Midterm Group Meeting

Seminar in Advanced Microeconometrics

Rémi Piatek

Department of Economics
University of Copenhagen

Seminar

Wikipedia definition

A seminar is a form of academic instruction, either at an academic institution or offered by a commercial or professional organization. It has the function of **bringing together small groups for recurring meetings, focusing each time on some particular subject, in which everyone present is requested to participate.** This is often accomplished through an ongoing Socratic dialogue with a seminar leader or instructor, or through a more formal presentation of research. It is essentially a place where assigned readings are discussed, questions can be raised and debates can be conducted.

Getting feedback, exchanging

- **Seminar:** No individual supervision, **group collaboration**.
- Don't send me emails with questions, use **online forum** instead:
  - Ask your questions.
  - Reply to questions asked by others.
  - I will step in and give feedback to those who also contribute by replying to questions.
- You will all face the same kinds of problems: Sharing your experience will pay off.
- Everyone can benefit from the comments (contrary to emails).

Today's Midterm Group Meeting

Goals of this meeting

- **Status** of each project:
  - What are you doing?
  - How far are you?
  - What are your main problems?
- **Time constraint**:
  - Short presentation (3/4 min).
  - Group discussion/feedback.
  - **Total time:** 10 min/group, 7 min for individual projects.

Brief statement of **research question** and **motivation** of the topic

Research question: How do cognitive and non-cognitive factors during childhood influence young adults' likelihood of completing high school (in the US)?

- Logistic regression
- Data from NLSY79
- Focus on children aged 5, born in 1980 because we have most data from that subgroup

Motivation: both cognitive and non-cognitive factors have been shown to be important – which type of factor has the largest impact→ determine ”most important” explanatory variables using the Lasso approach, i.e. adding penalty term to the logistic regression

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i \log \Pr(Y=1 \mid x_i) + (1-y_i)\log \Pr(Y=0 \mid x_i)\} + \lambda\|\beta\|_1$$

Status of our project

- What have we done so far?
  - Spent a lot of time trying to understand the theory behind Lasso
  - Coding: normal logistic, weighted least squares, starting on coding Lasso loops

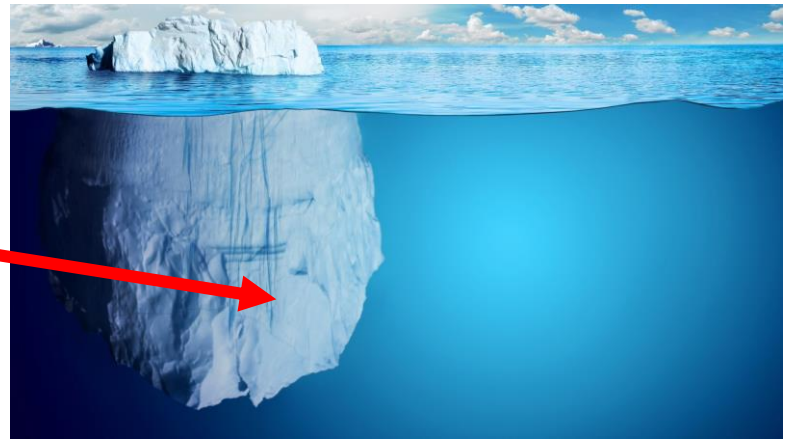
- What will we do next?
  - Code the loops
  - Look at standard errors

OUTER LOOP: Decrement λ.

MIDDLE LOOP: Update the quadratic approximation ℓ_Q using the current parameters $(\tilde{\beta}_0, \tilde{\beta})$.

INNER LOOP: Run the coordinate descent algorithm on the penalized weighted-least-squares problem (5.56).

- How far are we?

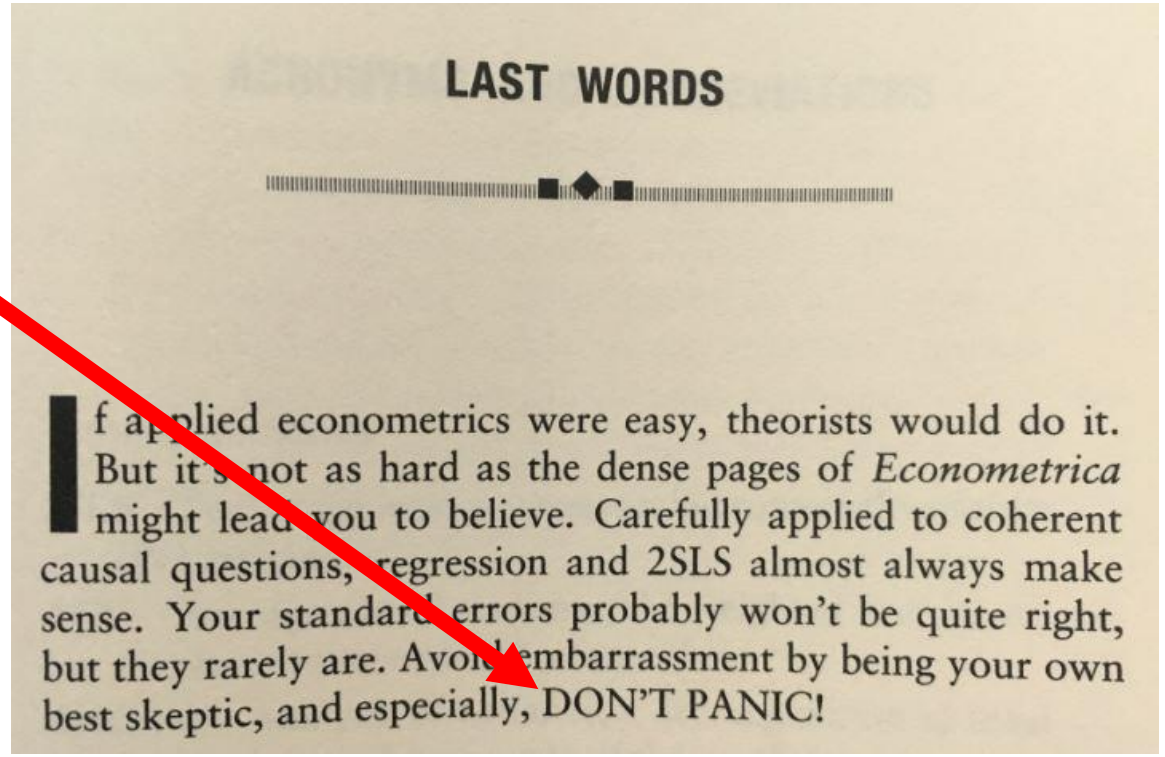


Any **problems**?

Challenges…

- coding loops in Matlab
- understanding exactly why Lasso is undertaken the way it is

Our work at the moment is characterized by the following words by Angrist and Pischke:



LAST WORDS

If applied econometrics were easy, theorists would do it. But it's not as hard as the dense pages of *Econometrica* might lead you to believe. Carefully applied to coherent causal questions, regression and 2SLS almost always make sense. Your standard errors probably won't be quite right, but they rarely are. Avoid embarrassment by being your own best skeptic, and especially, DON'T PANIC!

Returns to education in Denmark

A Bayesian Approach

Jonas & Nynne

Department of Economics

6 April, 2018

Motivation

Research Question:

What are the returns to education in Denmark?

The question has been heavily studied ever since it's popularization in the 1970's by Jacob Mincer. However, we consider the following questions to the standard formulation,

- Does prior knowledge about the returns to education and/or experience matter for the end results?
- Are the returns equal across the whole earnings distribution?

It is our intention to analyze whether returns differ across methods (standard vs. Bayesian) and distribution (high vs. low wages).

Progress

1) Data cleaning
2) Realized that data is too sparse to construct a social intergenerational mobility variable.
3) Successfully implemented a standard Gibbs sampler for the linear regression model using Jeffrey's prior.
4) Found literature on the implementation of Bayesian Quantile Methods Kozumi & Kobayshi (2011) and Yu & Moyeed (2001).

Preliminary results

Table: Returns to education (baseline = folkeskole)

	Level 2	Level 3	Level 4
OLS	$0,1641$	$0,3098$	$0,5901$
Bayesian	$0,1644$	$0,3099$	$0,5904$

Education as a continuous variable and Bayesian Quantiles

- *How do we turn our education variable from a categorical (i.e. Bachelor, Masters etc.) to a continuous variable (year of education)?*
  - Possible candidate: Method of Simulated Moments (MSM) using aggregate moments from Statistics Denmark?
- *How to code the Gibbs sampler for the Quantile regression following Kozumi Kobayashi (2011)?*
  - The data augmentation (z_i) step is causing problems. See below.
- Posteriors for the linear regression model with standard priors (normal and inverse-gamma) look similar, however the simulated data from the exponential distribution has a posterior distribution as,
  - $z_i|\boldsymbol{y}, \boldsymbol{\beta}_p, \sigma^2 \sim \mathcal{GIG}\left(\frac{1}{2}, \tilde{\delta}_i, \tilde{\gamma}_i\right)$, where $\tilde{\delta}_i^2 = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2/\tau^2\sigma^2$ and $\tilde{\gamma}_i^2 = 2/\sigma^2 + \theta^2/\tau^2\sigma^2$, with τ, θ being properties of a given quantile.

1. Mid-term meeting

Research question

- What is the effect of parental leave on wages?

Motivation

- Political interest in trying to get men to take more parental leave. But what happens if parental leave of men is increased? Or equivalently if parental leave of women is decreased?

Goal

- Extend the Roy model to a panel data setting and estimate the dynamics of wages in response to parental leave

1. Mid-term meeting

Status of the project

- Data cleaning in progress - works fine
- Problems with the Gibbs sampler

Sample $Y^* \sim p(Y^*|\beta, \Sigma, W, y)$
Sample $\beta \sim p(\beta|\Sigma, Y^*, W)$
Sample $\Sigma \sim p(\Sigma|\beta, Y^*, W)$

- Each step taken separately seems to work fine.
- But when putted together it doesn't work

1. Mid-term meeting

Feedback

- Any good advice on how to debug a Gibbs sampler?
- How to sample from a truncated normal distribution? - we have just found the program **rtnorm** (don't really understand it)

Estimating Child-Morbidity in Egypt by Use of Multinomial Logit Models

Maria Andreasen and Anna Granau

University of Copenhagen

April 6th, 2018

Project description

Replicate article which estimates child morbidity by the MNL:

$$P = (Y_{ijk} = k) = \frac{\exp(\eta_{ijk})}{1 + \sum_{l=1}^{k} \exp(\eta_{ijk})}, k = 0, 1, 2, ..., 7$$

where

$$\eta_{ijk} = z_{ij}\beta_k + f_k(\nu_{ij}) + S_{ik}$$

We want to extend model by random parameters model - including correlation between alternatives.

Status on the project

We have

- Cleaned data and recreated summary statistics
- Created a standard multinomial model with random effects, $\eta_{ijk} = z_{ij}\beta_k + S_{ik}$

We still need to

- Implement the P-splines, $f_k(\nu_{ij})$, either with a Bayesian or frequentist approach
- Extend the model to a random parameters model

Problems and Questions

- Difference between frequentistic and Bayesian results
- Are we allowed to use built-in-functions in R?

Determinants of educational choice

A micro econometric study of the individual's decision regarding field of study in a multinomial probit model using marginal data augmentation

Research question

- Investigate the influence of gender and the parental educational background on the individual's choice of field of study

Motivation

- More than 2/3 of graduates from humanities are women (2017)
- Less than 1/2 of science graduates are women (2017)
- Prev. results suggest that the educational level of the parents plays a significant role in determining educational level of the child

Status

- (A lot of) Literature search
- Access to data
- Begun data cleaning
- Begun programming of the algorithm in R
- Some derivations for the multinomial probit

Problems

- Derivations for the mixed probit, since practical identification poses a problem in the multinomial case
- Inclusion of field of study average wage as the alternative specific characteristic – reasonable?
- Sampling of the variance-covariance matrix – rejection sampling as proposed by Imai and van Dyk or Bartlett decomposition proposed by Nobile?

Project title (tentative): Returns to the same education: does gender matter?

Project description: We investigate the presence of a private sector gender wage gap among cand.polit graduates using a sample selection model on Danish register data from 1980-2010.

Motivation: The empirical literature shows that…

- The large convergence in the gender wage gap of the 1980s has slowed down. In a relative sense, especially among the highly educated *(Blau and Kahn, 2017).*

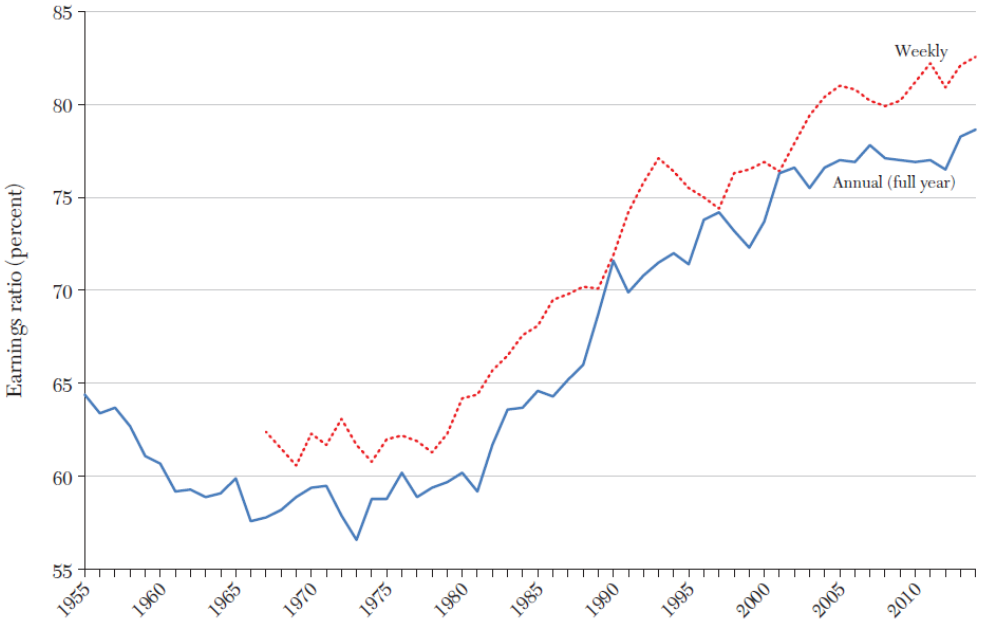


Figure 1. Female-to-Male Earnings Ratios of Full-Time Workers 1955–2014

Notes: Workers aged sixteen and over from 1979 onward, and fourteen and over prior to 1979.

Source: Fig. 7-2 "Evidence on Gender-Differences in Labor Market Outcomes," Francine D. Blau and Anne E. Winkler, *The Economics of Women, Men, and Work*, eighth edition. (New York: Oxford University Press 2018), p.173. By permission of Oxford University Press, USA.

- Childbirth disproportionally affects women's career path. Accumulation of human capital (*Gupta and Smith, 2002)*, wages *(Ejrnæs and Kunze, 2013)*, and sector choice *(Nielsen, Simonsen, and Verner, 2004)* is affected.

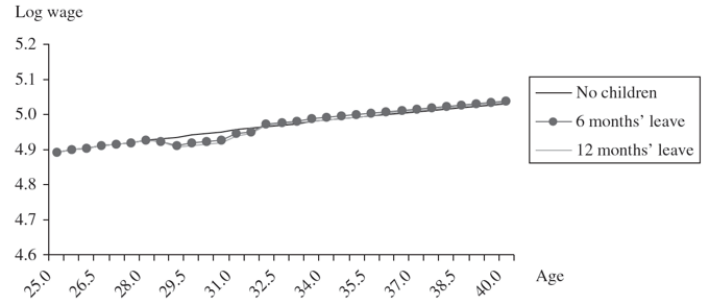


Fig. 2. Wage profiles, public sector (15–16 years of education)

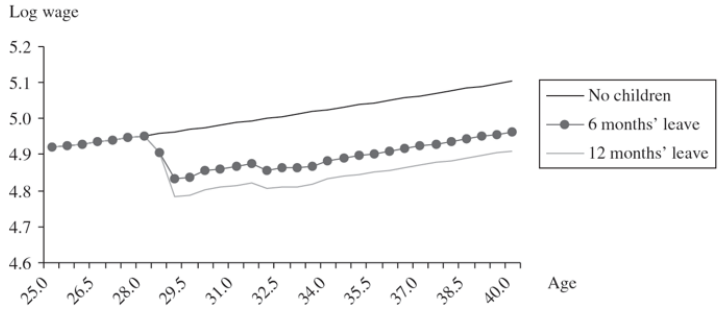


Fig. 3. Wage profiles, private sector (15–16 years of education)

Source: Nielsen, H. S., Simonsen, M. and Verner, M. (2004). *Scandinavian Journal of Economics* 106(4), 721-744

How we are going to do it:

Selection Models *(the parametric version)*

- **Idea** is to correct for a missing data problem (incidental truncation) arising from selection due to e.g. unobservable skills or preferences.
- **1st Estimation Approach:** joint ML estimation of selection and outcome equations. Fully efficient. Requires assumption of joint normality of error terms. Estimation of likelihood function.
- **2nd Estimation Approach:** Heckman two-step estimation. Loss of efficiency and need to correct standard errors, but requires weaker assumption on Ɛ. Estimation of first step probit and second step conditional truncated means.
- **Both Approaches:** Strong distributional asumptions that in theory ensure identification. In practice, issues of collinearity when same regressors included in selection and outcome equation → Exclusion restriction needed.

Our Model

- It is a Roy/Tobit Type 5/Endogenous Switching Model…

$$\ln w_{1i} = x_i\beta_1 + \epsilon_{1i}$$

$$\ln w_{2i} = x_i\beta_2 + \epsilon_{2i}$$

$$y_i^* = z_i\gamma + u_i$$

- Where y_i^* is our selection equation/a latent variable taking values according to:

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \quad \text{private sector employment} \\ 0 & \text{if} \quad y_i^* \leq 0 \quad \text{public sector employment} \end{cases}$$

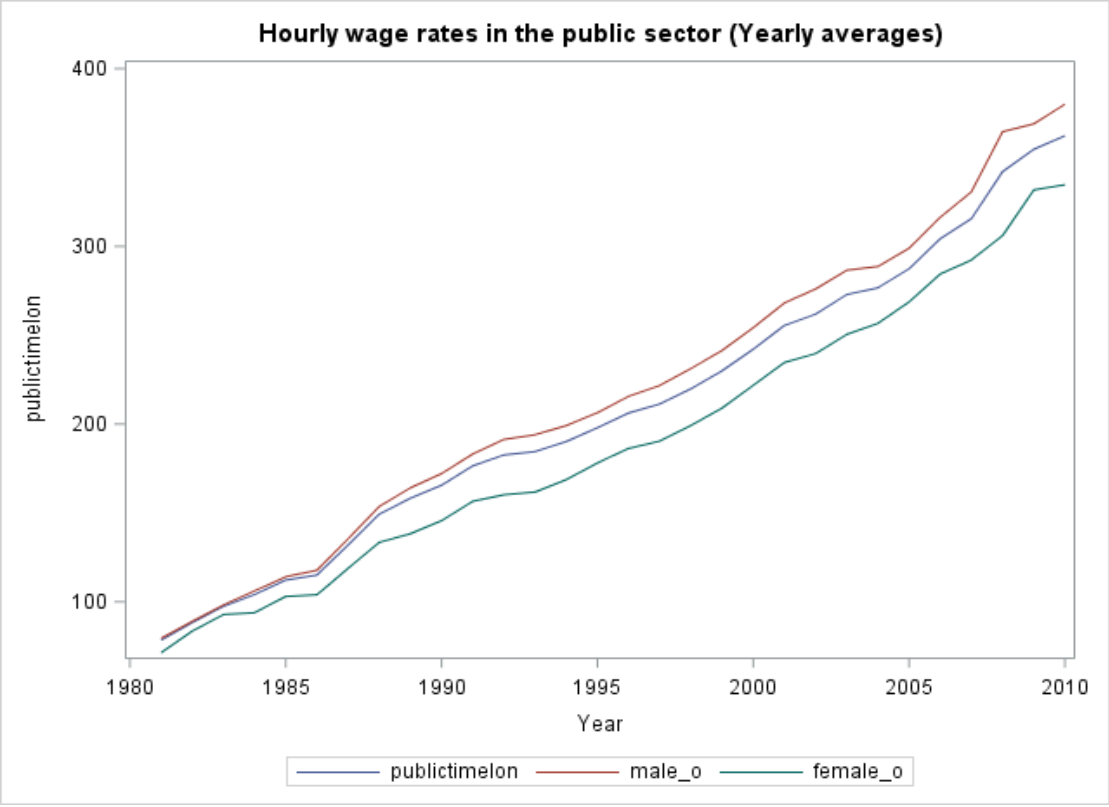
- And we assume our error terms to be trivariate normal with mean zero and covariance:

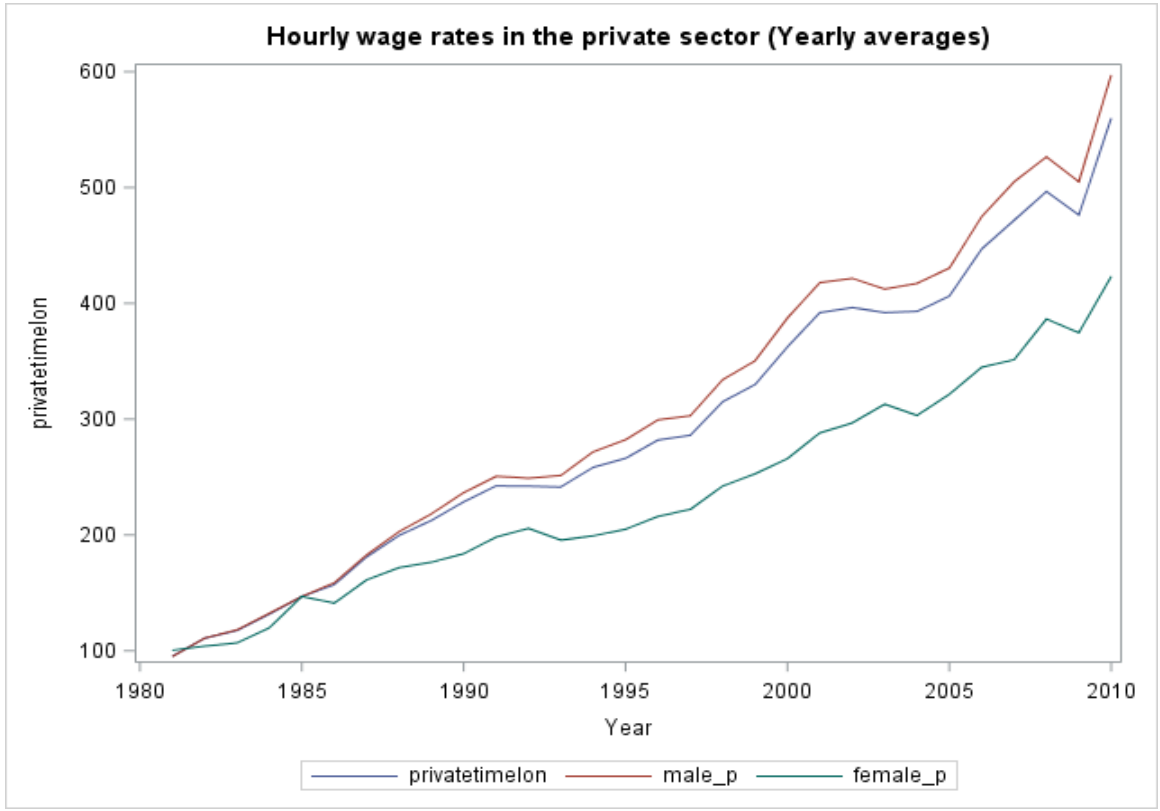
$$\text{Cov}\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ u_i \end{pmatrix} = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{1u} \\ \sigma_{12} & \sigma_2^2 & \sigma_{2u} \\ \sigma_{1u} & \sigma_{2u} & \sigma_u^2 \end{bmatrix}$$

Where we are at in the process:

- Our dataset is complete and we can show you the following:

	Female		Male	
	Private	Public	Private	Public
Share emp. in sector (pct.)	43.31	56.69	62.20	37.80
Mean wage (in DKK)	300.30	246.86	392.96	262.92
No. of individual-year observations	5290	6957	21573	13273





- Estimation of our model is going less smoothly…
- **Fragility of identification is also true in our case:** *MC simulation results of our model gives us a singular Hessian. This is where we are at the moment, computationally.*

The cross-nested logit model

Estimating partially nested structures

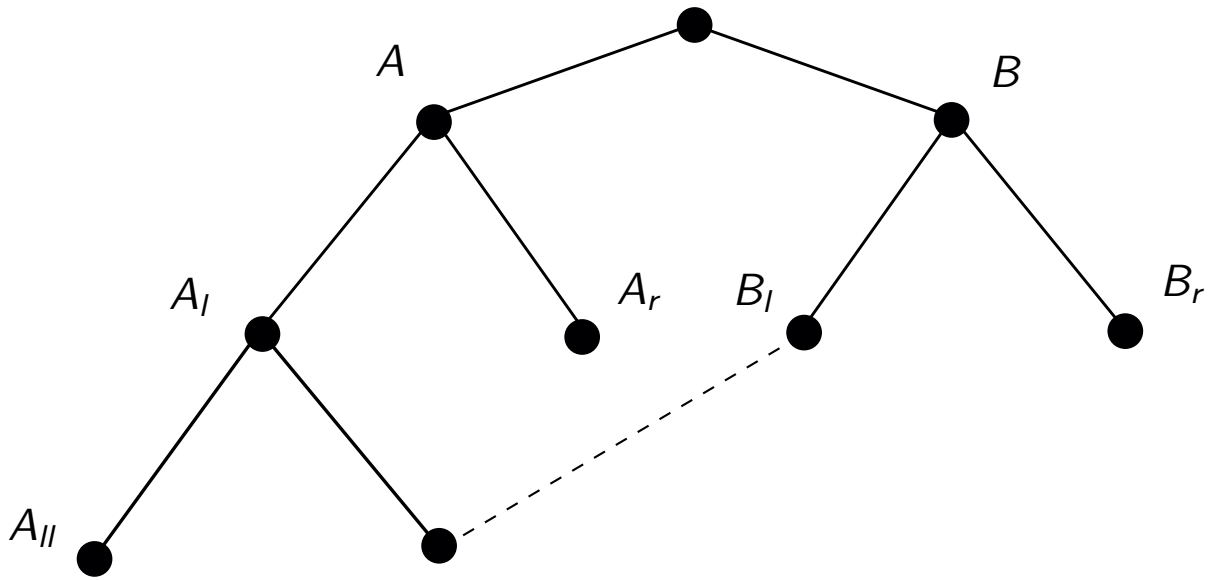
Thor Donsby Noe & Kristian Urup Larsen

April 5, 2018

Department of Economics, University of Copenhagen

Research question

- Show how the cross-nested logit model can extend the concepts of nested choices to a range of complex choice puzzles.
- Implement an estimator for the cross-nested logit on synthetic and real data (for the Danish unemployment benefits systems).



Motivation:

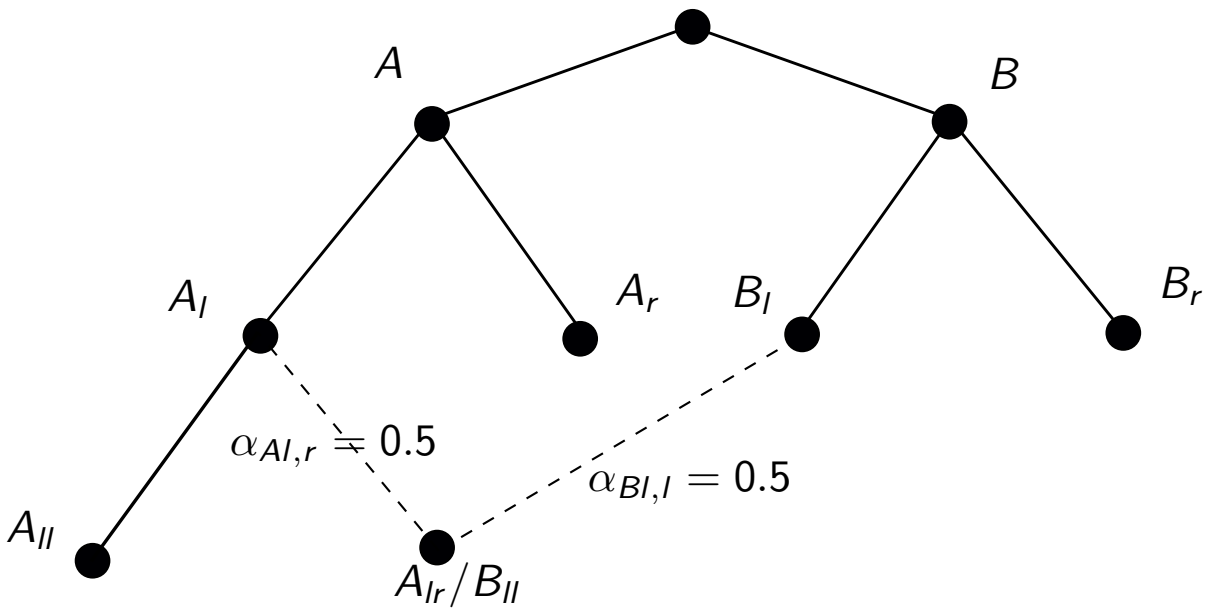
- To loosen the restriction that each option is only accessible through one tree-path, but not duplicated.

The cross-nested logit model

The CNL models are simply a generalization of the Multinomial Logit model, within the GEV (Generalized Extreme Value) class where

$$G(x_1, ..., x_J) = \sum_m \left(\sum_{j \in \mathcal{J}} \alpha_{jm} x_j^{\mu_m} \right)^{\frac{\mu}{\mu_m}} \tag{1}$$

Where m is a nest index, α_{jm} gives how much choice j is in nest m and is restricted to $\sum_m \alpha_{jm} > 0 \ \forall j$. It is somewhat common to require $\sum_m \alpha_{jm} = 1$.



So what's next?

- Simulate the DGP described by the CNL - we want a visual understanding of the data. Also sucessfully code the estimator.
- It's typically assumed that α's are a priori known (?) to keep the number of parameters down $\rightarrow$ can parameter tuning give optimal α's?
- The estimation requires either heavy computer power or analytically derived derivatives from complex functions (although they do exits!) $\rightarrow$ how sensitive is numerical optimization in this setting?
- Use CNL to estimate choice probabilities for those on sick leave within the *unemployment benefits system* using Danish registry data from the DREAM group.

Random coefficient logit estimation of PC game demand within the Steam community – Bayesian approach

A replication of the paper:

Jiang, R., Manchanda, P., & Rossi, P. (2009). Bayesian analysis of random coefficient logit models using aggregate data. Journal of Econometrics, 149(2), 136-148.

Pápai Dénes

2018. április 5.

Research question

- Something like this: What is the expected (steam) demand of a newly released PC game in a given genre?
- Goal:
  - Using random coefficients logit to fit an aggregate demand model, acquiring price elasticities

$$U_{ijt} = f\left(X_{jt}\middle|\theta^i\right) + \eta_{jt} + \varepsilon_{ijt} = X_{jt}\theta^i + \eta_{jt} + \varepsilon_{ijt} \tag{1}$$

$$\begin{aligned} s_{jt} &= \int s_{ijt}\phi\left(\theta^i \mid \bar{\theta}, \Sigma\right) d\theta^i \\ &= \int \frac{\exp\left(X_{jt}\theta^i + \eta_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(X_{kt}\theta^i + \eta_{kt}\right)} \phi\left(\theta^i \mid \bar{\theta}, \Sigma\right) d\theta^i \end{aligned} \tag{2}$$

Status

- Data on approximately 18 000 Steam apps
  - *appName, developer, playtime, owners, players, price, critics score* (Steamspy)
  - System requirements of the apps (Steam, only Windows) – not structured properly yet
    - *RAM, Processor, Graphics, OS, Network, DirectX, Storage, Sound*
  - Genre and game category data – easy enough to get, the problem is the proper classification of the games

- Implementation
  - Started to reproduce a simulated example of the paper to get started, only generated the data so far

Problems

- With the data
  - Will lose some observations while trying to merge system requirements with the other part of the data
  - have to have an assumption of the size of the whole market to define share of the outside good
  - Choosing share variable: *playtime, players, or owners*?
  - The variation of the prices could be better - an alternative is to separate markets by region and not by time (only in the US.)
- Implementation
  - Time to practice Bayesian econometrics
  - Must be manageable if I follow the authors

Research question: Can endogenous persistence of volatility capture underlying dynamics in stock market time-series?

- A Stochastic Volatility (SV) model with jumps is given as

$$R_t \sim N(J_t Z_t^s, \exp(\lambda_t))$$

$$\lambda_t \sim N(\mu + \phi\lambda_{t-1}, \sigma_v^2)$$

$$Z_t^s \sim N(\mu_s, \sigma_s^2)$$

$$J_t \sim Bern(\delta),$$

- I then change the model by making the persistence of the volatility dependent on the jump variable J

$$\lambda_t \sim N(\mu + (\phi + \rho_\lambda J_{t-1})\lambda_{t-1}, \sigma_v^2)$$

- Motivation: Any SV Model is closely related to models that are used to price financial assets and any significant modification of the model is therefore of high interest.

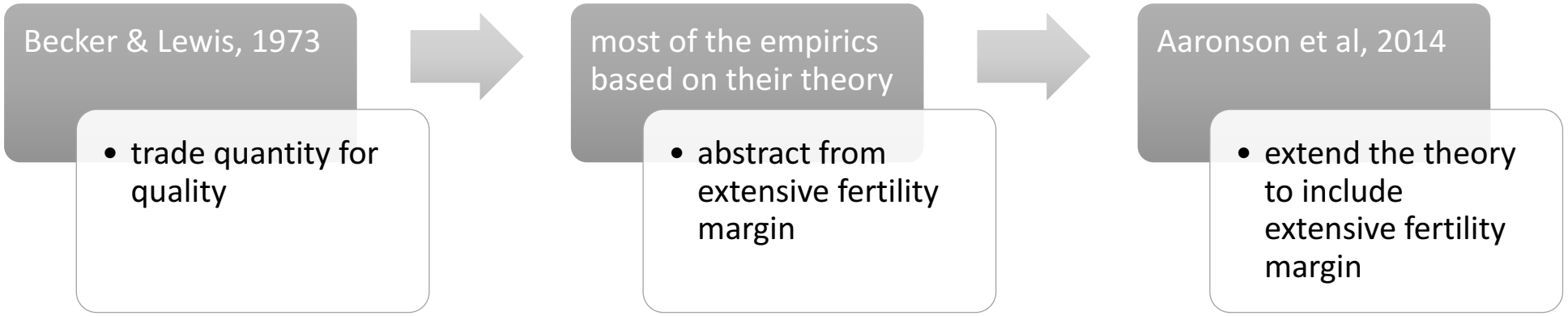
Project Status: Stuck

- I have calculated all the posterior distributions of the parameters and all latent variables. However, there might be a possible identification problem or I have miscalculated the posterior of the Jump variable distribution.

Title:

Effect of education cost on Danish **extensive** and **intensive** fertility margin

Motivation:



Apply to Danish setting

- Childlessness rate around 15% in 1955-1959 birth cohort
- Variation of day care facility using payment in municipal level
- “There is considerable variation in how the different municipalities tackle this issue, which means that the women’s situation depends on where they live” (The Danish Equal Status Council, 1997)

Empirical method

$$FertilityIndicator_i = \beta_0 + \beta_1 EducationCost_i + \beta X_i + \varepsilon_i$$

Data

Registry data from Statistics Denmark

- Panal data 1980-2010, only use 2007-2010
- Individual level

Independent variable:

"ESTIMATED CHILD CARE EXPENSE, region", 0-5years

- Rate of children enrolled in different day care facilities
- Yearly payment of using different day care facilities
- Aggregated to regional level due to data limitation

Other control variables

Mother's age, education level, work experience…

Dependent variable:

1. Extensive fertility
   - Binary outcome encoded as 0 and 1
   - Probit model
   - Preliminary result: significantly negative effect of estimated child care expense
2. Intensive fertility
   - Count data 1, 2, 3, 4, 5…
   - Poisson model
   - Preliminary result: insignificantly negative effect of estimated child care expense
   - (undone) Test for overdispersion problem, auxiliary derivation
3. Total fertility rate
   - Count data 0, 1, 2, 3, 4, 5… with observations of 0 child taking around 40%
   - Negative binomial 2 (variance is qudratic in mean: $w_i = \mu_i + \alpha\mu_i^2$)
   - Preliminary resuit: problem, standard deviation with complex numbers

Analyzing the impact of an increase in the maternity leave duration on labor outcomes in the UK using a semiparametric approach

BORJA ÁLVAREZ

Research question and motivation

Research question: Did an increase in the maternity leave duration lead to a worsening of the market outcomes of females relative to the men in the UK in 2007? And did this effect vary across income quantiles?

Motivation:

-Do an estimation avoiding basic models.

-Increase my knowledge on the analytical part in this field within public policy.

-Broadly known that an increase in maternity leave (relative to men's) increase gender pay gap and other labor outcomes. But does it have a greater impact depending on socieconomic status?

-Observe the impacts of this atypical measure.

Status of my project

Dataset: Six .tab datasets obtained from the UK Data Service. Each has the same structure. Cleaned and ready for the estimation in Matlab.

Estimation model: Chosen, subject to slight modifications when the estimation code is developed.

-I will opt to use a mixed method within Semiparametric framework, combining weighting and regression on the propensity score to estimate the average treatment effects.

-To estimate the effects depending on the socioeconomic status I will employ quartile treatment effects (QTE).

Code: So far just cleaning and loding.

Problems? Feedback?

Probems I have overcomed: I got stuck for some days with data cleaning. I used different softwares trying to solve it. The solution was do part of the job manually.

Problems I haven't overcomed yet: None

Feedback: Any experience going beyond average treatment effects? Analyzing some data stressing on quantile effects?