Gender pay gap: analysis 2021

May, 2021, Statistics Iceland

Content

- Formulation of the problem and goal of the analysis
- Solution
- Implementation
- What is new
- Examples

Formulation of the problem / Goal of the analysis

To calculate

- the *estimated difference* in hourly wages between gender groups
  - as mean and/or other distributional aspects
  - when accounting for
    - known characteristics (demographic, social, economic) of individuals and their environment
    - participation effects (employment related)
- the *uncertainty* of this estimate
  - due to model parameters
  - due to the choice of model
- their *time evolution*

Solution

MLM (linear or generalised, with discrete or continuous variables)

May include cross-correlations, grouping dependencies, dynamical aspects (e.g. auto-correlations)

Most general formulation: $y = F(t, x, z, ...)$ and $y \sim \mathcal{P}$ or in *levels*, e.g.

$$y = F(t|A, B) + e$$
$$A = f_A(x, z|a^1)$$
$$B = f_B(x, z|b^1)$$

$A^1 = ... + a^2$ and $B^1 = ...$, where $e, a^1, a^2, ...$ are distributed according to:

- (multi-) variate Normal (when frequentist) distributions
- (multi-) variate Prior (when Bayesian) distributions

Example: simplest model

Each subject is observed many times.

The response (y) of each subject is a linear function of time (at time points i).

The parameters (intercept and slope) of these functions have a normal distribution with higher level parameters $\mu_\alpha, \mu_\beta, ...$

$$y_i \sim N\left(\alpha_{j[i]} + \beta_{1j[i]}(\text{time}), \sigma^2\right)$$

$$\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\alpha_j} \\ \mu_{\beta_{1j}} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\alpha_j} & \rho_{\alpha_j\beta_{1j}} \\ \rho_{\beta_{1j}\alpha_j} & \sigma^2_{\beta_{1j}} \end{pmatrix}\right), \text{ for Subject j} = 1, \ldots, \text{J}$$

Implementation

The R-implementation of these models

$$(g)lmer \quad and \quad stan\_lmer \quad or \quad brms$$

$$qmlm \quad and \quad bma$$

What is new

- MLM: more general models
- Bayesian MLM
- Bayesian model averaging (similar to ensemble *ML*)
- May include: more details of differences in distributions plus participation effects

What is new: why more general

Note on time-varying and time-constant predictors, while groups/clusters present

$$y \sim t + x_{tc} + x_{tv}^{B} + x_{tv}^{W} + z$$

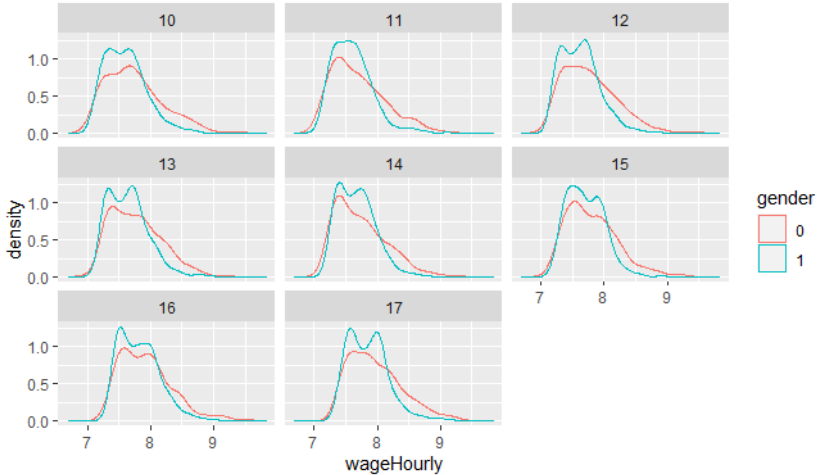
$$+ interact(t, x, z) + (1 + t|g) + (1 + x_{tv}^{W}|g)$$

What is new: why Bayesian

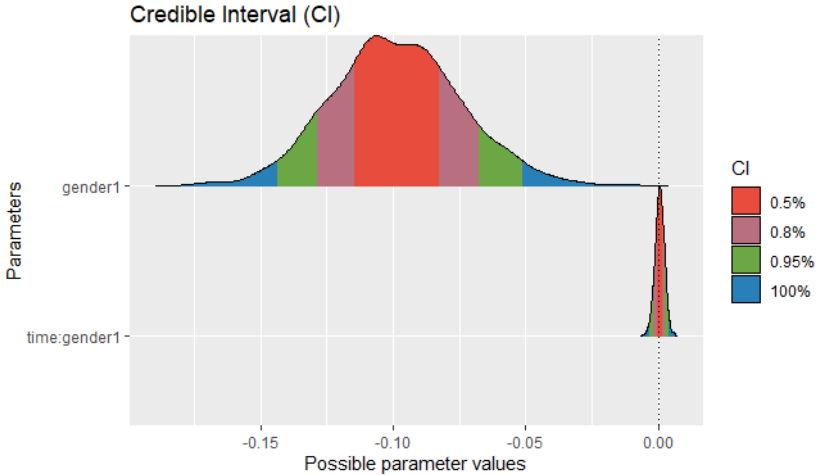
The advantages of using Bayesian approach

- interpretation of results (while in frequentist approach is difficult, unless the model is linear, with no inverse link function and no interaction terms OR! unless we do *simulations*):
  - by inspecting the posterior distribution at different levels of predictors
  - being able to make probabilistic statements about a scientific hypothesis
- combining *all possible model*, according to:
  - posterior probability of models, given the data and
  - posterior probability of parameters, given all models and data, which gives
  - posterior mean and standard deviation of parameter of interest <-> point estimate and uncertainty

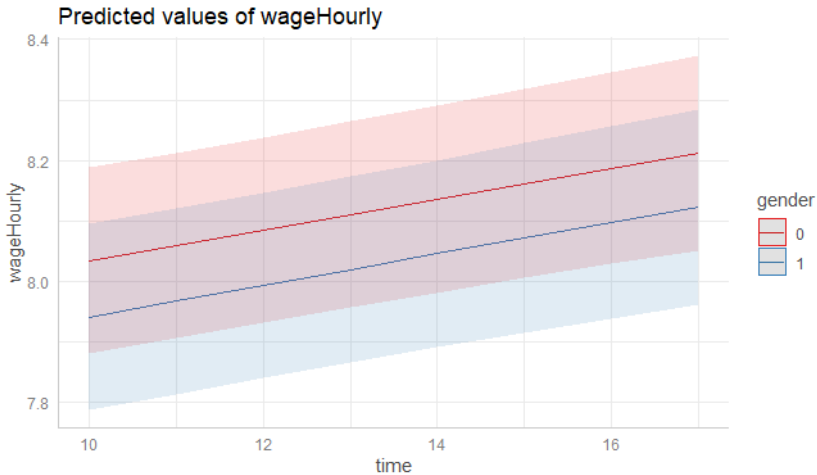
Examples, density distributions through time



Effect of gender



Adjusting (log) wages for all factors except gender



Transparent!

https://github.com/violetacln/GIW

Thank you!