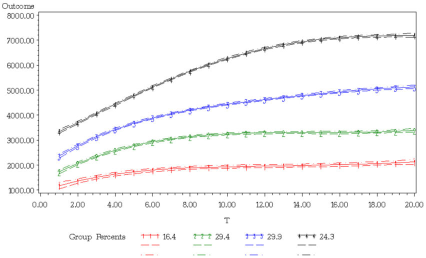
The statistical shape analysis approach

Are these set of trajectories different?

The statistical shape analysis approach

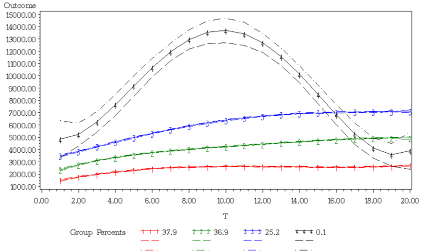
Are these set of trajectories different?

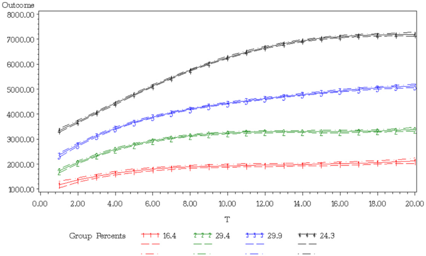




The statistical shape analysis approach

Are these set of trajectories different?





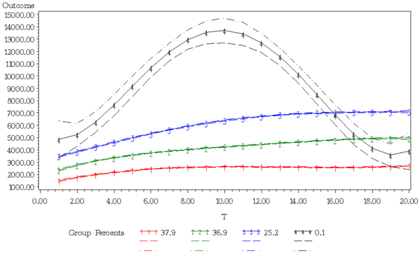
Shape Analysis says yes!

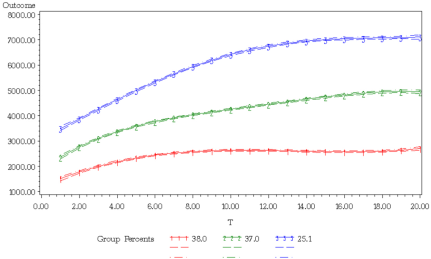
The statistical shape analysis approach

Are these set of trajectories different?

The statistical shape analysis approach

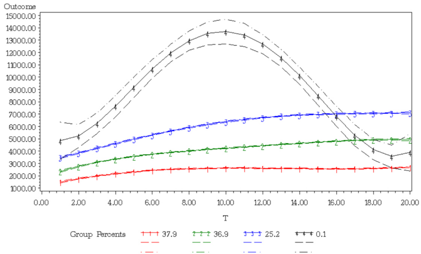
Are these set of trajectories different?

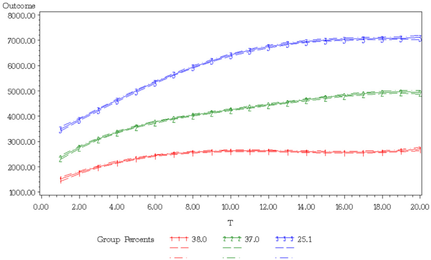




The statistical shape analysis approach

Are these set of trajectories different?

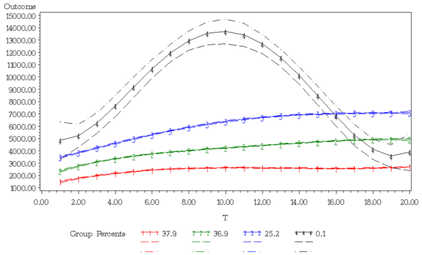


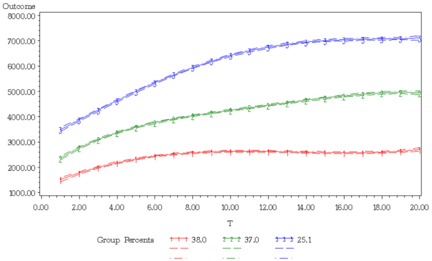


Shape Analysis says yes,

The statistical shape analysis approach

Are these set of trajectories different?





Shape Analysis says yes, but are they really?

The classical statistics approach

The classical statistics approach

Compare the estimated parameters:

The classical statistics approach

Compare the estimated parameters:

- Performing the Wald test to see if the parameters differ between two models.

The classical statistics approach

Compare the estimated parameters:

- Performing the Wald test to see if the parameters differ between two models.
- Compare the confidence intervals of the parameters and see if they have an intersection.

Functional Data Analysis Approach

Functional Data Analysis Approach

Compare the set of trajectories as functions:

Functional Data Analysis Approach

Compare the set of trajectories as functions:

Consider a metrical space on the continuous functions defined on the time interval of the trajectories and use tests on functional data to analyze the time stability of the results.

Outline

LUXEMBOURG SCHOOL OF FINANCE

Predictors of trajectory group membership

Predictors of trajectory group membership

x_i : vector of variables potentially associated with group membership (measured before t_1).

Predictors of trajectory group membership

x_i : vector of variables potentially associated with group membership (measured before t_1).

Multinomial logit model:

$$\pi_j(x_i) = \frac{e^{x_i \theta_j}}{\sum_{k=1}^{r} e^{x_i \theta_k}}, \tag{14}$$

where θ_j denotes the effect of x_i on the probability of group membership.

LUXEMBOURG SCHOOL OF FINANCE

Predictors of trajectory group membership

x_i : vector of variables potentially associated with group membership (measured before t_1).

Multinomial logit model:

$$\pi_j(x_i) = \frac{e^{x_i\theta_j}}{\sum_{k=1}^{r} e^{x_i\theta_k}}, \tag{14}$$

where θ_j denotes the effect of x_i on the probability of group membership.

$$L = \frac{1}{\sigma}\prod_{i=1}^{N}\sum_{j=1}^{r}\frac{e^{x_i\theta_j}}{\sum_{k=1}^{r} e^{x_i\theta_k}}\prod_{t=1}^{T}\phi\left(\frac{y_{i_t} - \beta^j t_{i_t}}{\sigma}\right). \tag{15}$$

Group membership probabilities

Group membership probabilities

The Wald test which indicates whether any number of ocefficients is significally different, allows the statistical testing of the predictors.

Group membership probabilities

The Wald test which indicates whether any number of ocefficients is significally different, allows the statistical testing of the predictors.

Confidence intervals for the probabilities of group membership can be computed by a parametric bootstrap technique.

Group membership probabilities: macro

Proc Traj Macro:

Group membership probabilities: macro

Proc Traj Macro:

```
DATA TEST;
      INPUT ID O1-O20 T1-T20 NATIO SEXE;
      CARDS;
data
RUN;
```

Group membership probabilities: macro

Proc Traj Macro:

```
DATA TEST;
      INPUT ID O1-O20 T1-T20 NATIO SEXE;
      CARDS;
data
RUN;

PROC TRAJ DATA=TEST OUTPLOT=OP OUTSTAT=OS OUT=OF
OUTEST=OE ITDETAIL;
      ID ID; VAR O1-O20; INDEP T1-T20;
      MODEL CNORM; MAX 15000; NGROUPS 3; ORDER 4 4 4;
      RISK NATIO SEXE; RUN;
```

LUXEMBOURG SCHOOL OF FINANCE

Group membership probabilities: results

LUXEMBOURG SCHOOL OF FINANCE

Group membership probabilities: results

Maximum Likelihood Estimates
Model: Censored Normal (CNORM)

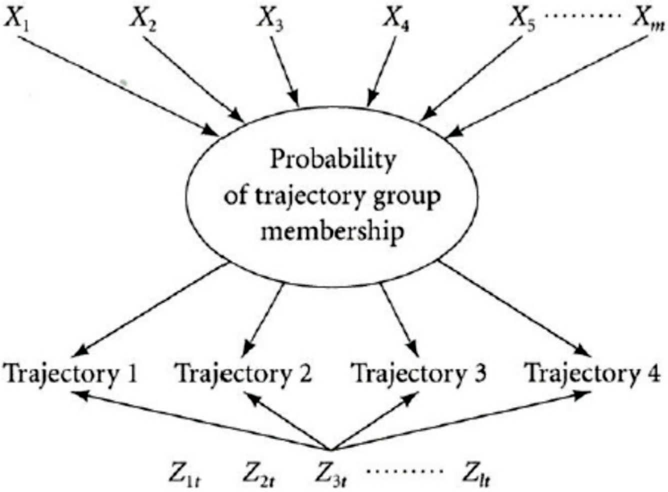
Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	1275.04779	62.99065	20.242	0.0000
	Linear	389.21202	38.92642	9.999	0.0000
	Quadratic	-39.15162	7.30285	-5.361	0.0000
	Cubic	1.59498	0.51761	3.081	0.0021
	Quartic	-0.02140	0.01226	-1.746	0.0808
2	Intercept	2222.80424	54.84704	40.527	0.0000
	Linear	424.05294	34.28352	12.369	0.0000
	Quadratic	-37.06840	6.43629	-5.759	0.0000
	Cubic	1.86457	0.45611	4.088	0.0000
	Quartic	-0.03661	0.01079	-3.392	0.0007
3	Intercept	3320.52407	69.05348	48.086	0.0000
	Linear	404.79252	43.10582	9.391	0.0000
	Quadratic	-4.60135	8.09472	-0.568	0.5697
	Cubic	-0.80156	0.57341	-1.398	0.1622
	Quartic	0.02479	0.01356	1.828	0.0675
	Sigma	931.29644	3.56268	261.403	0.0000
	Group membership				
1	Constant	(0.00000)	.	.	.
2	Constant	0.42351	0.11171	3.791	0.0002
	NATIO	0.11996	0.02712	4.424	0.0000
	SEXE	-0.76451	0.12195	-6.269	0.0000
3	Constant	0.13833	0.11999	1.153	0.2490
	NATIO	0.21875	0.03024	7.233	0.0000
	SEXE	-2.08600	0.15090	-13.823	0.0000

BIC=-285105.7 (N=34320) BIC=-285072.7 (N=1716) AIC=-285012.8 L=-284990.8

LUXEMBOURG SCHOOL OF FINANCE

Adding covariates to the trajectories (1)

Adding covariates to the trajectories (1)



Adding covariates to the trajectories (2)

Adding covariates to the trajectories (2)

Y^*: latent variable measured by Y.

Adding covariates to the trajectories (2)

Y^*: latent variable measured by Y.

$$y^*_{i_t} = \beta^j_0 + \beta^j_1 t + \beta^j_2 t^2 + \beta^j_3 t^3 + \beta^j_4 t^4 + \alpha^j_1 z_{1t} + ... + \alpha^j_L z_{Lt} + \varepsilon_{i_t}, \quad (16)$$

where $\varepsilon_{i_t} \sim \mathcal{N}(0, \sigma)$, σ being a constant standard deviation and z_{lt} are covariates that may depend or not upon time t.

Adding covariates to the trajectories (2)

Y^*: latent variable measured by Y.

$$y^*_{i_t} = \beta^j_0 + \beta^j_1 t + \beta^j_2 t^2 + \beta^j_3 t^3 + \beta^j_4 t^4 + \alpha^j_1 z_{1t} + ... + \alpha^j_L z_{Lt} + \varepsilon_{i_t}, \quad (16)$$

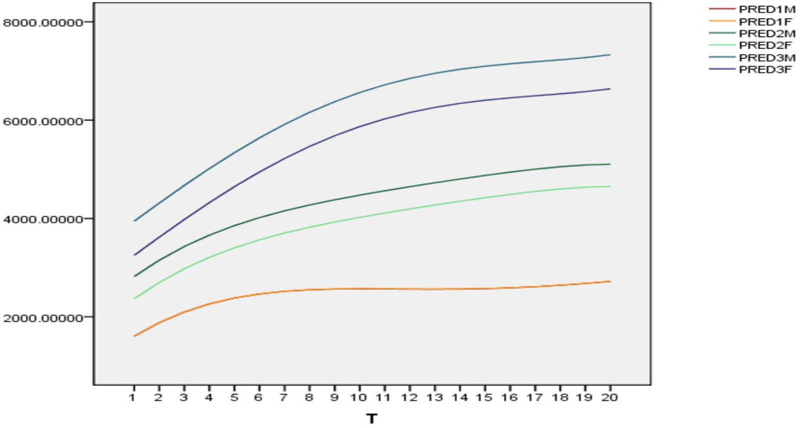
where $\varepsilon_{i_t} \sim \mathcal{N}(0, \sigma)$, σ being a constant standard deviation and z_{lt} are covariates that may depend or not upon time t.

Unfortunately the estimation of parameters α^j_l is not implemented in proc traj procedure; it is just possible to plot the impact of the covariates.

LUXEMBOURG SCHOOL OF FINANCE

Adding covariates to the trajectories (3)

Adding covariates to the trajectories (3)



Functional PLS regression with functional response

Functional PLS regression with functional response

Response $Y = \{Y_t\}_{t \in \mathcal{T}_Y}$ and predictor $X = \{X_t\}_{t \in \mathcal{T}_X}$ of functional type.

Functional PLS regression with functional response

Response $Y = \{Y_t\}_{t \in \mathcal{T}_Y}$ and predictor $X = \{X_t\}_{t \in \mathcal{T}_X}$ of functional type.

The Escoufier operators associated to X and Y are defined by

$$W^X Z = \int_{\mathcal{T}_X} \mathbb{E}(X_t Z) X_t dt, \quad \forall \text{ r.v } Z \tag{17}$$

and

$$W^Y Z = \int_{\mathcal{T}_Y} \mathbb{E}(Y_t Z) Y_t dt, \quad \forall \text{ r.v } Z.$$

Functional PLS regression with functional response

Response $Y = \{Y_t\}_{t \in \mathcal{T}_Y}$ and predictor $X = \{X_t\}_{t \in \mathcal{T}_X}$ of functional type.

The Escoufier operators associated to X and Y are defined by

$$W^X Z = \int_{\mathcal{T}_X} \mathbb{E}(X_t Z) X_t dt, \quad \forall \text{ r.v } Z \tag{17}$$

and

$$W^Y Z = \int_{\mathcal{T}_Y} \mathbb{E}(Y_t Z) Y_t dt, \quad \forall \text{ r.v } Z. \tag{18}$$

Theorem (C.Preda, J.S., 2011)

At each step of the PLS regression, the PLS components $t_h, h > 1$ are eigenvectors of the product of the two Esoufier operators i.e.

$$W^X W^Y t_h = \lambda_h. \tag{19}$$

Bibliography

- Nagin, D.S. 2005: *Group-based Modeling of Development*. Cambridge, MA.: Harvard University Press.
- Jones, B. and Nagin D.S. 2007: Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them. *Sociological Research and Methods*, **35** p.542-571.
- Guigou, J.D, Lovat, B. and Schiltz, J. 2012: Analysis of the salary trajectories in Luxembourg : a finite mixture model approach. To appear.
- Giebel S. 2011: *Zur Anwendung der statistischen Formanalyse*. Phd Thesis, University of Luxembourg.
- Preda, C. and Schiltz, J. 2012: Functional PLS regression with functional response: the basis expansion approach. To appear.

LUXEMBOURG SCHOOL OF FINANCE