Bayesian Statistics | Project Presentation

ANALYSING THE IMPACT OF UNIONISM OVER THE WAGE

Presenter

Eduardo Gonnelli

Professors

Francesco Pauli

Gioia Di Credico

Leonardo Egidi

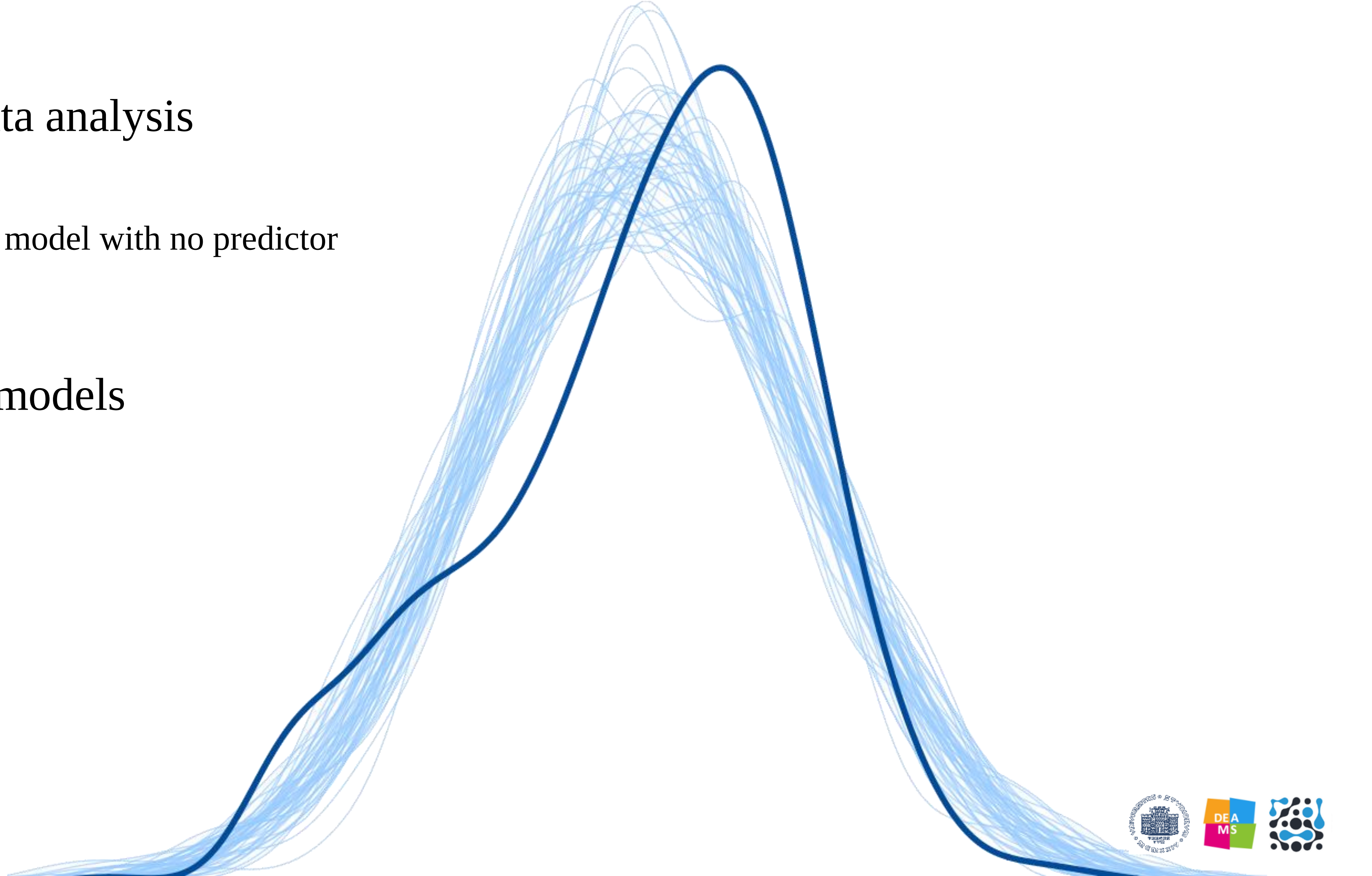
University of Trieste

Data Science and Scientific Computing

2020

SUMMARY

1. Introduction
2. Exploratory data analysis
3. Models
   - 3.1 Varying intercept model with no predictor
   - 3.2 Varying intercept
   - 3.3 Varying slope
4. Evaluating the models
5. Discussion
6. Conclusions



1 - INTRODUCTION

- The aim of this project is to build a model for studying the effects of covariates on wage with Rstan package.

- The dataset: Males.

- It is available in R-project by the Econometrics package called Ecdat.

- National Longitudinal Survey (NLS Youth Sample).

- Published by Vella, F. and M. Verbeek. - *Journal of Applied Econometrics*.

Details of the dataset:

- a panel of 545 observations from 1980 to 1987.
- *number of observations* : 4360.
- *observation* : individuals.
- *country* : United States.

1 - INTRODUCTION

- The analysis took into account the hierarchical structure of the data.

- Give a special attention to the relation between **unionism** and **wage**.

- Implement the effect of time in the Model.

- Check the model fit by using the proper tools of posterior predictive checking

The Bayesian approach was employed to analyse the data.

All the visualization steps of the Bayesian approach was obtained by the bayesplot (Bayesian workflow).

- Structure of the data

```
data(Males)
str(Males)
```

```
## 'data.frame':    4360 obs. of  12 variables:
##  $ nr        : int  13 13 13 13 13 13 13 13 17 17 ...
##  $ year      : int  1980 1981 1982 1983 1984 1985 1986 1987 1980 1981 ...
##  $ school    : int  14 14 14 14 14 14 14 14 13 13 ...
##  $ exper     : int  1 2 3 4 5 6 7 8 4 5 ...
##  $ union     : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
##  $ ethn      : Factor w/ 3 levels "other","black",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ maried    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ health    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ wage      : num  1.2 1.85 1.34 1.43 1.57 ...
##  $ industry  : Factor w/ 12 levels "Agricultural",..: 7 8 7 7 8 7 7 7 4 4 ...
##  $ occupation: Factor w/ 9 levels "Professional, Technical_and_kindred",..: 9 9 9 9 5 2 2 2 2 2 ...
##  $ residence : Factor w/ 4 levels "rural_area","north_east",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
Males$nr <- as.factor(Males$nr)
Males$year <- as.factor(Males$year)
Males$school <- as.factor(Males$school)
Males$exper <- as.factor(Males$exper)
```

2 – EXPLORATORY DATA ANALYSIS

- The analysis of the unionism behaviour was made for the employee number 166.

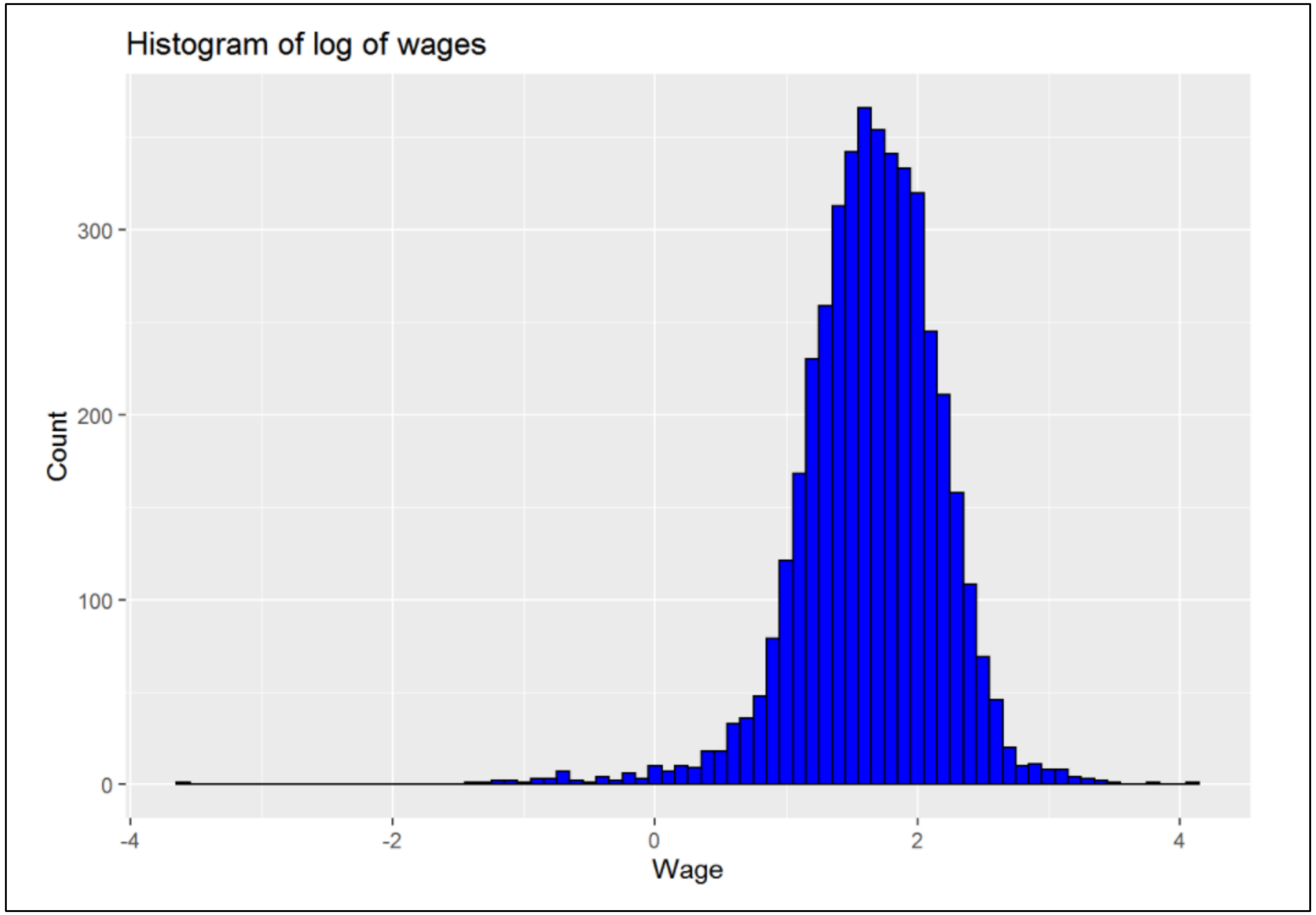
- After 1984 he left the union

```
Males %>%
  filter(nr == "166")
```

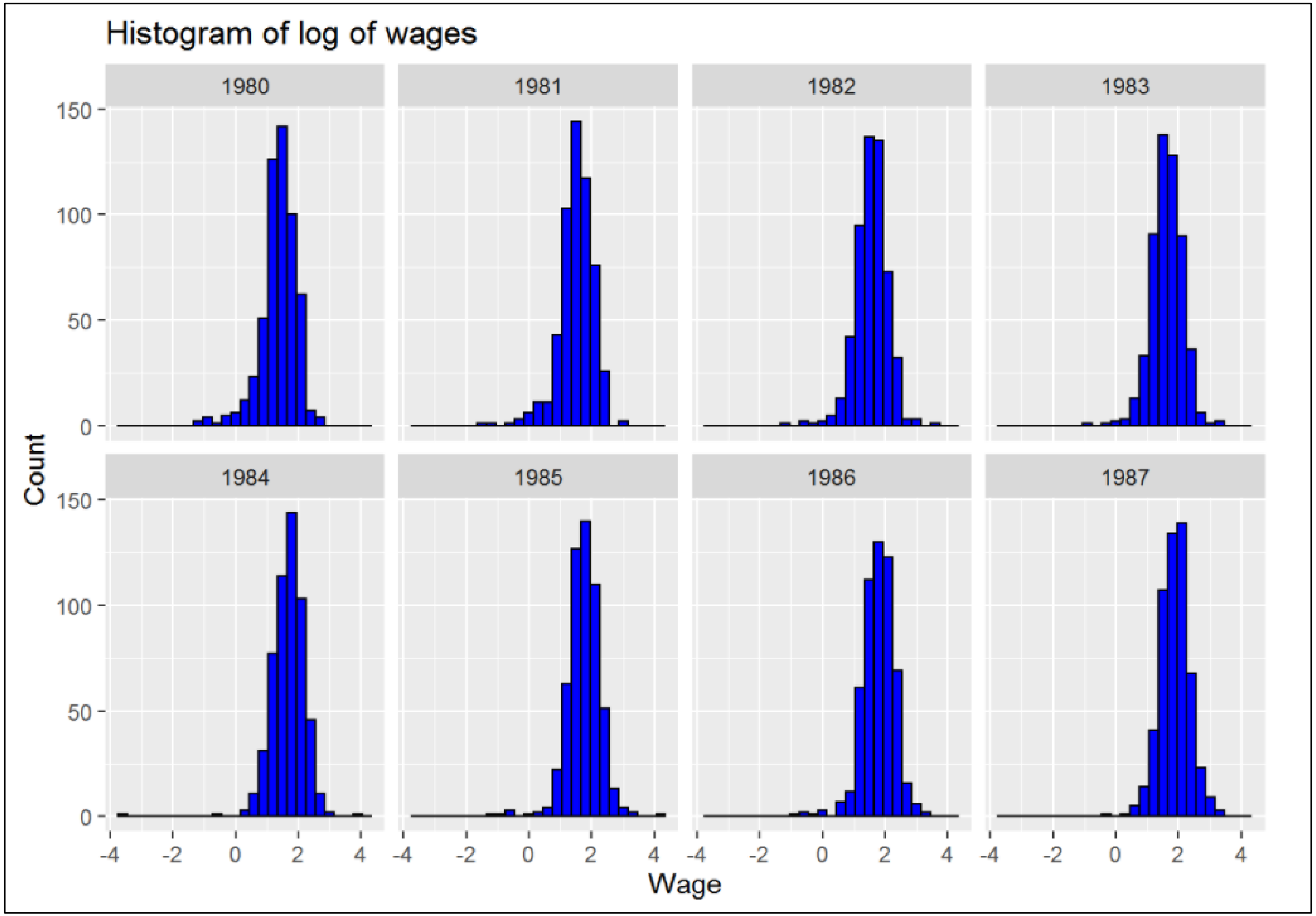
```
##     nr year school exper union  ethn maried health      wage industry
## 73 166 1980     10     2   yes other     no     no 1.0052940    Trade
## 74 166 1981     10     3   yes other     no     no 0.9376874    Trade
## 75 166 1982     10     4   yes other     no     no 1.3674877    Trade
## 76 166 1983     10     5   yes other     no     no 1.3780088    Trade
## 77 166 1984     10     6   yes other     no     no 1.1701835    Trade
## 78 166 1985     10     7    no other     no     no 1.7282214    Trade
## 79 166 1986     10     8    no other    yes     no 1.7098647    Trade
## 80 166 1987     10     9    no other    yes     no 1.8510748    Trade
##                           occupation       residence
## 73                Laborers_and_farmers      north_east
## 74                Laborers_and_farmers      north_east
## 75                Laborers_and_farmers      north_east
## 76                Clerical_and_kindred      north_east
## 77                Clerical_and_kindred      north_east
## 78        Craftsmen, Foremen_and_kindred      north_east
## 79 Managers, Officials_and_Proprietors      north_east
## 80        Craftsmen, Foremen_and_kindred nothern_central
```

2 – EXPLORATORY DATA ANALYSIS

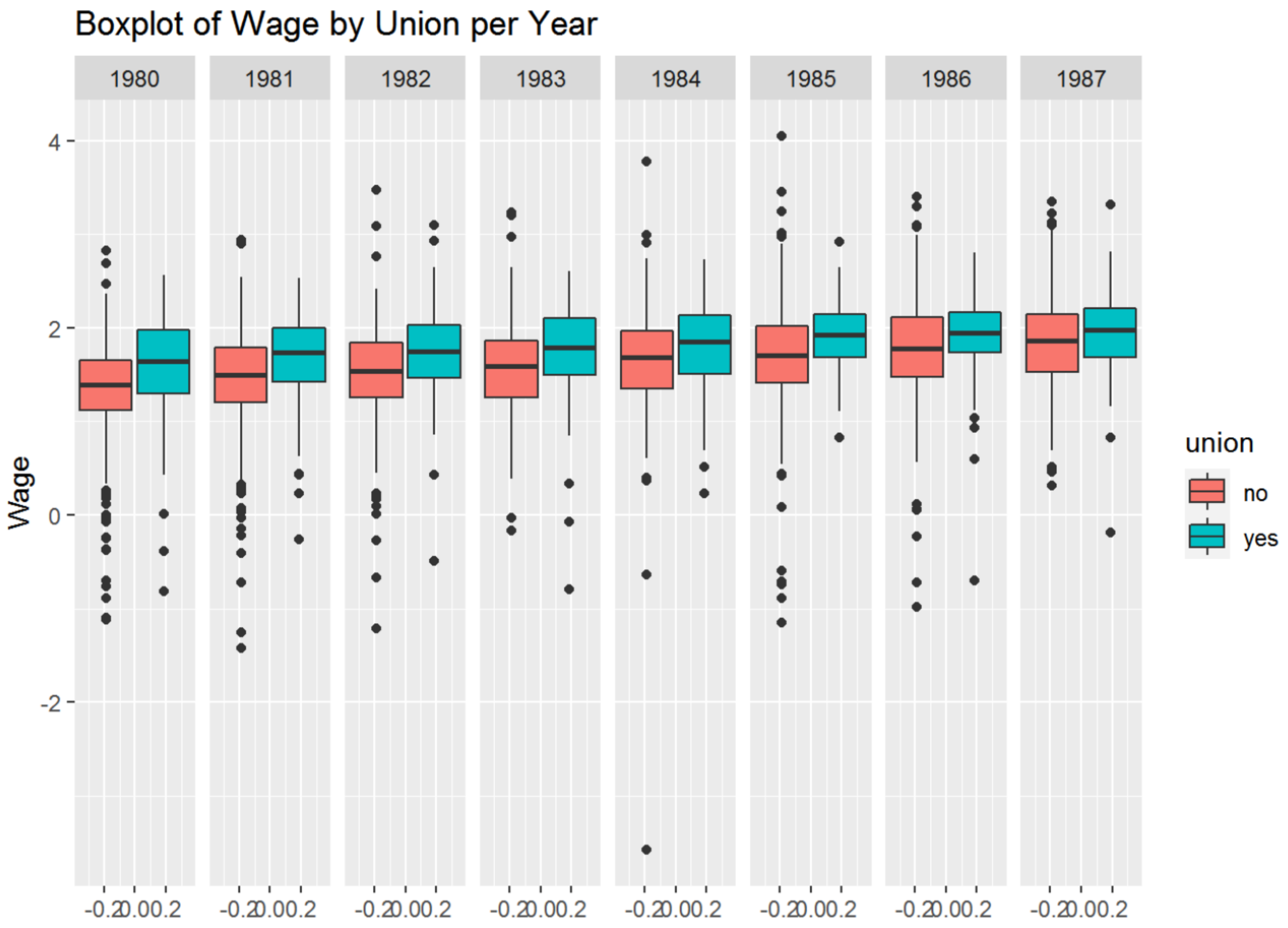
- Wage distribution show that the wage follows the normal distribution.



- Wage distribution and grouping by years



2 – EXPLORATORY DATA ANALYSIS



2 – EXPLORATORY DATA ANALYSIS

- Year with more employees in the union was 1987

```
a = round(100*(sum(union.by.year$UnionYes)/(sum(union.by.year$UnionYes)+sum(union.by.year$UnionNo))),2)
b = '\U0025'
paste("For all years the percentage of employees affiliated to union is:", a,b)
```

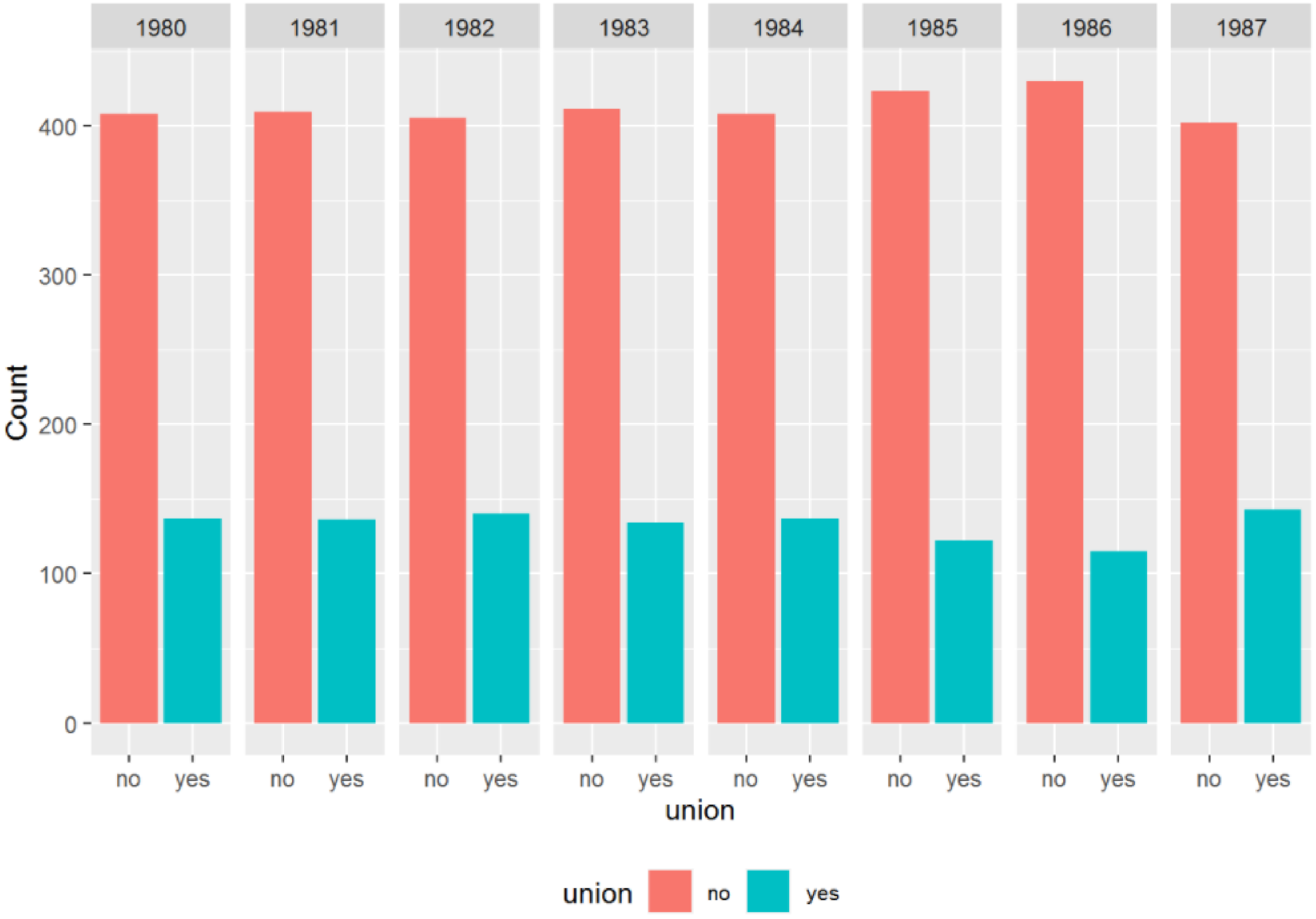
```
## [1] "For all years the percentage of employees affiliated to union is: 24.4 %"
```

The following analysis shows that there were more employees on the union in 1987.

```
a1 = round(100*(union.by.year$UnionYes/(union.by.year$UnionNo+union.by.year$UnionYes)),2)
names_years <- c('1980', '1981', '1982', '1983', '1984', '1985', '1986', '1987')
paste(names_years, ":", a1, b)
```

```
## [1] "1980 : 25.14 %" "1981 : 24.95 %" "1982 : 25.69 %" "1983 : 24.59 %"
## [5] "1984 : 25.14 %" "1985 : 22.39 %" "1986 : 21.1 %"  "1987 : 26.24 %"
```

2 – EXPLORATORY DATA ANALYSIS

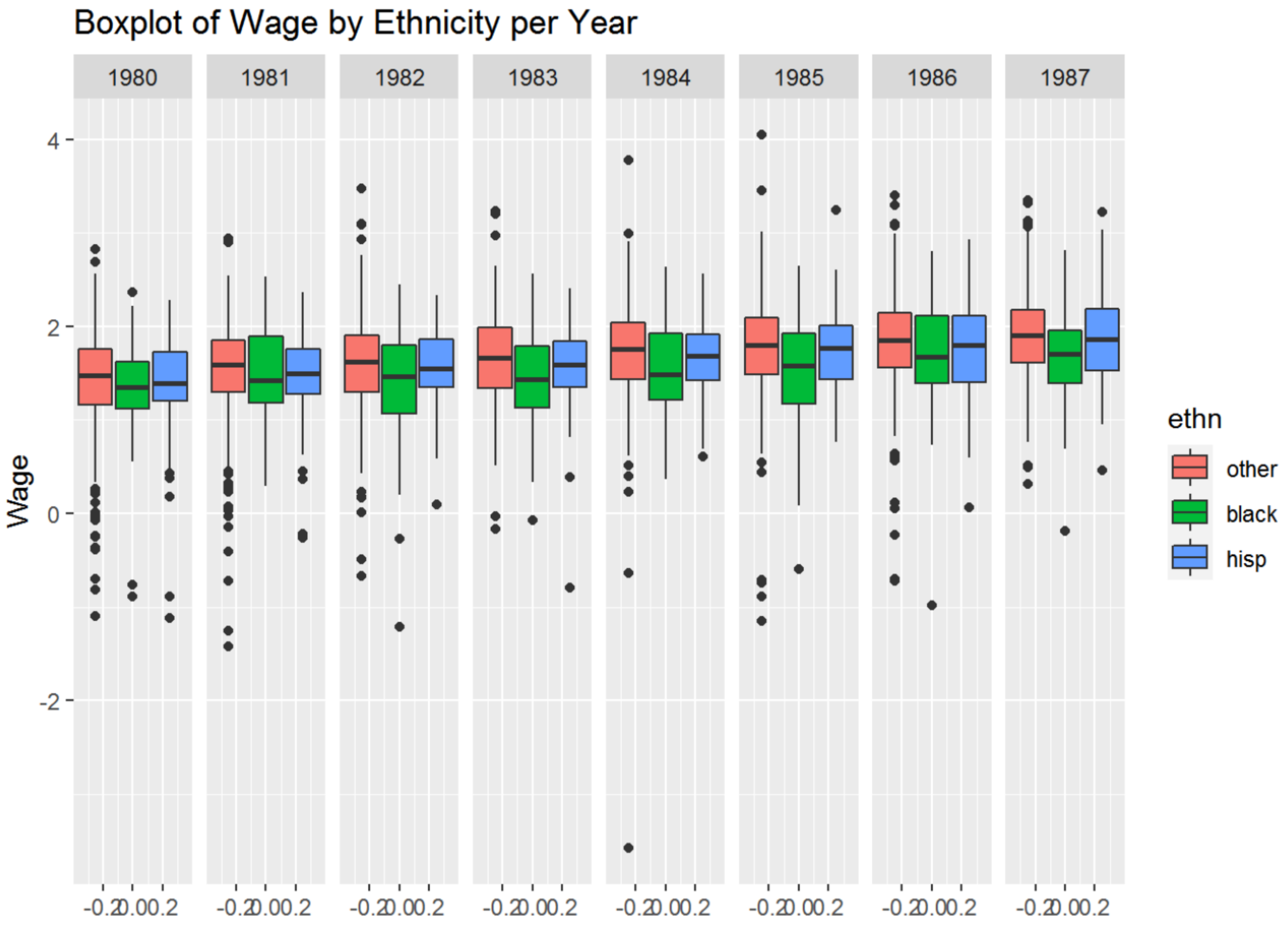


2 – EXPLORATORY DATA ANALYSIS

- Analysis of wage and ethnicity

People from Black and Hispanic ethnicities have lower wage income.

This covariate was considered in the analysis because it is an individual characteristic and does not be modified through the years.



3 – MODELLING

- The modelling of the problem take into account the hierarchical structure of the data.

- The variables **union** and **ethnicity** were selected as fixed-effect variables.

- The variable **years** was selected as random-effect variable.

- Wage is the response variable.

- The modelling process is twofold:
  1. The Intraclass correlation coefficient (ICC) was calculated for the years group.
  2. It was built the varying-intercept and varying-slope model.

- The **loo** function was employed to select the best model.

3. Models

3.1 Varying intercept model with no predictor

3.2 Varying intercept

3.3 Varying slope

$$y_i = \alpha_{j[i]} + \varepsilon_i$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

$$\alpha_{j,[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

3 – MODELLING

3.1 Varying intercept model with no predictor

Stan

```
data {
  int<lower=0> N; // number of obs
  int<lower=0> J; // Number of years
  vector[N] y; // outcome data (log wage)
  int year[N]; //year ID variable
}

parameters {
  real a[J]; // year j intercept
  real mu_a; // prior on alpha
  real<lower=0> sigma_y; // prior sigma of y in year j
  real<lower=0> sigma_a; // hyperparameter sigma of alpha
}

model {
  a ~ normal(mu_a, sigma_a);
  for (n in 1:N)
    y[n] ~ normal(a[year[n]], sigma_y);
}
```

3.1 Varying intercept model with no predictor

```
Males$union <- as.numeric(Males$union == "yes")

## arrange data into a list

year_random_effect_stan <- list(
  N = nrow(Males),
  J = length(unique(Males$year)),
  y = Males$wage,
  year = as.numeric(Males$year)
)

str(year_random_effect_stan)
```

```
## List of 4
##  $ N   : int 4360
##  $ J   : int 8
##  $ y   : num [1:4360] 1.2 1.85 1.34 1.43 1.57 ...
##  $ year: num [1:4360] 1 2 3 4 5 6 7 8 1 2 ...
```

```
model.no.predictor.stan <- stan(file='year_random_effect_stan.stan',
                                      data=year_random_effect_stan,
                                      iter=1000,
                                      chains=4)
```

3.1 Varying intercept model with no predictor

```
#RStan output
print(model.no.predictor.stan, pars = c('sigma_a','sigma_y', 'mu_a'))
```

```
## Inference for Stan model: 01_year_random_effect_stan.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##         mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## sigma_a 0.19       0 0.08 0.11 0.15 0.18 0.22  0.39  1376    1
## sigma_y 0.51       0 0.01 0.50 0.51 0.51 0.52  0.52  2976    1
## mu_a    1.65       0 0.07 1.51 1.61 1.65 1.69  1.79  1760    1
##
## Samples were drawn using NUTS(diag_e) at Mon Sep 07 00:02:03 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Intercept (mean) is 1.65 and that the credible interval ranges from 1.51 to 1.79.

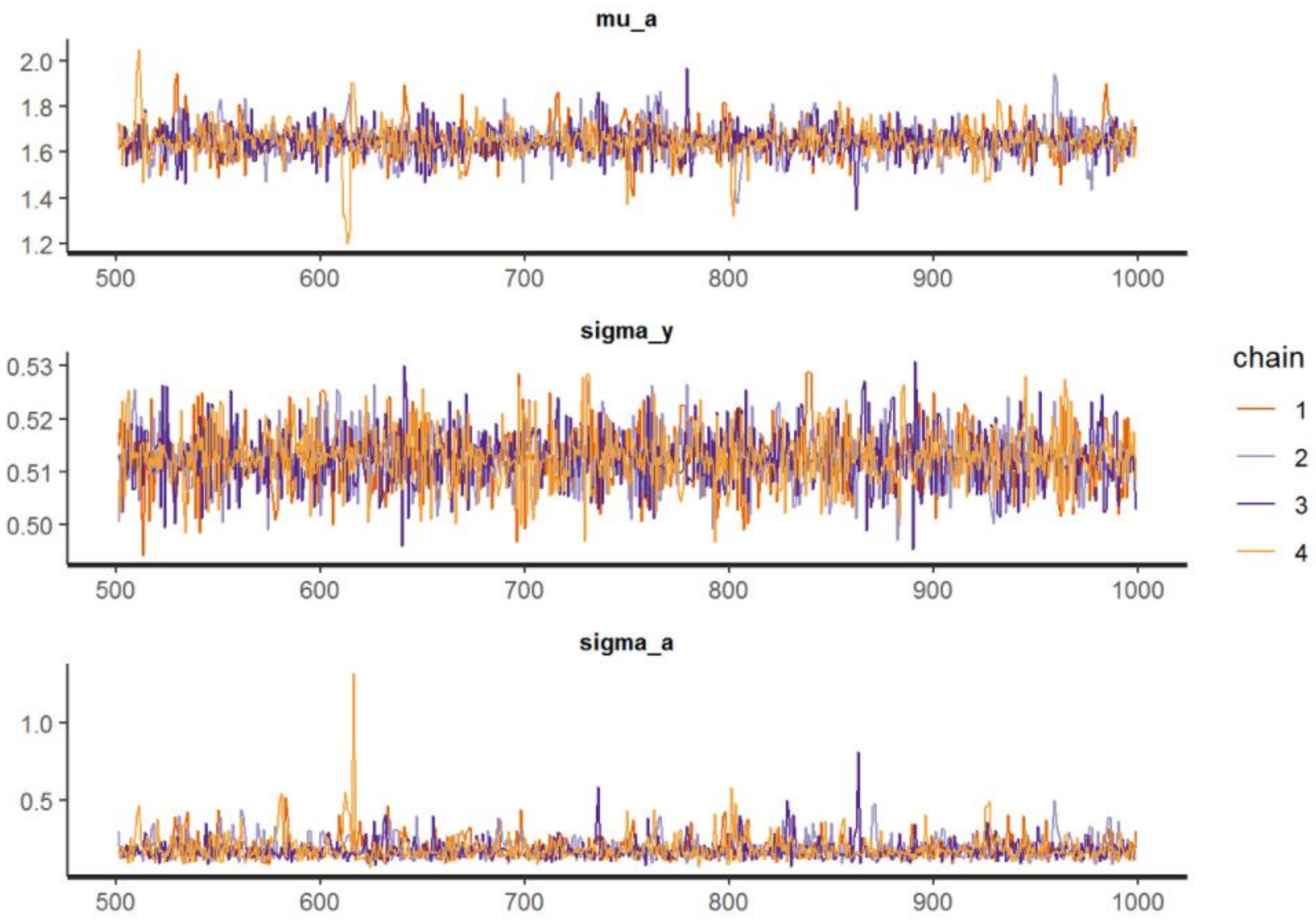
The Intraclass correlation coefficient (ICC): $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$ $\frac{0.19^2}{0.19^2 + 0.51^2} = 0.13$

So years explain 13% of the raw variation in this dataset.

In other words, it means that a multilevel model is warranted.

3.1 Varying intercept model with no predictor

```
traceplot(model.no.predictor.stan, pars= c("mu_a","sigma_y","sigma_a"), nrow = 3)
```



3. Models

3.1 Varying intercept model with no predictor

3.2 Varying intercept

3.3 Varying slope

$$y_i = \alpha_{j[i]} + \beta_{union} * x_{union} + \beta_{black} * x_{black} + \beta_{hisp} * x_{hisp} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_y^2)$$

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

3.2 Varying intercept

Stan

```
data {
  int<lower=1> N; // number of obs
  int<lower=0> J; // Number of years
  vector[N] y; // outcome data (log wage)
  int<lower=0,upper=1> x_union[N]; // Input data (union)
  int<lower=0,upper=1> x_black[N]; // Input data (black ethn)
  int<lower=0,upper=1> x_hisp[N];  // Input data (hisp ethn)
  int year[N]; //year ID variable
}
parameters {
  real a[J]; // year j intercept
  real b_union_yes; // beta slope for union data
  real b_hisp; // beta slope for hisp data
  real b_black; // beta slope for black data
  real mu_a; // prior on alpha
  real<lower=0> sigma_y; // prior sigma of y in year j
  real<lower=0> sigma_a; // hyperparameter sigma of alpha
}
model {
  a ~ normal(mu_a, sigma_a);
  for (n in 1:N)
    y[n] ~ normal(a[year[n]] + b_union_yes * x_union[n] + b_black * x_black[n] + b_hisp * x_hisp[n], sigma_y);
}
generated quantities {
 vector[N] log_lik;
 vector[N] y_hat;
  for (i in 1:N){
    y_hat[i] = normal_rng(a[year[i]] + b_union_yes * x_union[i] + b_black * x_black[i] + b_hisp * x_hisp[i],
sigma_y);
    log_lik[i] = normal_lpdf(y[i] | a[year[i]] + b_union_yes * x_union[i] + b_black * x_black[i] + b_hisp *
x_hisp[i], sigma_y);
    }
}
```

3.2 Varying intercept

```
model.stan.union.ethn.data <- list(
  N = nrow(Males),
  J = length(unique(Males$year)),
  y = Males$wage,
  x_union = Males$union,
  x_black = data.ethn$V1_black,
  x_hisp = data.ethn$V1_hisp,
  year = as.numeric(Males$year)
  )

str(model.stan.union.ethn.data)
```

```
## List of 7
##  $ N      : int 4360
##  $ J      : int 8
##  $ y      : num [1:4360] 1.2 1.85 1.34 1.43 1.57 ...
##  $ x_union: num [1:4360] 0 0 0 0 0 0 0 0 0 0 ...
##  $ x_black: int [1:4360] 0 0 0 0 0 0 0 0 0 0 ...
##  $ x_hisp : int [1:4360] 0 0 0 0 0 0 0 0 0 0 ...
##  $ year   : num [1:4360] 1 2 3 4 5 6 7 8 1 2 ...
```

```
model.varying.intercept.stan <- stan(file='varying_intercepts.stan',
                     data=model.stan.union.ethn.data,
                     iter=1000,
                     chains=4)
```

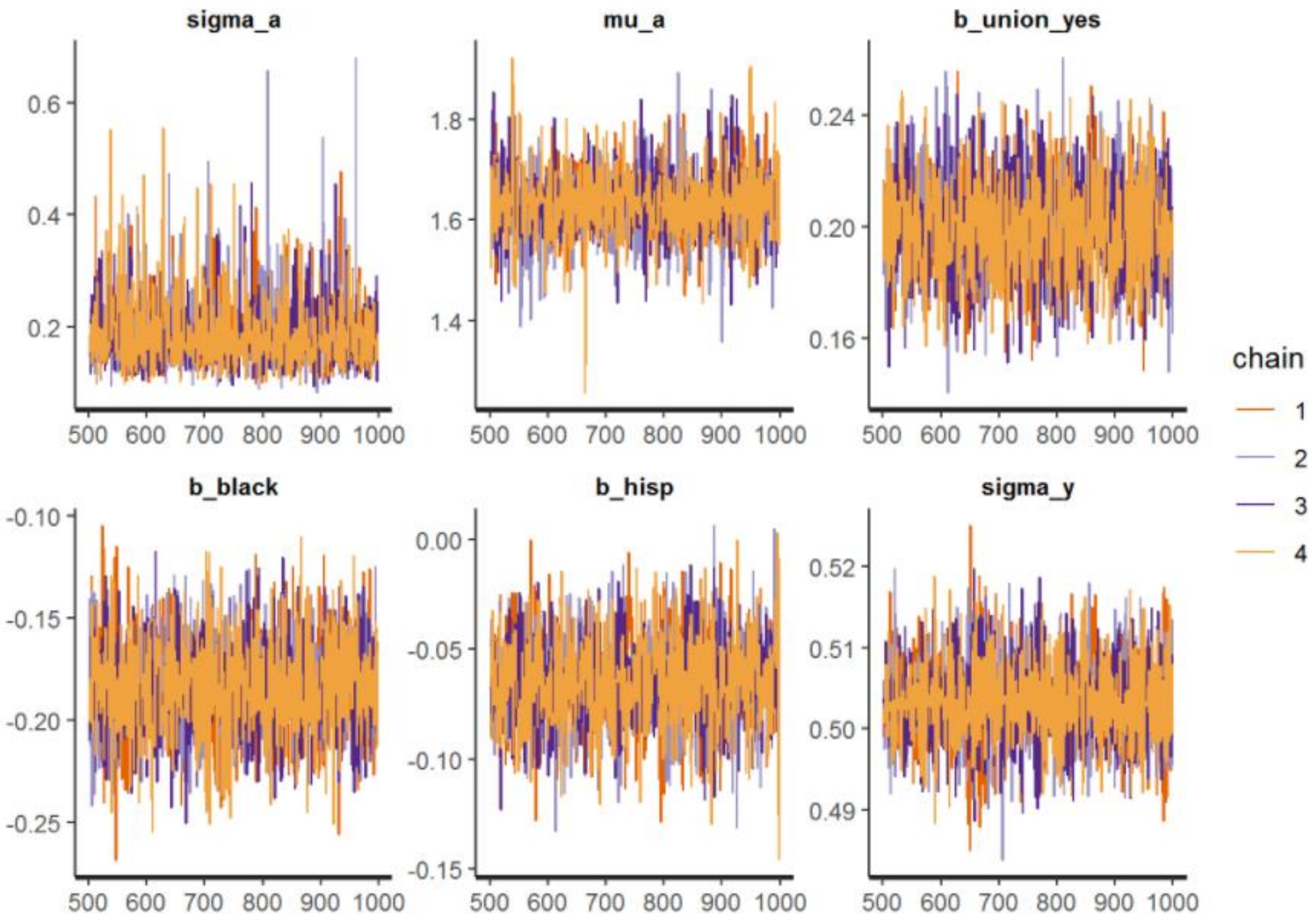
3.2 Varying intercept

```
#RStan output
print(model.varying.intercept.stan, pars = c('sigma_a','mu_a', 'b_union_yes','b_black','b_hisp', 'sigma_y' ))
```

```
## Inference for Stan model: 02_varying_intercepts.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##              mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## sigma_a      0.19       0 0.06  0.11  0.15  0.18  0.22  0.36  1749    1
## mu_a         1.63       0 0.07  1.50  1.59  1.63  1.67  1.78  1769    1
## b_union_yes  0.20       0 0.02  0.16  0.19  0.20  0.21  0.24  2445    1
## b_black     -0.18       0 0.02 -0.23 -0.20 -0.18 -0.17 -0.14  3588    1
## b_hisp      -0.06       0 0.02 -0.11 -0.08 -0.06 -0.05 -0.03  2092    1
## sigma_y      0.50       0 0.01  0.49  0.50  0.50  0.51  0.51  4565    1
##
## Samples were drawn using NUTS(diag_e) at Mon Sep 07 00:22:47 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

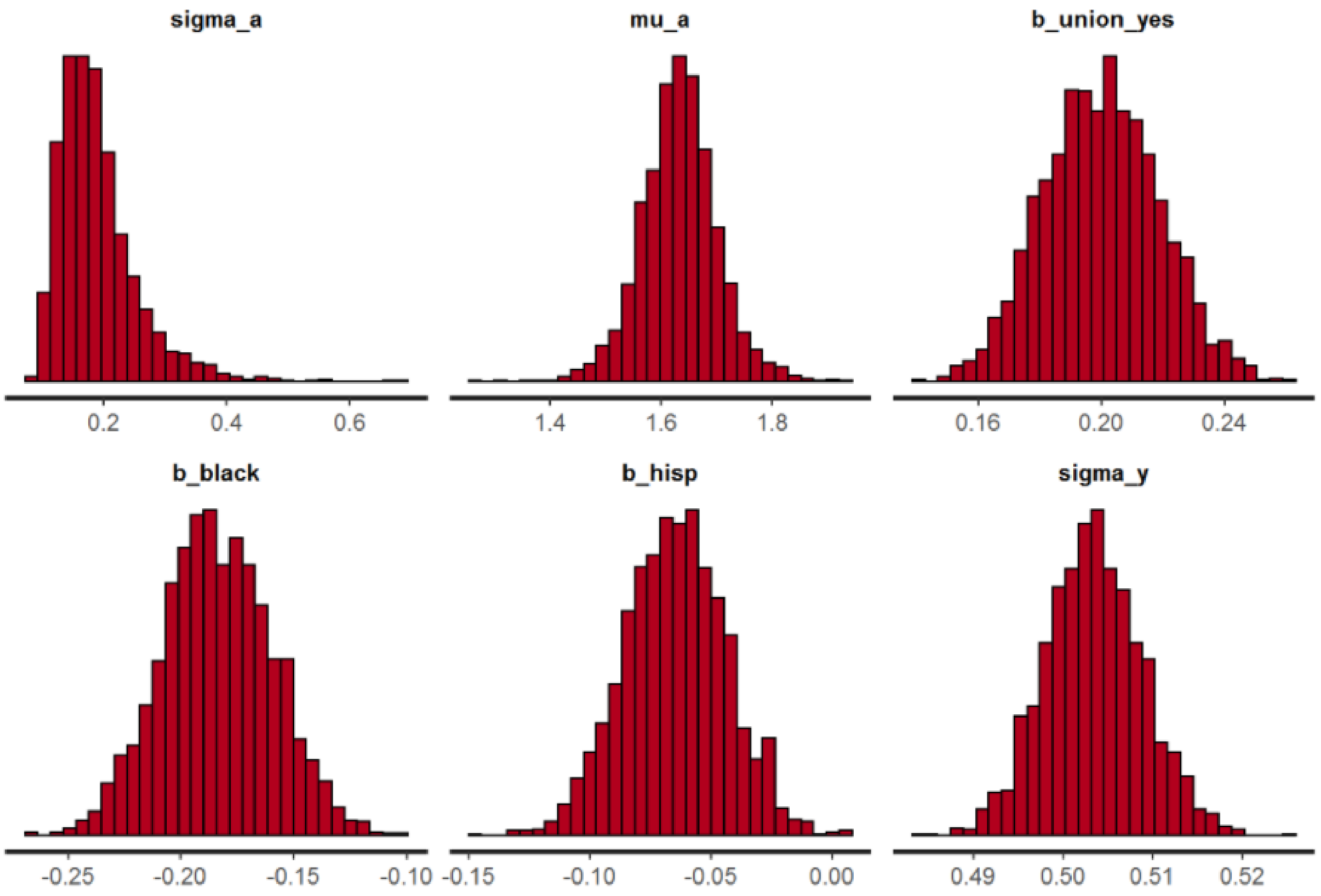
3.2 Varying intercept

```
traceplot(model.varying.intercept.stan, pars = c('sigma_a','mu_a', 'b_union_yes','b_black','b_hisp', 'sigma_y' ))
```



3.2 Varying intercept

```
stan_hist(model.varying.intercept.stan, pars = c('sigma_a','mu_a', 'b_union_yes','b_black','b_hisp', 'sigma_y' ))
```



3.2 Varying intercept

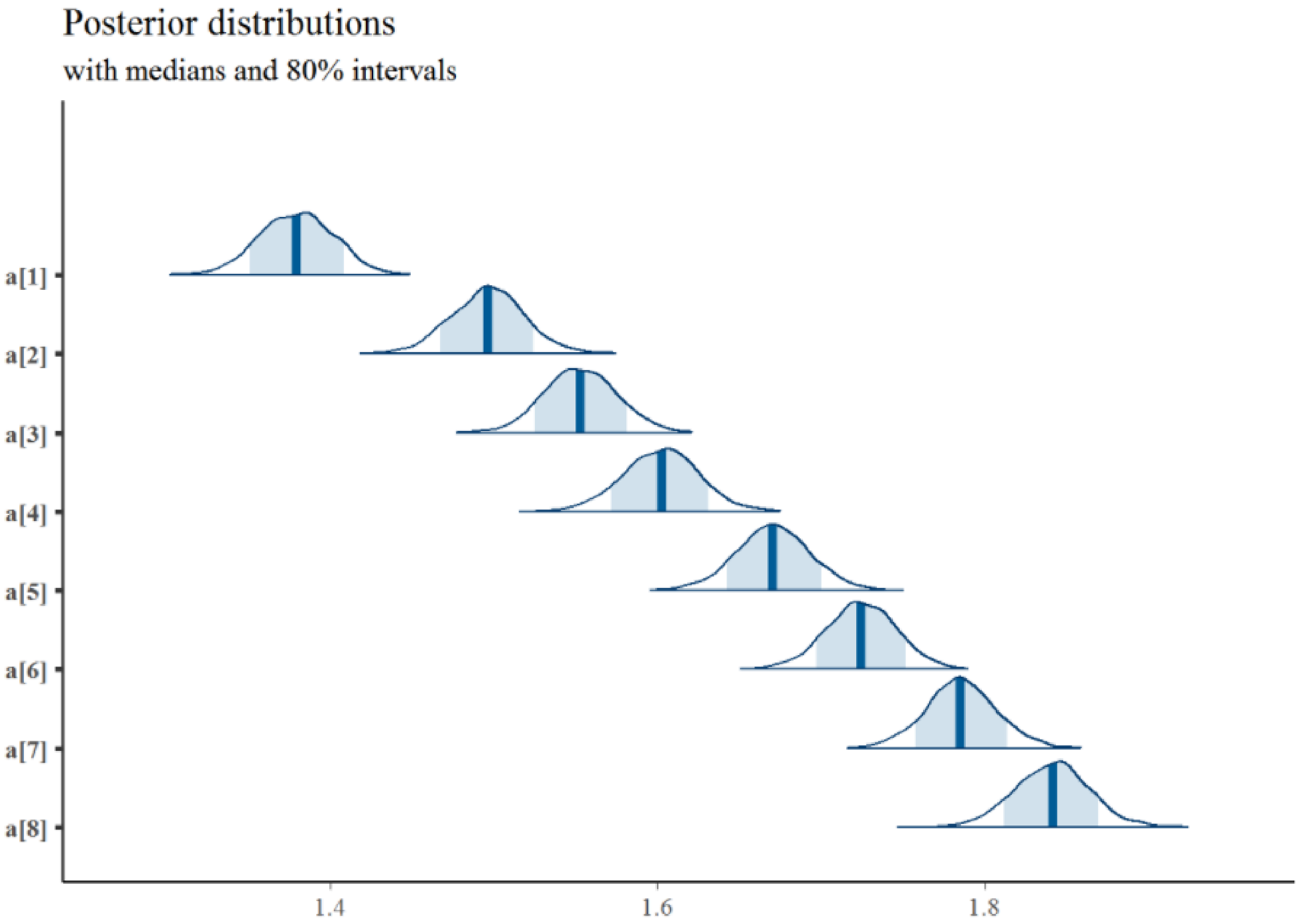
- mcmc_areas graphs

```
array.intercept <- as.array(model.varying.intercept.stan)
```

```
plot.title.intercept <- ggtitle("Posterior distributions",
                                "with medians and 80% intervals")

mcmc_areas.intercept <- mcmc_areas(array.intercept,
           pars = c('a[1]',
                    'a[2]',
                    'a[3]',
                    'a[4]',
                    'a[5]',
                    'a[6]',
                    'a[7]',
                    'a[8]'),
           prob = 0.8) + plot.title.intercept

mcmc_areas.intercept
```

3.2 Varying intercept

- **mcmc_areas** graphs



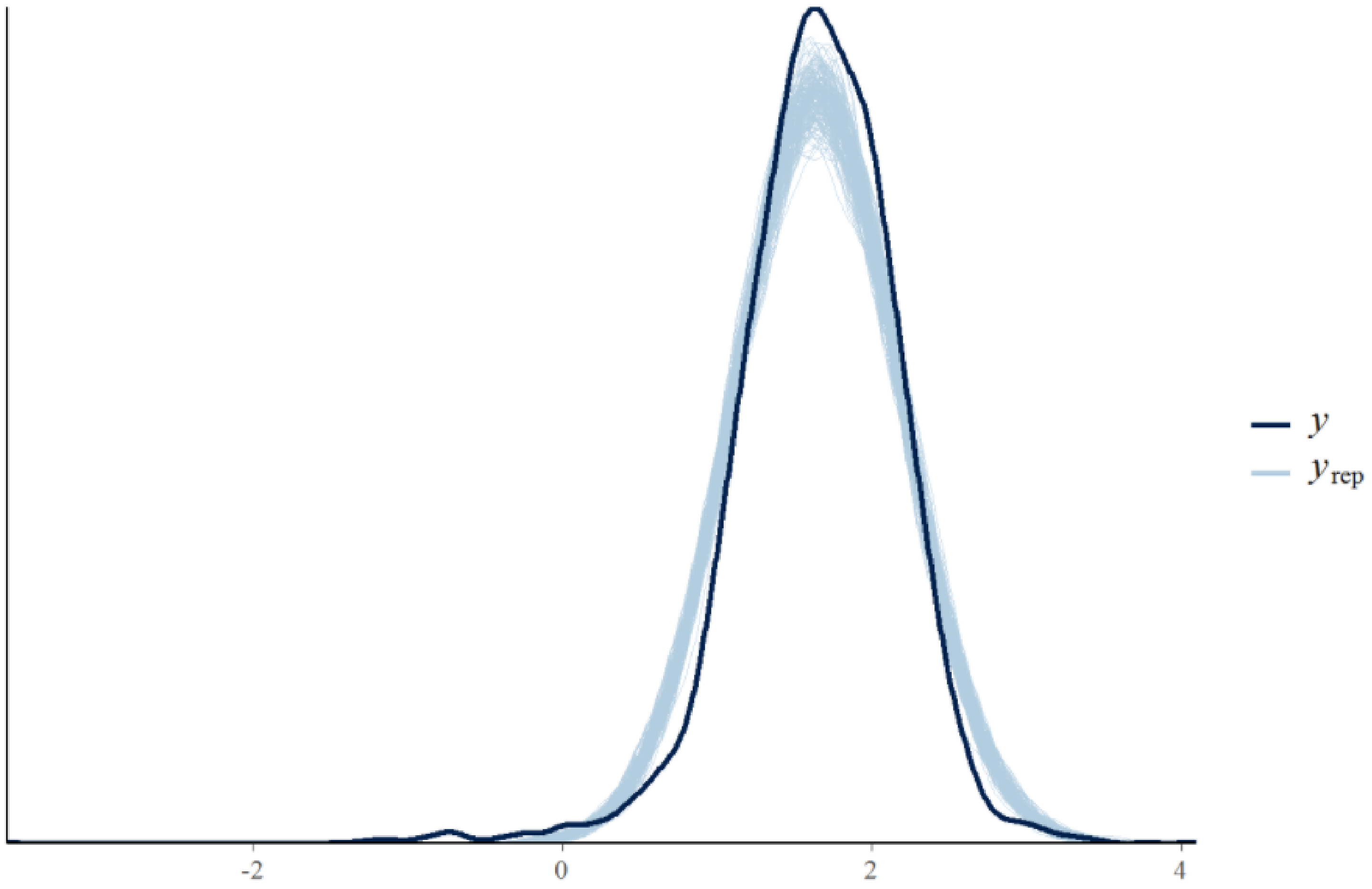
3.2 Varying intercept

```
print(model.varying.intercept.stan, pars = c('a[1]','a[2]','a[3]','a[4]','a[5]','a[6]','a[7]','a[8]'))
```

```
## Inference for Stan model: 02_varying_intercepts.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##      mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## a[1] 1.38       0 0.02 1.34 1.36 1.38 1.40  1.42  2610    1
## a[2] 1.50       0 0.02 1.45 1.48 1.50 1.51  1.54  2752    1
## a[3] 1.55       0 0.02 1.51 1.54 1.55 1.57  1.60  3080    1
## a[4] 1.60       0 0.02 1.56 1.59 1.60 1.62  1.65  3253    1
## a[5] 1.67       0 0.02 1.63 1.66 1.67 1.69  1.72  2846    1
## a[6] 1.72       0 0.02 1.68 1.71 1.72 1.74  1.77  3066    1
## a[7] 1.79       0 0.02 1.74 1.77 1.79 1.80  1.83  2991    1
## a[8] 1.84       0 0.02 1.80 1.83 1.84 1.86  1.89  2705    1
##
## Samples were drawn using NUTS(diag_e) at Mon Sep 07 00:22:47 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

3.2 Varying intercept

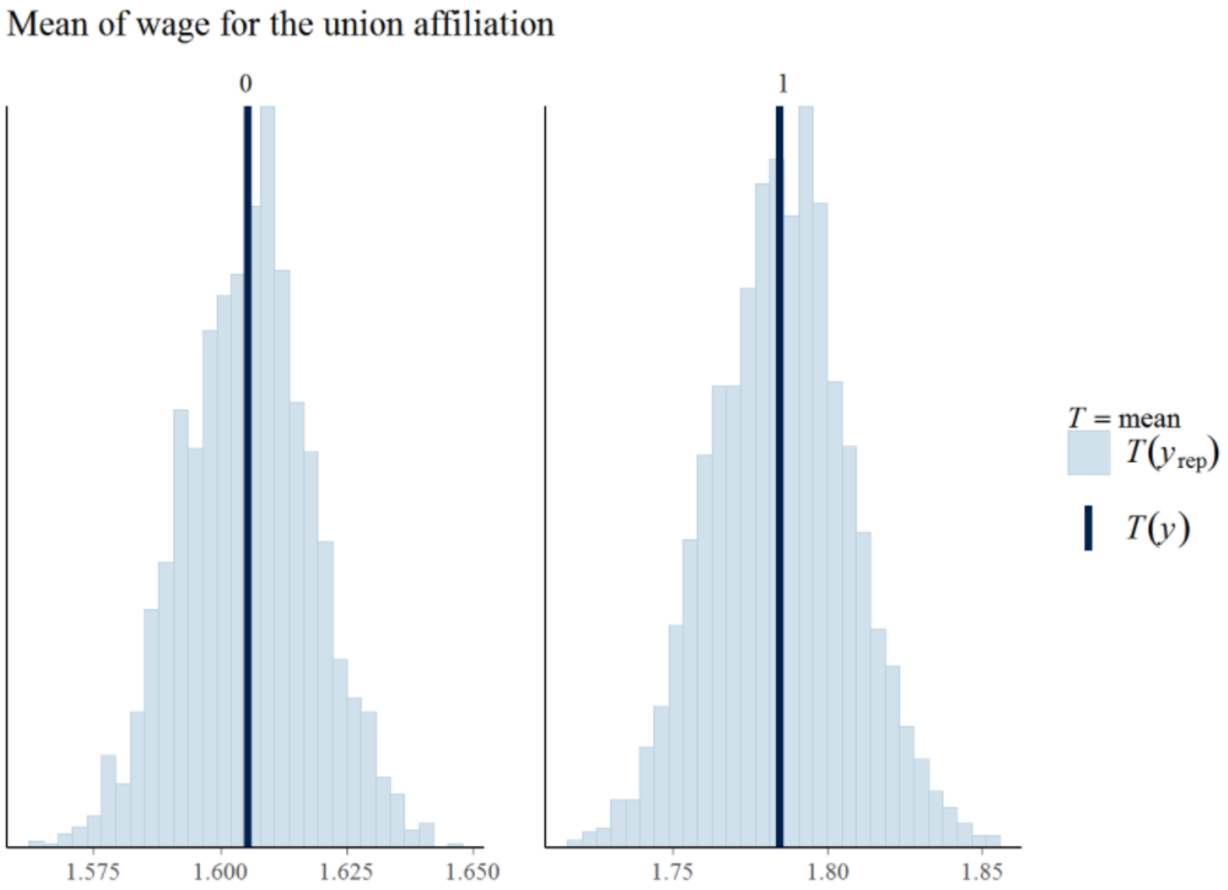
- **ppc_dens_overlay** graphs
- Comparison between empirical distribution of the data y to the distributions of simulated data y_rep
- Simulation distributions for **the varying intercept model**



3.2 Varying intercept

```
ppc_stat_grouped.intercept <- ppc_stat_grouped(
  y = model.stan.union.ethn.data$y,
  yrep = y_rep.intercept,
  group = model.stan.union.ethn.data$x_union,
  stat = 'mean')
```

```
ppc_stat_grouped.intercept + ggtitle("Mean of wage for the union affiliation")
```



3. Models

3.1 Varying intercept model with no predictor

3.2 Varying intercept

3.3 Varying slope

$$y_i = \alpha + \beta_{j[i]union} * x_{union} + \beta_{black} * x_{black} + \beta_{hisp} * x_{hisp} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_y^2)$$

$$\beta_{j[i]} \sim N(\mu_\beta, \sigma_\beta^2)$$

3.3 Varying slope

Stan

```
data {
  int<lower=1> N; // number of obs
  int<lower=0> J; // Number of years
  vector[N] y; // outcome data (log wage)
  int<lower=0,upper=1> x_union[N]; // Input data (union)
  int<lower=0,upper=1> x_black[N];
  int<lower=0,upper=1> x_hisp[N];
  int year[N]; //year ID variable
}
parameters {
  real a;
  real b_union[J];  // year j slopes
  real b_hisp;
  real b_black;
  real mu_b; // prior on beta
  real<lower=0> sigma_y; // prior sigma of y in year j
  real<lower=0> sigma_b; // hyperparameter sigma of beta
}
model {
  b_union ~ normal(mu_b, sigma_b);
  for (n in 1:N)
    y[n] ~ normal(a + b_union[year[n]] * x_union[n] + b_black * x_black[n] + b_hisp * x_hisp[n], sigma_y);
}
generated quantities {
 vector[N] log_lik;
 vector[N] y_hat;
  for (i in 1:N){
    y_hat[i] = normal_rng(a + b_union[year[i]] * x_union[i] + b_black * x_black[i] + b_hisp * x_hisp[i],
sigma_y);
    log_lik[i] = normal_lpdf(y[i] | a + b_union[year[i]] * x_union[i] + b_black * x_black[i] + b_hisp *
x_hisp[i], sigma_y );
    }
}
```

3.3 Varying slope

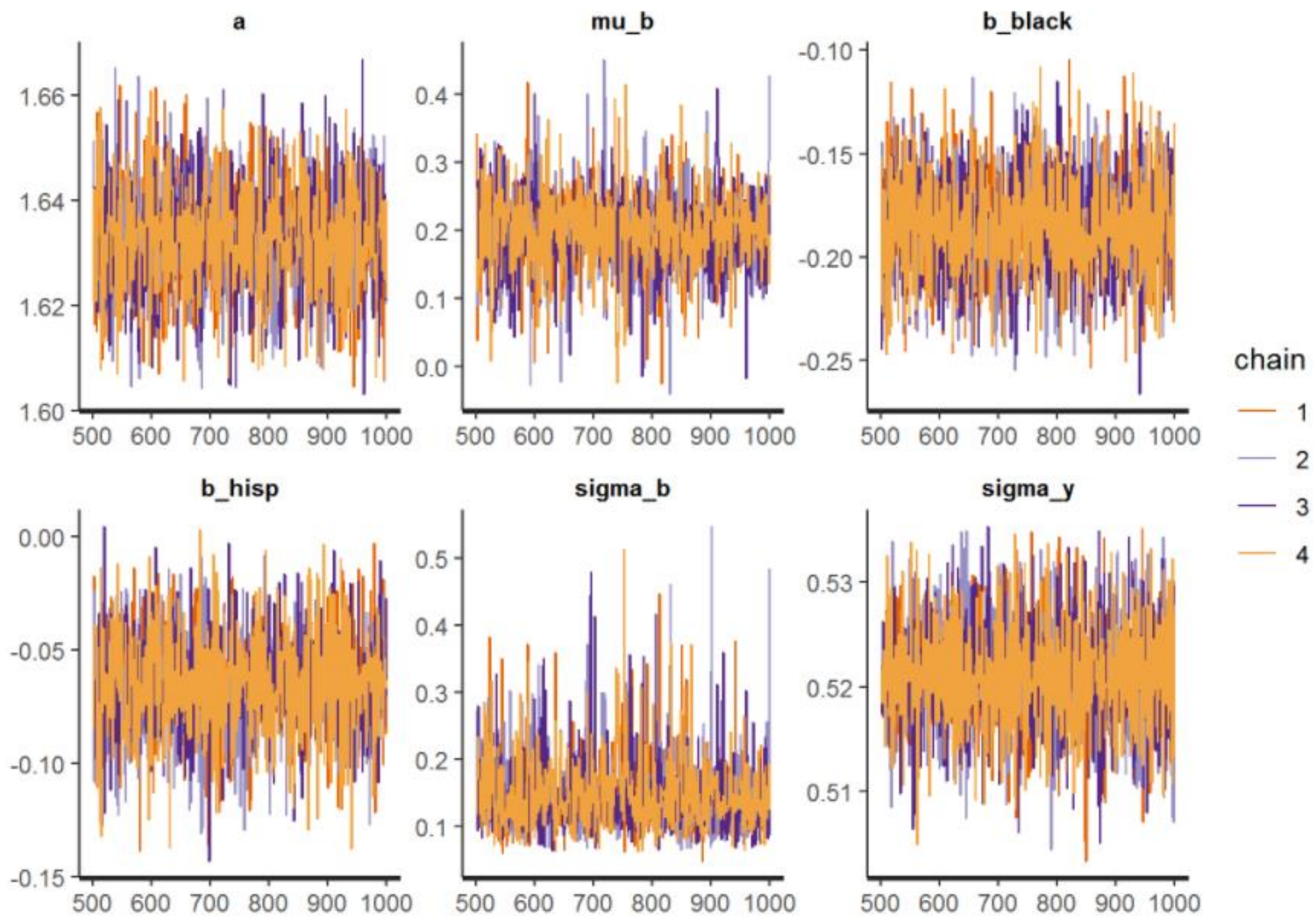
```
model.varying.slope.stan <- stan(file='varying_slopes.stan',
                                 data=model.stan.union.ethn.data,
                                 iter=1000,
                                 chains=4)
```

```
#RStan output
print(model.varying.slope.stan, pars = c('a', 'mu_b', 'b_black', 'b_hisp','sigma_b', 'sigma_y' ))
```

```
## Inference for Stan model: 03_varying_slopes.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##          mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## a        1.63       0 0.01  1.61  1.63  1.63  1.64  1.65  1842    1
## mu_b     0.20       0 0.06  0.07  0.16  0.20  0.23  0.31  1487    1
## b_black -0.18       0 0.03 -0.23 -0.20 -0.18 -0.17 -0.14  2506    1
## b_hisp  -0.07       0 0.02 -0.11 -0.08 -0.07 -0.05 -0.02  2266    1
## sigma_b  0.15       0 0.06  0.07  0.11  0.14  0.17  0.30  1422    1
## sigma_y  0.52       0 0.01  0.51  0.52  0.52  0.52  0.53  4101    1
##
## Samples were drawn using NUTS(diag_e) at Mon Sep 07 01:05:05 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

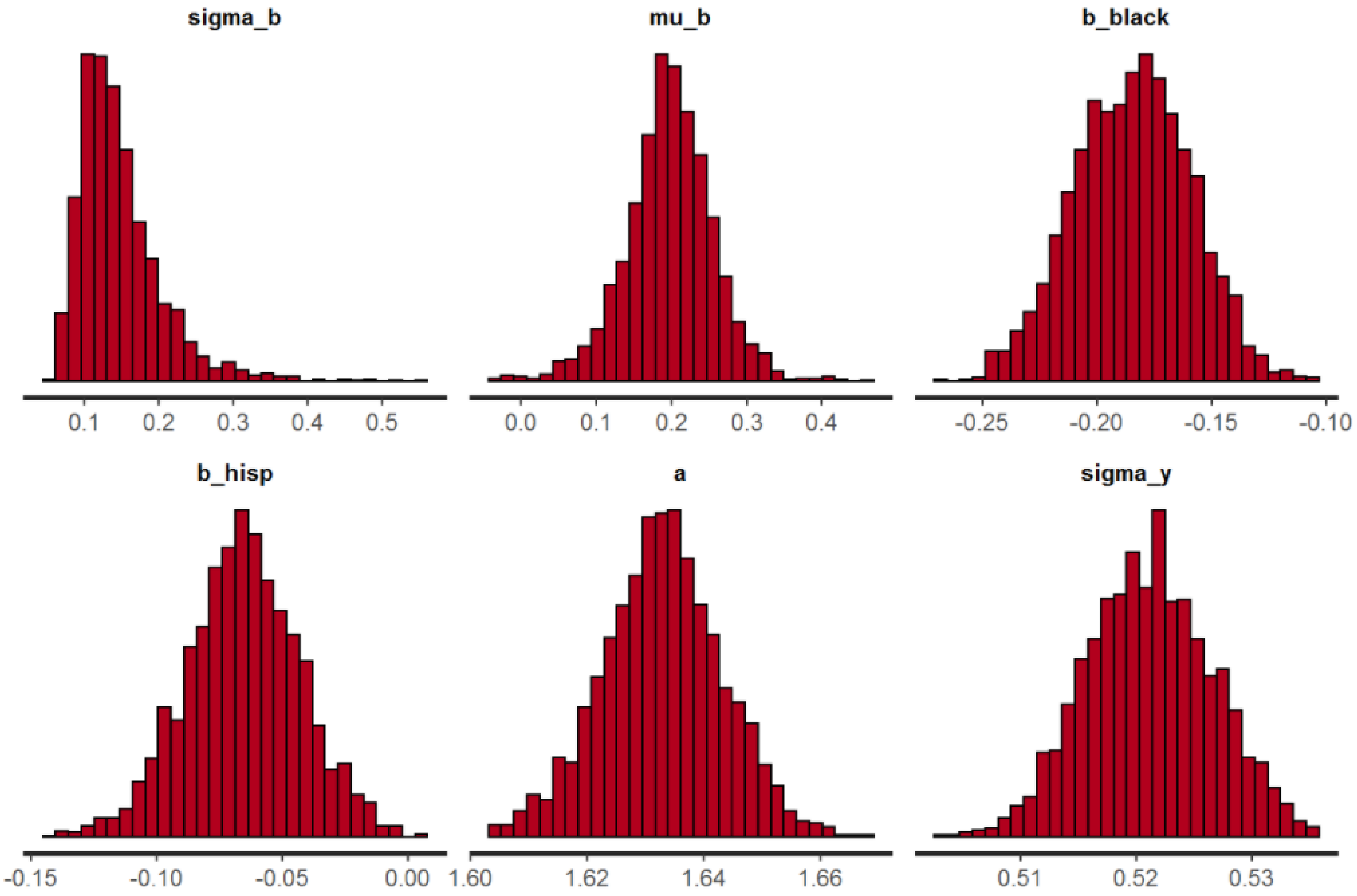
3.3 Varying slope

```
traceplot(model.varying.slope.stan, pars = c('a', 'mu_b', 'b_black', 'b_hisp','sigma_b', 'sigma_y'))
```



3.3 Varying slope

```
stan_hist(model.varying.slope.stan,  pars = c('sigma_b', 'mu_b', 'b_black', 'b_hisp','a', 'sigma_y' ))
```



3.3 Varying slope

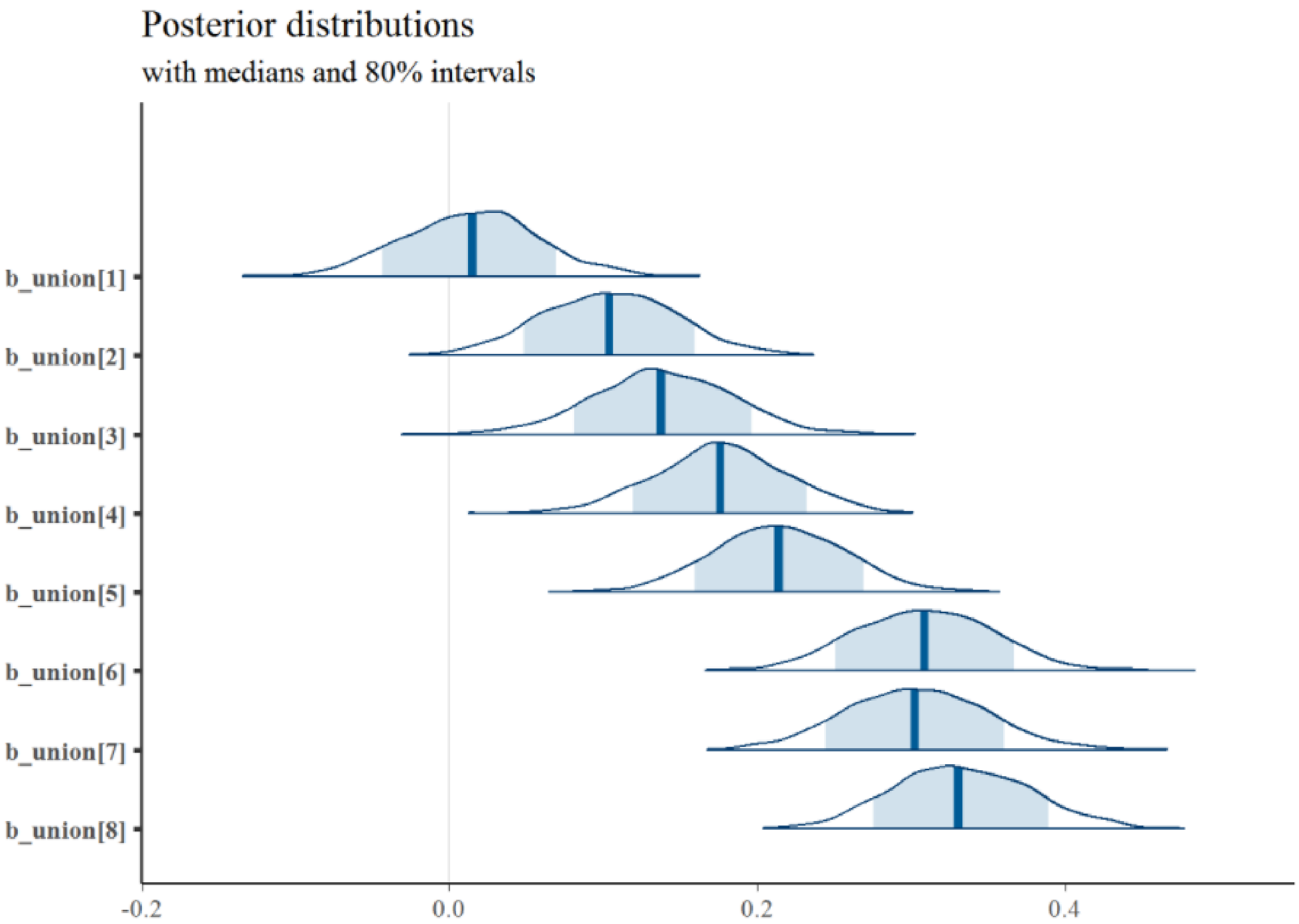
- **mcmc_areas** graphs

```
array.slope <- as.array(model.varying.slope.stan)


plot.title.slope <- ggtitle("Posterior distributions",
                            "with medians and 80% intervals")

mcmc_areas.slope <- mcmc_areas(array.slope,
           pars = c("b_union[1]",
                    "b_union[2]",
                    "b_union[3]",
                    "b_union[4]",
                    "b_union[5]",
                    "b_union[6]",
                    "b_union[7]",
                    "b_union[8]"),
           prob = 0.8) + plot.title.intercept
```

```
mcmc_areas.slope
```

3.3 Varying slope

- **mcmc_areas** graphs



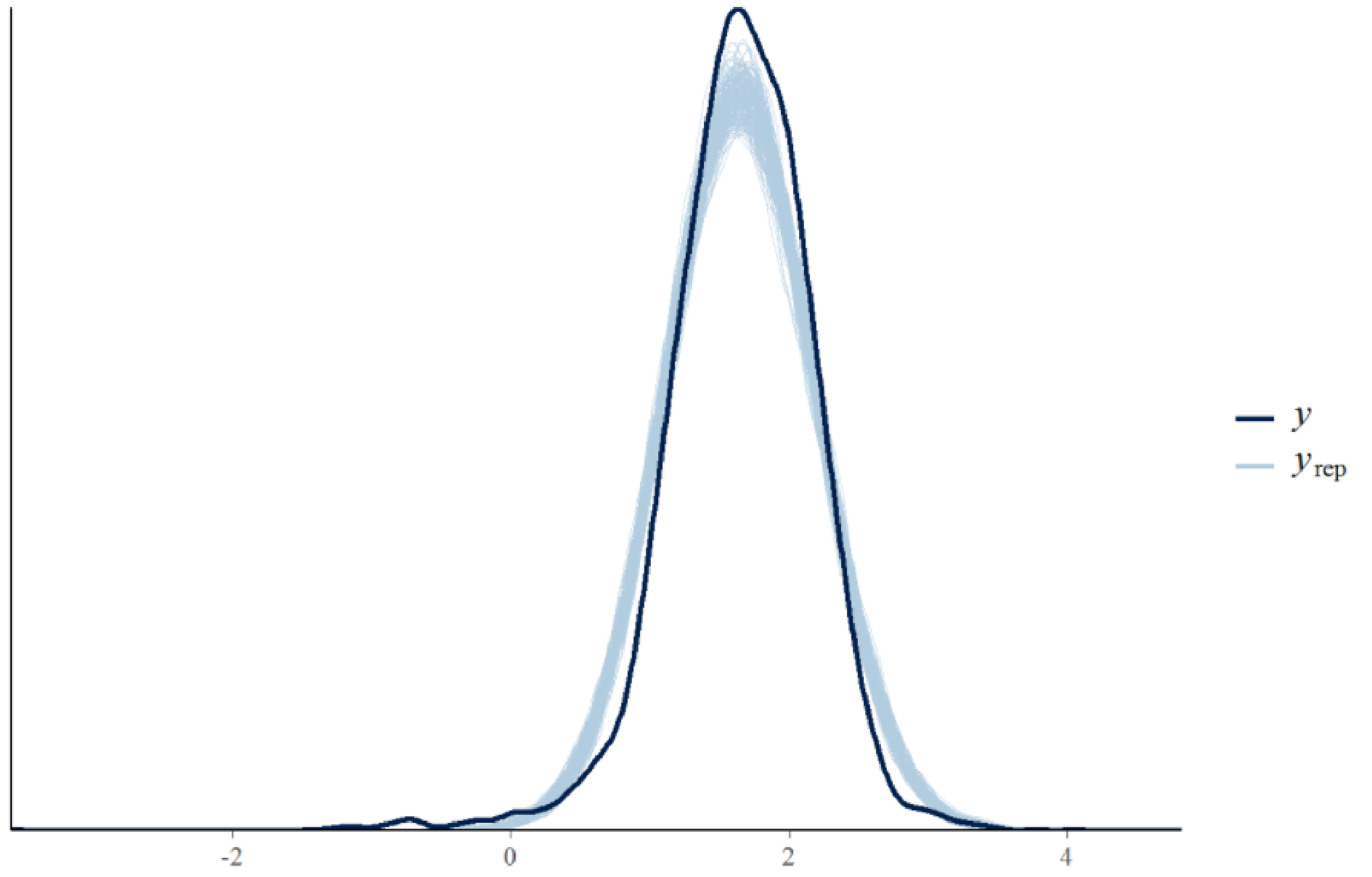
3.3 Varying slope

```
print(model.varying.slope.stan, pars = c("b_union[1]","b_union[2]",
                                         "b_union[3]","b_union[4]",
                                         "b_union[5]","b_union[6]",
                                         "b_union[7]","b_union[8]"))
```

```
## Inference for Stan model: 03_varying_slopes.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##             mean se_mean   sd  2.5%   25%  50%  75% 97.5% n_eff Rhat
## b_union[1] 0.01       0 0.04 -0.07 -0.02 0.01 0.04  0.10  2000    1
## b_union[2] 0.10       0 0.04  0.02  0.07 0.10 0.13  0.19  2184    1
## b_union[3] 0.14       0 0.05  0.05  0.11 0.14 0.17  0.22  2884    1
## b_union[4] 0.18       0 0.04  0.09  0.15 0.18 0.20  0.26  2555    1
## b_union[5] 0.21       0 0.04  0.13  0.19 0.21 0.24  0.30  2834    1
## b_union[6] 0.31       0 0.05  0.22  0.28 0.31 0.34  0.40  2925    1
## b_union[7] 0.30       0 0.05  0.22  0.27 0.30 0.33  0.40  2295    1
## b_union[8] 0.33       0 0.04  0.25  0.30 0.33 0.36  0.42  2821    1
##
## Samples were drawn using NUTS(diag_e) at Mon Sep 07 01:05:05 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

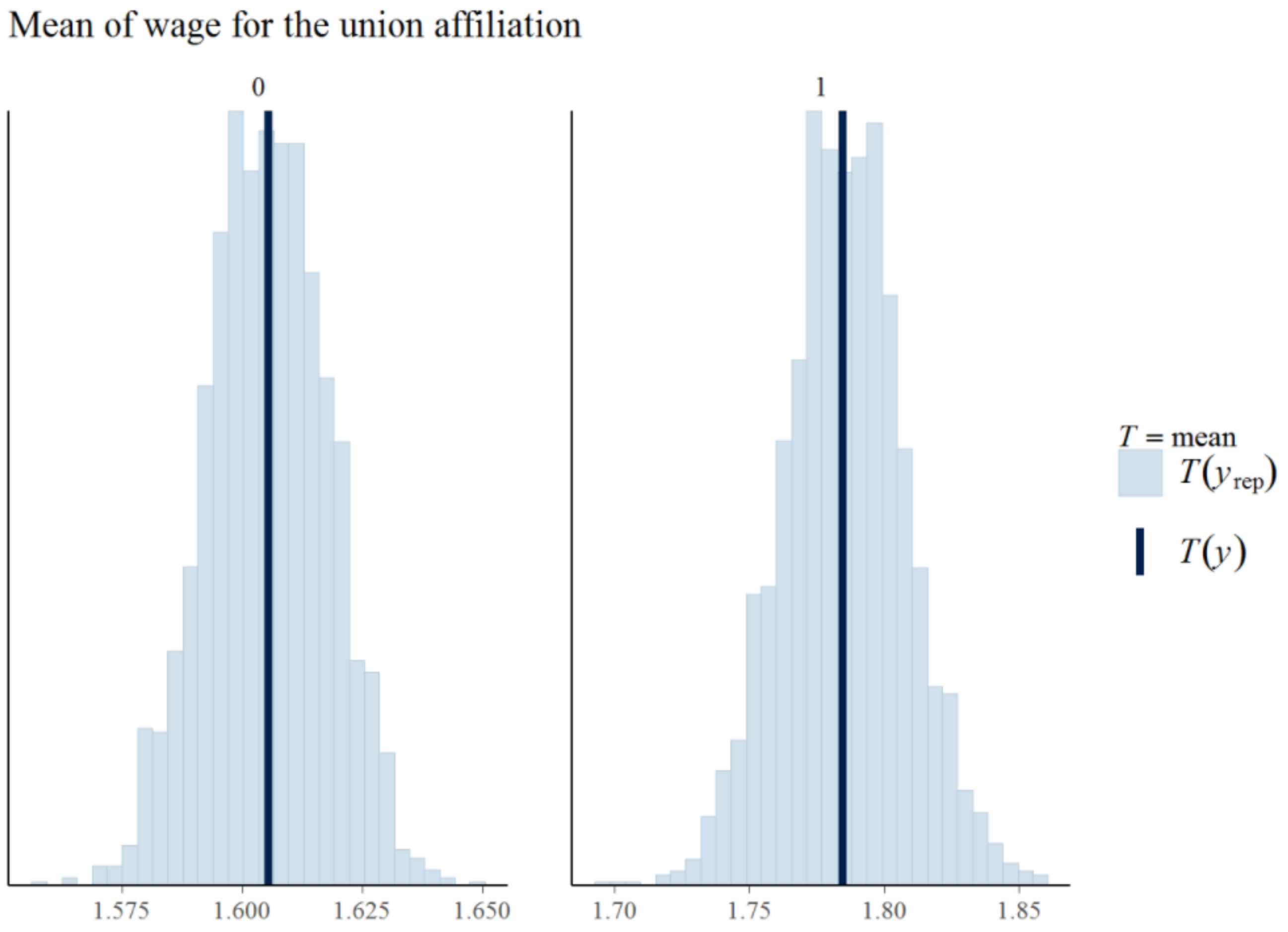
3.3 Varying slope

- **ppc_dens_overlay** graphs
- Comparison between empirical distribution of the data y to the distributions of simulated data y_rep
- Simulation distributions for **the varying slope model**



3.3 Varying slope

```
ppc_stat_grouped.slope <- ppc_stat_grouped(
  y = model.stan.union.ethn.data$y,
  yrep = y_rep.slope,
  group = model.stan.union.ethn.data$x_union,
  stat = 'mean')
```



- Evaluating the models:

```
log.lik.intercepts <- extract_log_lik(model.varying.intercept.stan)
loo.intercepts <- loo(log.lik.intercepts)
```

```
print(loo.intercepts)
```

```
##
## Computed from 2000 by 4360 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -3200.2  92.8
## p_loo        14.7   1.7
## looic      6400.4 185.7
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     4359  100.0%  1652
##  (0.5, 0.7]   (ok)          1    0.0%  349
##    (0.7, 1]   (bad)         0    0.0%  <NA>
##    (1, Inf)   (very bad)    0    0.0%  <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

4 – EVALUATING THE MODELS

- Evaluating the models:

```
log.lik.slopes <- extract_log_lik(model.varying.slope.stan)
loo.slopes <- loo(log.lik.slopes)
```

```
print(loo.slopes)
```

```
##
## Computed from 2000 by 4360 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -3351.0  89.8
## p_loo        12.5   1.4
## looic      6702.0 179.5
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

4 – EVALUATING THE MODELS

- The varying-intercept model is preferred in terms of lower LOOIC.

```
model.eval <- loo_compare(loo.interceps, loo.slopes)
print(model.eval, simplify = FALSE)
```

```
##         elpd_diff se_diff elpd_loo se_elpd_loo p_loo   se_p_loo looic   se_looic
## model1     0.0       0.0 -3200.2     92.8        14.7     1.7   6400.4   185.7
## model2  -150.8      17.0 -3351.0     89.8        12.5     1.4   6702.0   179.5
```

5 - DISCUSSION

- The loo function was employed to select the best model, and the varying-intercept model is slightly favourite in terms of lower LOOIC.

- $\beta_{[1987]union}$ = 0.33 in the Credible Interval of 95% [0.25 – 0.42].

- In this specific problem, it can be interpreted whether the employee is affiliated to the union, the value of variable wage increases by 0.33 on average, for the year of 1987.

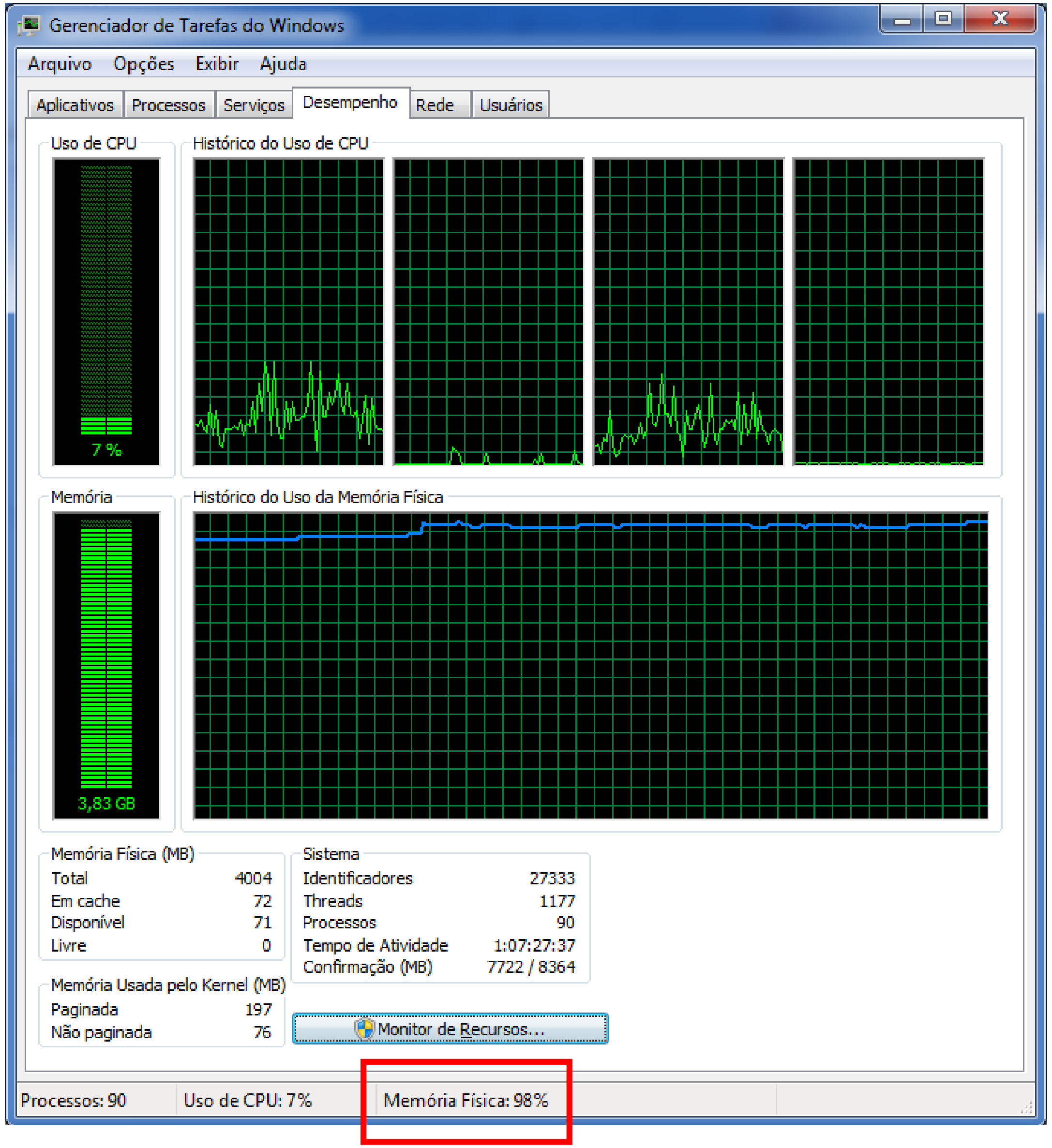
- The results are consistent with the observations made by Vella and Verbeek.

- It could be explained because Black and Hispanic workers may choose to bargain their wages through union membership rather than on an individual basis.

- They may reduce its impact via union membership. Few of the other individual related variables appear to have a statistically significant impact on union membership.

5 - DISCUSSION

- Due to the higher number of categories of the other variables (i.e., school, occupation, residence, etc.) the time taken to run the algorithm was very long. The most important bottleneck to conduct further analysis was the computational RAM memory available in the machine.



6 - CONCLUSIONS

- In this work, it was possible to employ the group-level effect for the year grouping factors and take into account the hierarchical structure of the data.

- Additional analysis was performed varying the slope for union covariate and intercept within groups.

- The wage is likely to be influenced by the union membership. It is more likely to increase their wage in 0.2, in average, for the Credible Interval of 95% [0.07 – 0.31] compared to non union membership.

- Compared to Other workers category, the Black workers were 0.18 less likely to have a higher wage in a Credible Interval of 95% [-0.13 – -0.23].

REFERENCES

[1] Stan Development Team, “RStan: The R interface to Stan.” 2020, [Online]. Available: http://mc-stan.org/.

[2] F. Vella and M. Verbeek, “Whose wages do unions raise ? A dynamic model of unionism and wage,” *Journal of Applied Econometrics*, vol. 13, pp. 163–183, 1998.

[3] Y. Croissant and S. Graves, *Ecdat: Data sets for econometrics*. 2020.

[4] P.-C. Bürkner, “Advanced Bayesian multilevel modeling with the R package brms,” *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018, doi: 10.32614/RJ-2018-017.

[5] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman, “Visualization in bayesian workflow,” *J. R. Stat. Soc. A*, vol. 182, no. 2, pp. 389–402, 2019, doi: 10.1111/rssa.12378.

[6] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*, 1st ed. Lawrence Erlbaum Associates, 2007.

[7] J. Hox, *Multilevel analysis - techniques and applications*, 2nd ed. Lawrence Erlbaum Associates, 2002.

Thank you very much for your attention

Grazie mille per la vostra attenzione

Muito obrigado pela sua atenção