Final Project Report

Luning Li

2021/4/15

Abstract

This paper focuses on the people aged 25-35 working for wages in Ohio in 2018, and shows that the gender gap exists even taken account in factors such as usual weekly working hours, occupation income scores, race, age and education level. This paper indicates that females are paid only 86.6% (84.3%,89.0%) of their male counterparts, conditioned on the independent variables mentioned.

1 Introduction

With the development of the feminist movement in the past decade, more and more women choose to enter the workplace. However, the gender pay gap has attracted wide attention. In 2017, women working full time and year-round in the United States typically were paid just 80 percent of what men were paid, a gap of 20 percent (Fontenot et al., 2018). The cause of gender pay gap is controversal. This kind of dispute mainly focuses on whether this gap is due to gender discrimination, or to the difference objective factors between male and female. The purpose of this project is to first investigate to what extend gender factor can affect the wage, and then provide women with advice to get a high-paying job by establishing a regression model on wage income. It may take another decade for women to get equal pay for equal job as men; however, they could contribute more to other independent variables of wage to get themselves a higher wage.

Focusing on people at the age of 25-35 in Ohio, who are not self-employed and worked 52 weeks in 2018, this paper tries to find the subjective factors that influence wage. Bayesian hierarchy model is introduced on county level means of wages in Ohio. The advantage of a hierarchy model is that it provides a flexible way to explain the county level means. The results from hierarchy model pulls the estimates of county level means towards the population mean and therefore, less sensitive to noises compared to no pooling method.

2 Data

The data set used in this research is the American Community Survey 2019 Sample extracted from IPUMS USA (www.ipums.org). It involves the following variables:

- `AGE`: Age of the respondents. Filter the young respondents (control the age to 25-35).
- `SEX`: Sex of the respondents. 1 if male, 2 if female.
- `WKSWORK1`: The number of weeks that respondents worked last year. Filter respondents with `WKSWORK1`$>$ 40.
- `UHRWORK1`: The number of hours per week that the respondent usually worked, if the person worked during the previous year. Filter respondents with `UHRWORK1`>0.
- `INCWAGE`: The total pre-tax wage and salary income - that is, money received as an employee - for the previous year with 999999 = N/A and 999998 = Missing. Filter respondents with `INCWAGE`>0.
- `OCCSCORE`: A constructed variable that assigns occupational income scores to each occupation.
- `PWSTATE2`: The state in which the respondent's primary workplace was located. Filter the respondents working in Ohio (code 39).
- `PWCOUNTY`: The county (or county equivalent) where the respondent worked. 0 if not available.
- `RACWHT`: A bivariate indicator of "White" race.
- `CLASSWKR`: Class of worker. 1 stands for self-employed and 2 stands for working for wages.

- `RACE`: The race of the respondents.

The dependent variable that we are interested in is the average weekly salary (on log scale) that one respondents received last year. That is:

- $Y = \log\left(\frac{\text{INCWAGE}}{\text{WKSWORK1}}\right)$.

The relevant independent variables are:

- `SEX`: gender of the respondents;
- `OCCSCORE`: occupational income score;
- `UHRSWORK`: usual hours worked per week;
- `education`: education level of the respondents;
- `age`: if the age of respondent is 30 years or older;
- `RACWHT`: if the race of the respondent is white or not.

In addition, a hierarchical structure will be applied on the county level mean. i.e. assume that the county level means of the weekly salary are independent samples coming from the same distribution.

2.1 Data Preprocessing:

When extracting the data from IPUMS, choose the case where age of respondents is between 25 and 35, and the primary working State to be Ohio. Then preprocess the data for further analysis:

- Select the respondents who worked in Ohio (this is done when downloading data from the IPUMS website).
- Select the respondents whose age is between 25-35 (this is done when downloading data from IPUMS website).
- Select the respondents whose worked at least 40 weeks last year (`WKSWORK`> 40).
- Select the respondents whose average hours worked per week (`UHRSWORK`) is greater than 0.
- Select the respondents whose wage and salary income (`INCWAGE`) is greater than 0.
- Select the respondents whose occupational income score (`OCCSCORE`) is greater than 0.
- Select the respondents who is working for wages but not self-employed.

After checking, missing values exist for variable `PWCOUNTY`. Those respondents worked in county not identifiable from public-use data was removed from our research. We only focused on the employees working within an identifiable county.

Then we need to modify some variables as following:

- Calculate the dependent variable `Y` by $Y = \log\left(\frac{\text{INCWAGE}}{\text{WKSWORK1}}\right)$.
- Construct a categorical variable `education`:
  - 0 if the respondent does not obtain high school diploma,
  - 1 if the respondent has a high school diploma, but does not have a bachelor degree,
  - 2 if the highest degree of the respondent is a Bachelor's degree,
  - 3 if the highest degree of the respondent is a Master's degree or higher.
- Construct a categorical variable `age` to indicate whether the respondent is elder than 30 years old.
- construct a categorical variable `County` to record the county code for further Bayesian analysis.

After the data processing, our original data looks like:

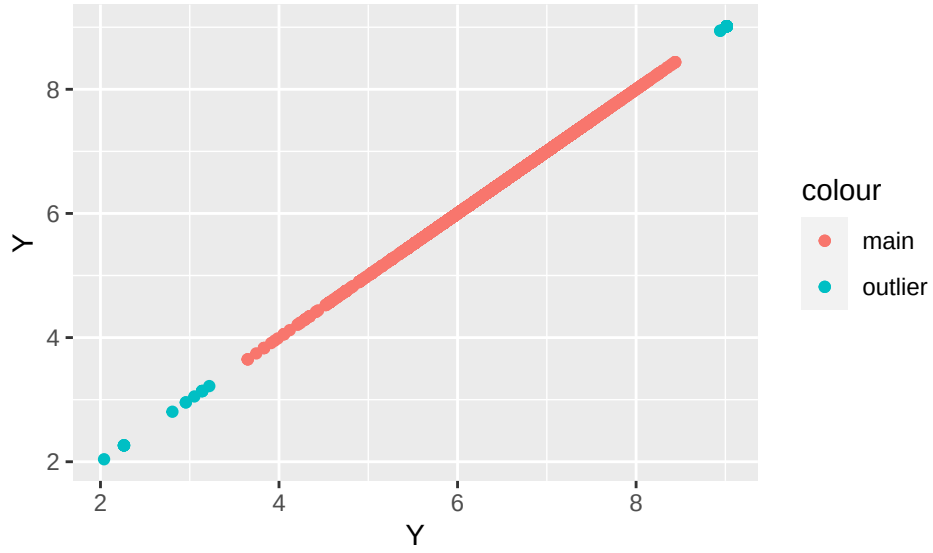
Table 1: Table continues below

PWCOUNTY	Y	SEX	gender	UHRSWORK	OCCSCORE	education	RACWHT
3	6.155	1	Male	35	11	1	2
3	6.512	2	Female	50	33	1	2
3	6.844	2	Female	38	36	3	2

PWCOUNTY	Y	SEX	gender	UHRSWORK	OCCSCORE	education	RACWHT
3	6.296	2	Female	40	20	2	2
3	7.233	1	Male	40	42	2	2
3	7.824	2	Female	40	40	3	2

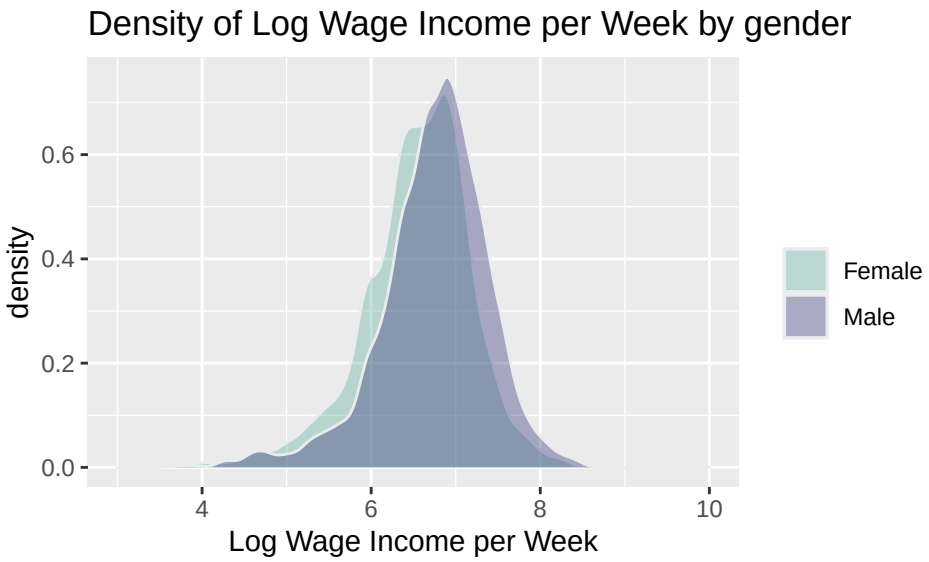
racwht	age	age1	RACE	AGE	COUNTY
white	0	less than 30	1	26	15
white	0	less than 30	1	27	15
white	0	less than 30	1	27	15
white	1	greater than or equal to 30	1	32	15
white	0	less than 30	1	28	15
white	1	greater than or equal to 30	1	34	15

But if we look at the scatter plot of our dependent variable `Y`, we can find that there are some ourliers. For our analysis, I would like to remove those outliers from our data.

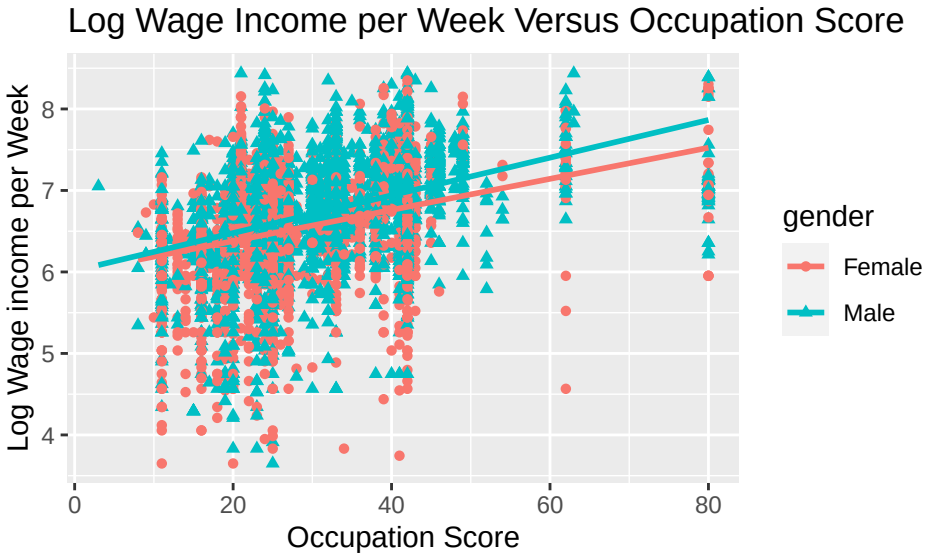


2.2 Exploratory Data Analysis

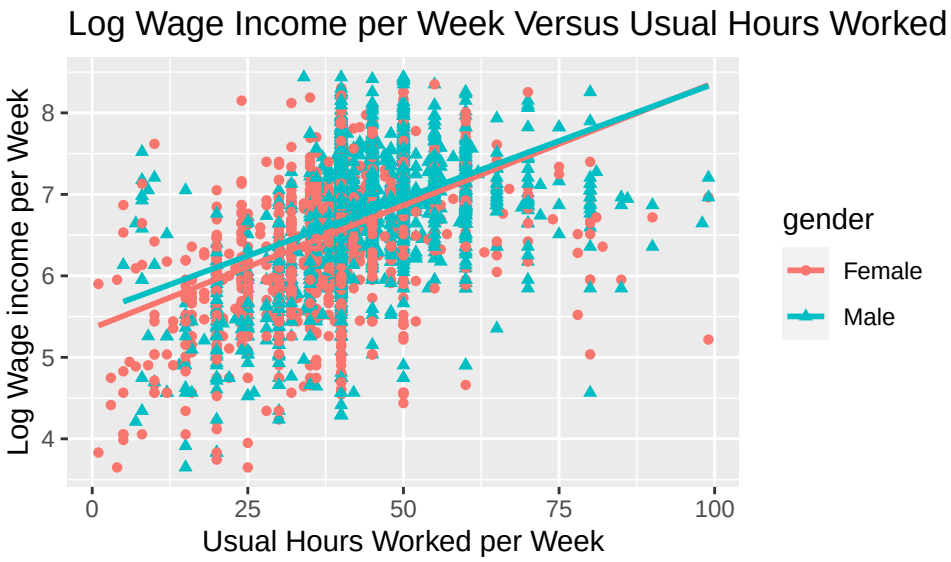
Before fitting the model, the following plots can be applied to investigate the correlations between our dependent variable Y and the corresponding independent variables.



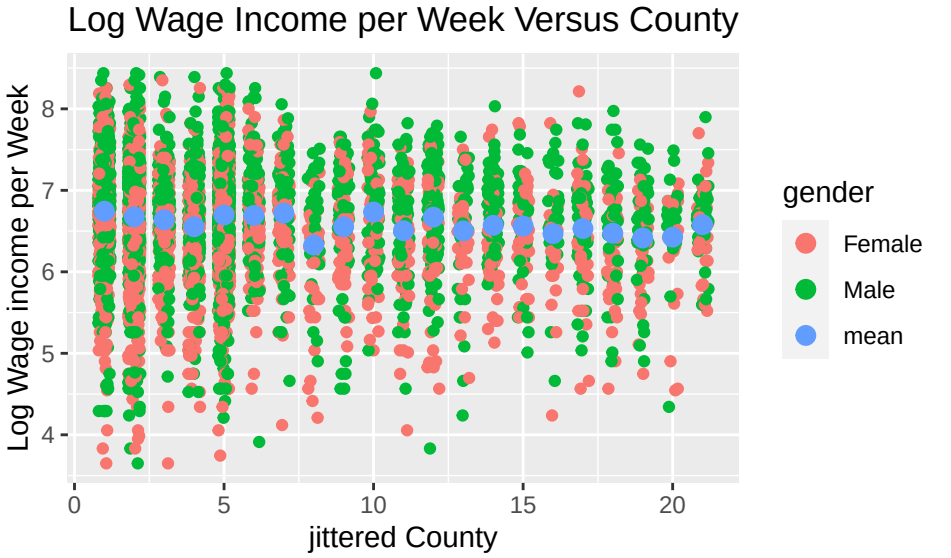
The density plot of log weekly wage income (`Y`) illustrates the gender pay gap exists in Ohio. This density from female is more left-tailed than that from male, and therefore, females are generally paid less than the males. Therefore, this research will treat `gender` as a predictor for wage.



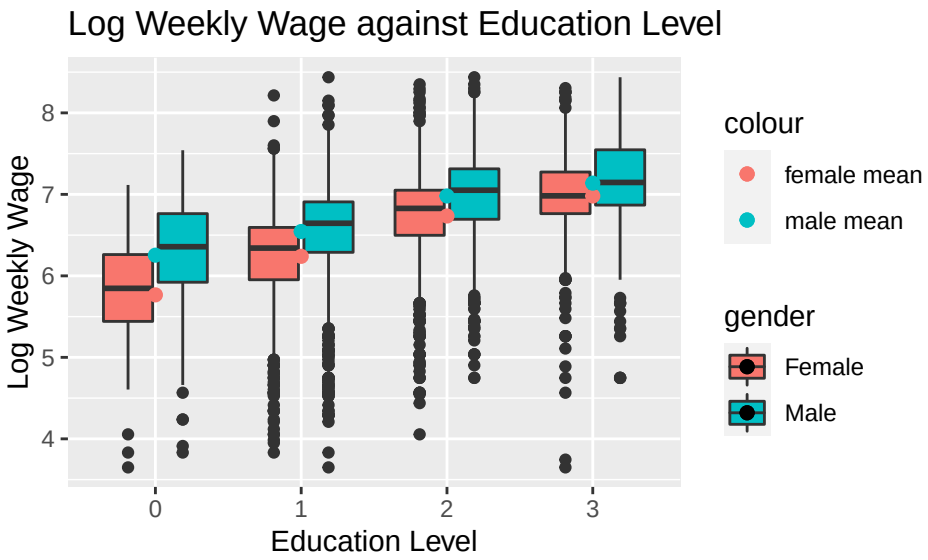
The scatter plot of `Y` against the occupation score (`OCCSCORE`) shows a positive correlation. Therefore, `OCCSCORE` is taken as a predictor for the wage. Moreover, the solid lines represent the linear regression lines between those two variables from each gender. The two regression lines are close to each other, so no interaction term between `OCCSCORE` and `gender` is considered in the model.



The scatter plot of `Y` against the usual hours worked per week (`UHRSWORK`) shows a positive correlation. Therefore, `UHRSWORK` is taken as a predictor for the wage. Again, the solid lines represent the linear regression lines between those two variables from each gender. The two regression lines are close to each other, so no interaction term between `UHRSWORK` and `gender` is considered in the model.



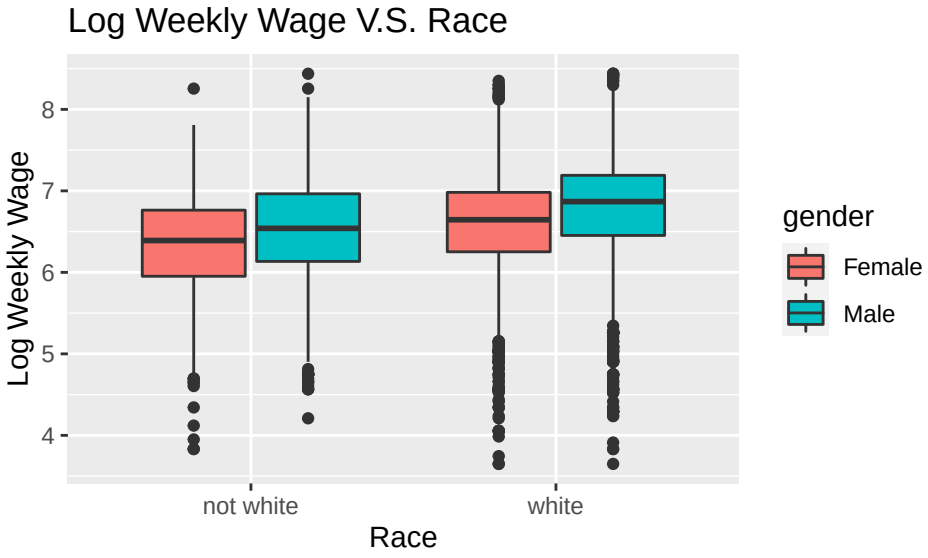
The county means of observed `Y` are distributed around 6.5. It is reasonable to assume that these means come from the same distribution. This justifies our hierarchical structure on county level mean of `Y`. Note that in this plot, the County code is jittered to make the observations from each county more visible.



In this education plot, the 4 education levels means:

- 0: does not have a high school diploma
- 1: have a high school diploma but does not have a Bachelor's degree
- 2: Bachelor's degree
- 3: Master degree or higher

we can see a linear relationship between the education and weekly wage. Moreover, the boxplot shows a linear relationship between education level and log weekly wage. An interesting observation is that when the level of education is raised by one level, the increase in log weekly wages is almost constant. Therefore, we might treat this variable as a continuous variable in our model.



Differences in weekly log wage income exists between white and non-white races. Therefore, the `RACWHT` variable is taken as an independent variable in our model (This will be expressed by a dummy variable). This difference is almost the same within across gender; therefore, no iteration term between age and race is introduced in our model.

The plot of dependent variable Y against age group is similar to the plot above, so I move it to the appendix.

3 Methods

In this report, I use multiple linear regression through Bayesian method, with the log weekly income as the response variable, and the possible influential factors as predictor variables. Moreover, a hierarchy structure is introduced to model the county mean of weekly income wage.

Mathematically, the model can be written as:

$$\begin{aligned} y_i|\eta_{c[i]}^{\text{county}} &\sim N\left(\beta_0 + \eta_{c[i]}^{\text{country}} + \textstyle\sum_{j=1}^{6} \beta_j x_{i,j}, \sigma_y^2\right) \\ \eta_c^{\text{country}} &\sim N\left(0, \left(\sigma_\eta^{\text{country}}\right)^2\right), \text{ for } c = 1, 2, \ldots, C. \end{aligned}$$

where

- y_i the log weekly income of the ith respondents
- $x_{i,1}$ gender of the ith respondents
  - 0: the respondent is male
  - 1: the respondent is female
- $x_{i,2}$ the education level of the ith respondents (treated as a continuous variable)
  - 0: does not have a high school diploma
  - 1: have a high school diploma but does not have a Bachelor's degree
  - 2: Bachelor's degree
  - 3: Master degree or higher
- $x_{i,3}$ race of the ith respondents is white or not
  - 0: the respondent is not white
  - 1: the respondent is white
- $x_{i,4}$ indicator that the ith respondents is 30 years old or older
  - 0: the respondent is younger than 30
  - 1: the respondent is older than 30 (include 30)
- $x_{i,5}$ the usual weekly work hours of the ith respondent
- $x_{i,6}$ the occupation income score, reflecting the median of the income of this occupation
- $\eta_{c[i]}^{\text{county}}$ the county level derivation of log weekly wage income in the county where the ith respondent worked.

Note that the continuous independent variables except the `education` are standardized to make the Bayesian model converges faster.

The reason for choosing this model is: by the density plot from the EDA part, the distribution of dependent variable Y (log weekly wage income) is approximately normal. And we see a linear relationship between the independent variables and Y in EDA. Therefore, a multiple linear regression is a natural choice. The hierarchy structure is introduced, because slight difference exist in county mean. But the sample size within some county is small. Under this case, hierarchy model is more effective than estimating the mean within each county independently. People from the same county may have similarity in the income structure; therefore, it is more appropriate to use hierarchical model on county means instead of taking the population mean for all counties.

In order to apply the Bayesian method, we have to set the priors for all parameters in this model:

- $\beta_0, \beta_1, \ldots, \beta_6 \sim N(0, 1)$,
- $\eta_c^{\text{county}} \sim N(0, 1)$,
- $\sigma_y, \sigma_\eta^{\text{country}} \sim N_+(0, 1)$.

Then the model is fitted using the `RStan` package, and it uses MCMC sampling.

4 Results

4.1 Estimated Coefficients

Using the mean as the estimated coefficients, and using the equi-tailed 95% credible interval, we have:

	estimated coefficients	95% CI
beta0	5.995	(5.94 , 6.05)
beta1	-0.1435	(-0.171 , -0.117)
beta2	0.2679	(0.25 , 0.286)
beta3	0.205	(0.169 , 0.243)
beta4	0.1551	(0.129 , 0.181)
beta5	0.2108	(0.197 , 0.224)
beta6	0.08835	(0.074 , 0.103)

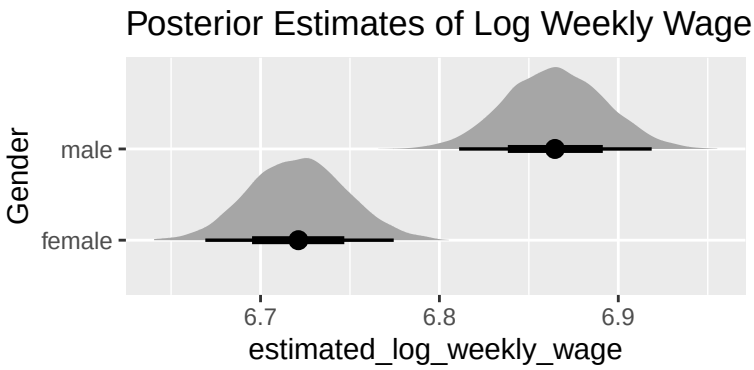
Interpretation of the coefficients:

- β_0: Estimated population mean of log weekly wage income of male, non-white respondents with average usual weekly working hours, average occupation scores and without high school diploma.
- β_1: When the other independent variables are exactly the same, the weekly wage income of female is $\exp(-0.1435) = 0.866$ $(0.843, 0.890)$ times of that of non-white people.
- β_2: If the respondent's education increased by one level, then his/her weekly wage income is expected to increase by $\exp(0.2679) = 1.307$ $(1.284, 1.331)$.
- β_3: When the other independent variables are exactly the same, the weekly income of white people is $\exp(0.250) = 1.284$ $(1.184, 1.275)$ times of that of non-white people.
- β_4: When the other independent variables are exactly the same, the weekly income of people older than 30 is $\exp(0.1551) = 1.168$ $(1.138, 1.198)$ times of that of people younger than 30.
- β_5: When the usual worked hours is increased by one standard derivation from the average, the weekly wage income is expected to increase by $\exp(0.2108) = 1.235$ $(1.218, 1.251)$.
- β_6: An interesting observation is that the influence of occupation income score on weekly income is not as large as one might expect. Occupation income score is determined by the median of the income for this occupation. With other factors fixed, choosing an occupation that has 1 standard derivation income score higher than the average, the weekly wage income will increase by $\exp(0.08835) = 1.092$ $(1.077, 1.108)$.

This illustrate that despite the influence of education, race, age, weekly working hours as well as the occupation income score, the gender pay gap still exists. Moreover, seeking an occupation that has high median pay does not contribute to the income as much as one might expect. Trying to get a higher education degree will help to increase the weekly wage income a lot.

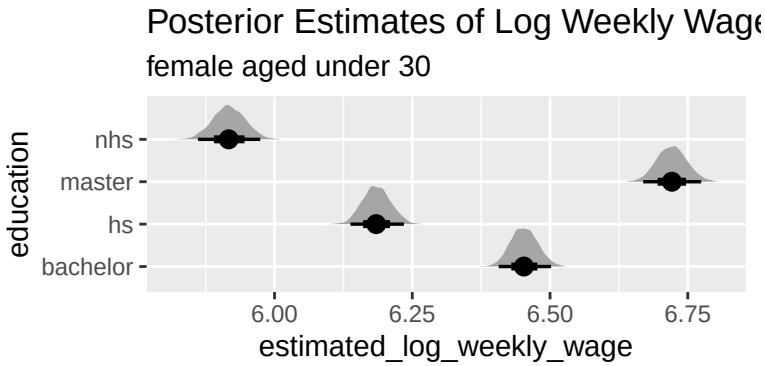
The more detailed information about posterior distribution of coefficients, and the prior v.s. posterior density plots of all parameters are in the Appendix.

More specifically, we can look at the posterior estimated mean of log weekly income non-white people of different gender but with average weekly working hours, average occupation income score, with a Bachelor's degree and of the same age group working in County Hamilton. We can see a significant pay gender gap.



The same trend will be see for the white race group, and the elder age group, since the model used does not contain an iterative term between `age`, `race`, and `gender`.

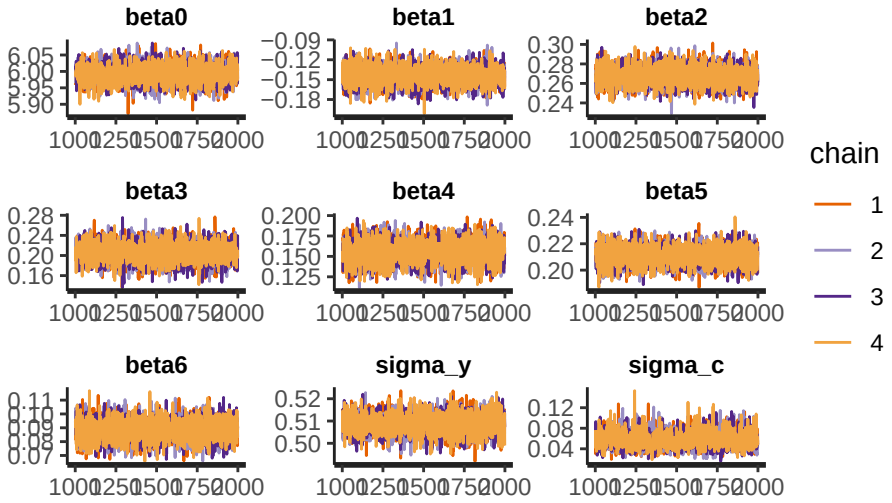
The posterior estimated mean of log weekly income non-white females with average weekly working hours, average occupation income score,and of the same age group working in County Hamilton indicates a significant pay gap in different education level.



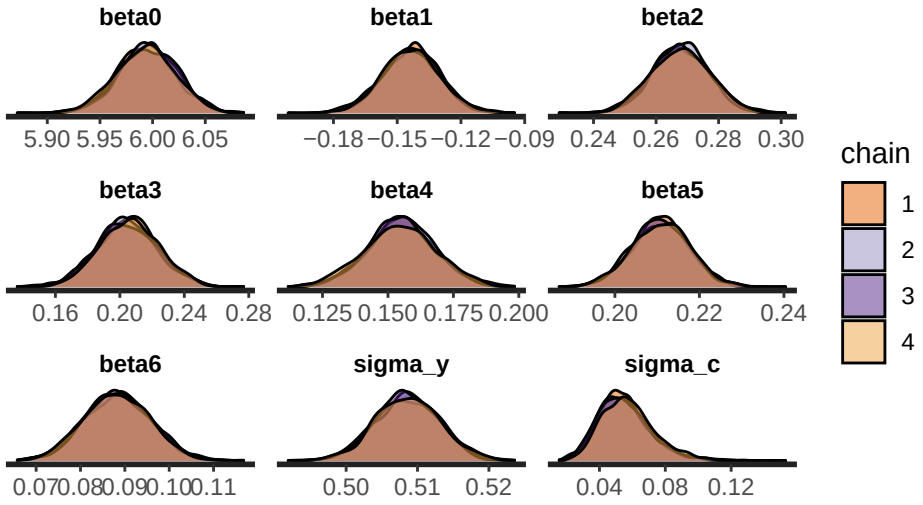
The similar plot in education level indicates that

4.2 Diagnostics

We have to first check that the samples converges well. It can be checked through the trace plots and the density plots of four chains for all parameters:

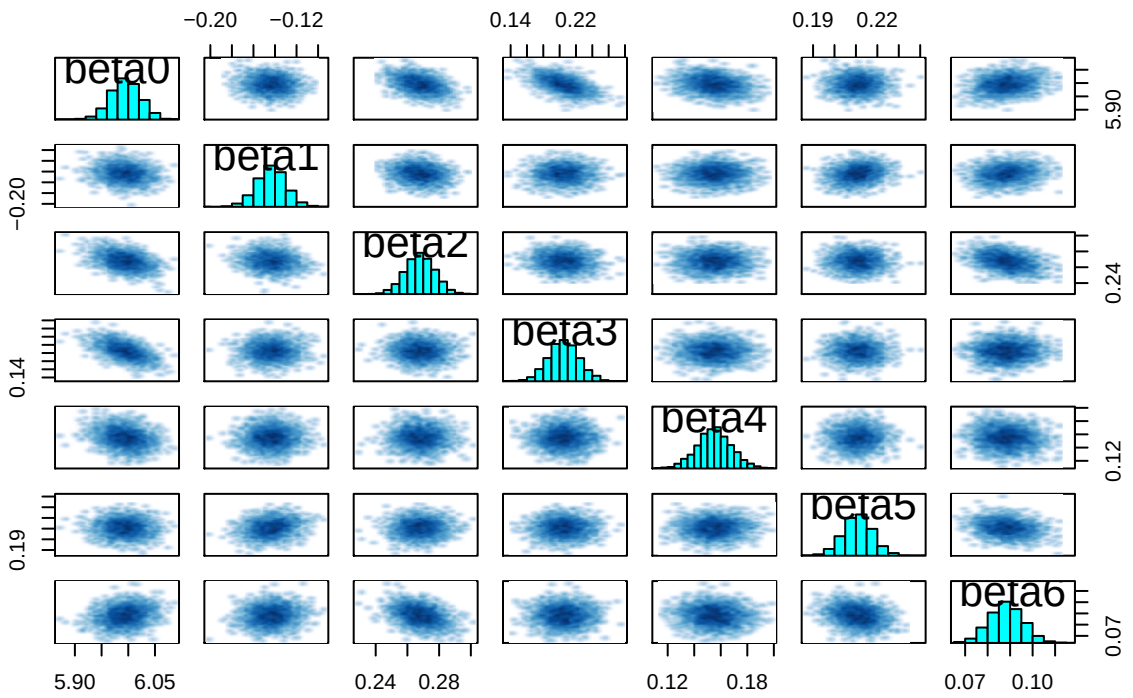


The trace plots of parameters are given as above. We can see that the four chains are well mixed, which indicates that our model converges well.



The trace plots together with the density plots above show that the sampling from the `Stan` model converges and the four chains are mixed well. This means that the model converges to a target distribution as we expected.

In addition, the target posterior distribution of all coefficients can be checked through the paired joint density:



From the plot above, we can see that the paired posterior joint distributions of all coefficients are reasonable, since the scatter plot above is cloud-like as we expected.

4.3 Compare Models

An alternative model:

Introduce hierarchical structures on County level, age level and race level.

$$
\begin{aligned}
y_i|\eta_{c[i]}^{\text{county}} &\sim N\left(\beta_0 + \eta_{c[i]}^{\text{country}} + \textstyle\sum_{j=1}^{4}\beta_j x_{i,j}, \sigma_y^2\right) \\
\eta_c^{\text{country}} &\sim N\left(0, \left(\sigma_\eta^{\text{country}}\right)^2\right), \text{ for } c = 1, 2, \ldots, C. \\
\eta_a^{\text{age}} &\sim N\left(0, \left(\sigma_\eta^{\text{age}}\right)^2\right), \text{ for } a = 1, 2, \ldots, A. \\
\eta_r^{\text{race}} &\sim N\left(0, \left(\sigma_\eta^{\text{race}}\right)^2\right), \text{ for } r = 1, 2, \ldots, R.
\end{aligned}
$$

where

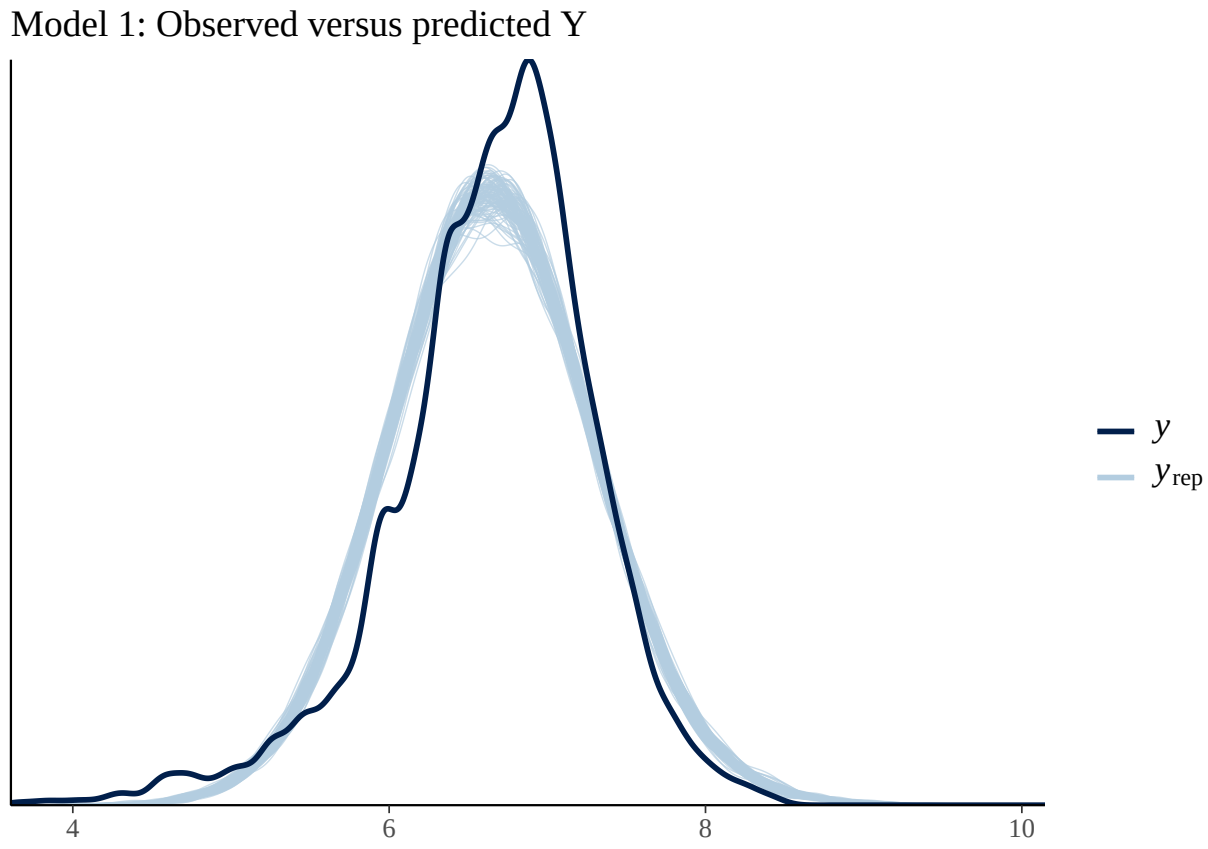
- y_i the log weekly income of the ith respondents
- $x_{i,1}$ gender of the ith respondents
  - 0: the respondent is male
  - 1: the respondent is female
- $x_{i,2}$ the education level of the ith respondents (treated as a continuous variable)
  - 0: does not have a high school diploma
  - 1: have a high school diploma but does not have a Bachelor's degree

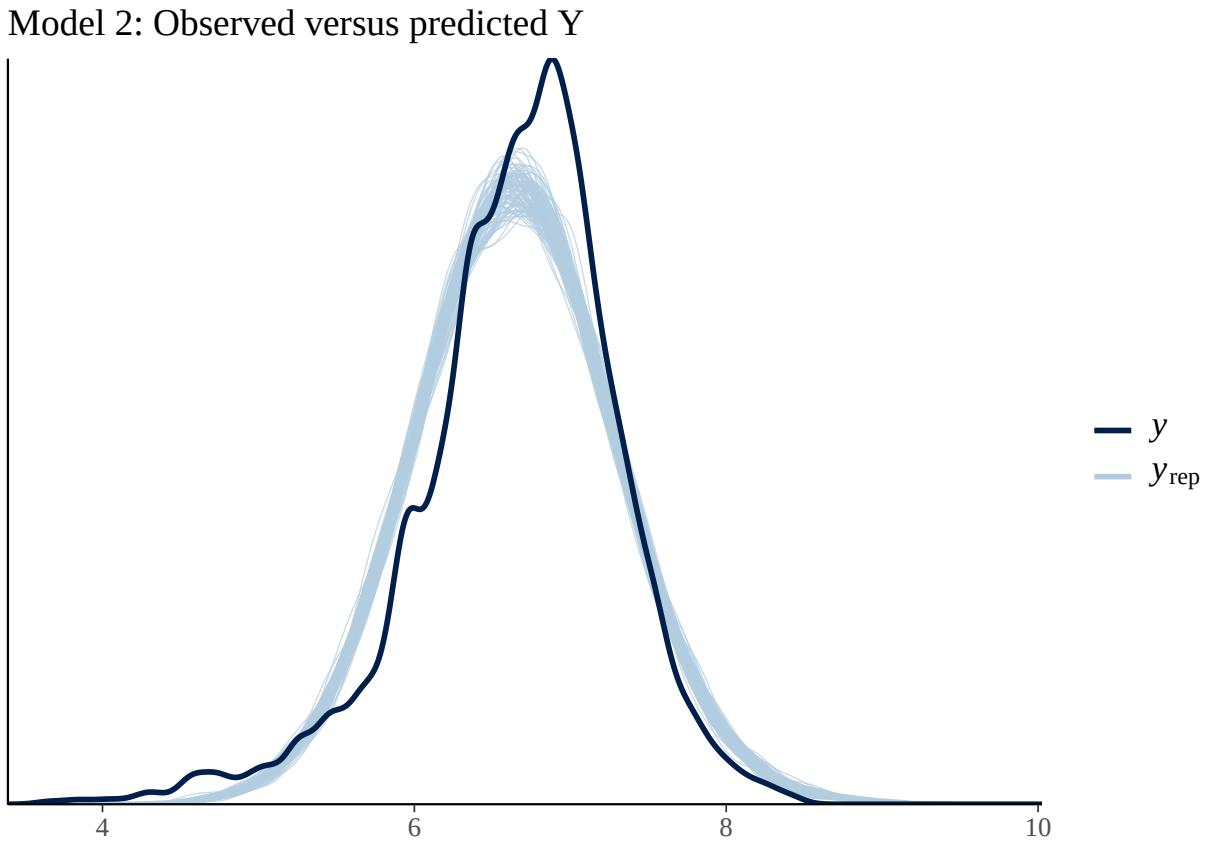
  - 2: Bachelor's degree
  - 3: Master degree or higher
- $x_{i,3}$ the usual weekly work hours of the ith respondent
- $x_{i,4}$ the occupation income score, reflecting the median of the income of this occupation
- $\eta_{c[i]}^{\text{county}}$ the county level derivation of log weekly wage income
- $\eta_{a[i]}^{\text{age}}$ the age level derivation of log weekly wage income
- $\eta_{r[i]}^{\text{race}}$ the race level derivation mean of log weekly wage income

Set the priors for all parameters in this model:

- $\beta_0, \beta_1, \ldots, \beta_6 \sim N(0, 1)$,
- $\eta_c^{\text{county}}, \eta_a^{\text{age}}, \eta_r^{\text{race}} \sim N(0, 1)$,
- $\sigma_y, \sigma_\eta^{\text{country}}, \sigma_\eta^{\text{age}}, \sigma_\eta^{\text{gender}} \sim N_+(0, 1)$.

One way to compare the fitness of the two models is to compare the replicated predicted density of independent variable to its observed density. The predicted density and observed density of our dependent variable Y (log weekly wage income) are given as below:



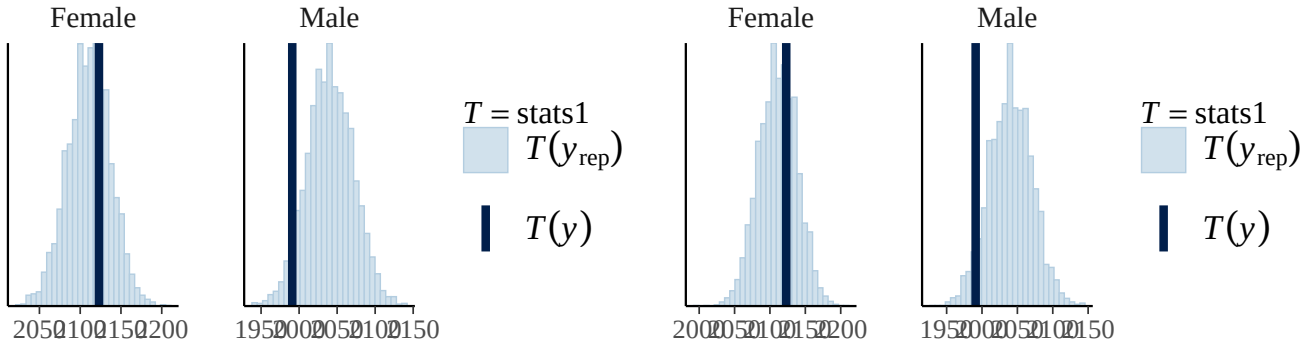


This compares our dataset with 100 replicates from our model. The trend of the replicated predicted densities from two models are similar and both are slightly different from the observed density of Y: the observed density has higher mode, and some fluctuations.

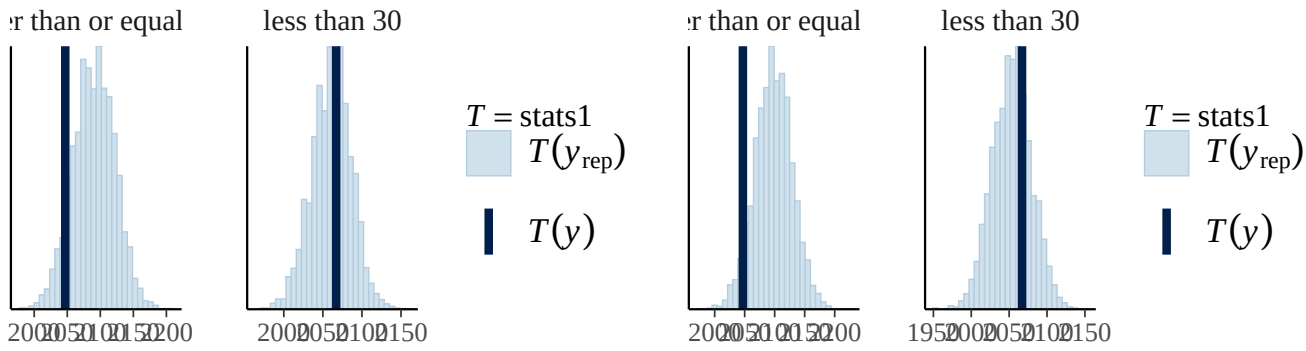
Moreover we can introduce test statistic T:

$$T = \text{proportion of people with weekly income less than 1000 dollars}$$

The density of test statistic T by gender groups:



The density of test statistic T by age groups:



The plot on the left is from Model 1 and the plot on the right is from Model 2. By the test statistic, both models seem to be fine. We cannot make statement about which model work better through this test statistic.

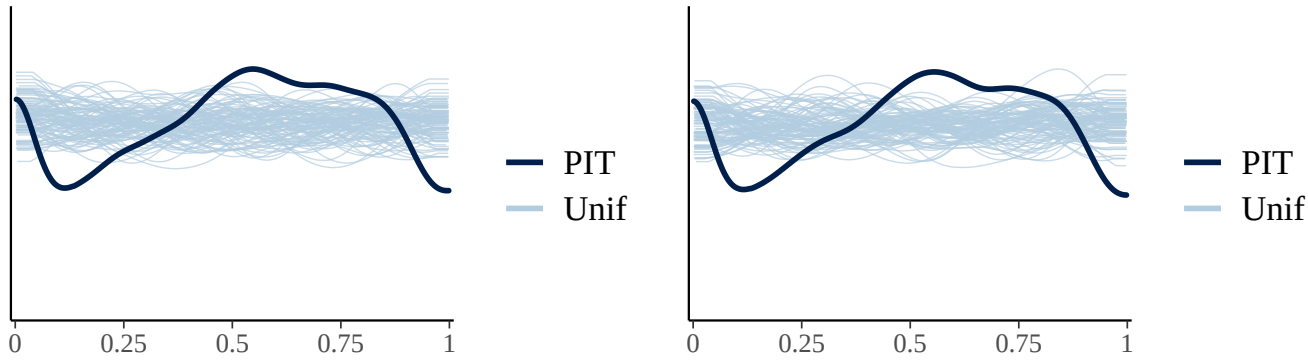
The Leave-one-out expected log pointwise predictive density can be used to compare the fitness of two models. But the table below still does not give enough information of which model works better.

Table 4: Table continues below

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo
model1	0	0	-4694	83.69	24.08	1.001
model2	-0.489	5.364	-4695	84.23	38.92	1.437

	looic	se_looic
model1	9388	167.4
model2	9389	168.5

Another way is to use leave-one-out probability integral transform (PIT). In order to get a sense of how well the model fits the data, one can compare the PIT to standard uniform distribution:



The PIT curves from two models are similar. The PIT curve fluctuates around the standard uniform distributions. From the diagnostic test above, we can see that our model works fine for fitting the data. However, more independent variables should be added into the model to improve the fitness. See more details in discussion part. (The summary of leave-one out expected pointwise log density is in the appendix.)

In fact, we cannot determine which model fits the dataset better. However, the fitting process of Model 2 is much more time-consuming compared to Model 1. This is the reason that we choose Model 1 over Model 2.

5 Discussion

From the Results part, we can see that given the same education level, the same race, the same age group, the same usual weekly work hours and the same occupation income score, gender is still a strong influential factor contributing to the weekly wage income. Women are paid 86.6/ (84.3%, 89.0%) of the men, taking into account the education level, the hours worked per week, the occupation income score, race and age.

The result also indicates that the occupation one chooses and the education one receives also influence the weekly wage income. However, the occupation income score does not have as large impact as one might expect. Choosing an occupation with 1 standard derivation higher only result in 1.092 (1.077, 1.108) increase in weekly wage income. It is more effective to try to pursue a higher degree, than to choose an occupation that has high median pay, in order to get a higher wage.

However, I am not satisfied with this model. The following independent variables can be introduced to make the model works better:

- the proficiency of the respondents in terms of his/her occupation,
- the time when the respondents started to take his/her current job,
- the rank of university that the respondents graduated from.

However, the IPUMS does not provide any variables described above.

Another interesting question is: do personality traits factors influence the income.Jordan B. Peterson states that the gender pay gap is not caused by gender discrimination; an important factor is the gender personality traits difference. (Jordan 2018). This brings a widely spread discussion. I wonder to what extent the personality treats will contribute to weekly income and if we include the personality traits variable in our model, then after controlling all the other independent variables, is gender still a statistically significant influential contributor to weekly wage income. Unfortunately, the personality traits are not provided in IPUMS. So I could not include it in this model.

Further analysis can also be down by applying a principal component analysis (PCA) before fitting the models. Since this model only contains 6 independent variables, I did not apply PCA. But if in future, the variables described above can be collected and added into the model, a PCA should be down to get better result.

Reference

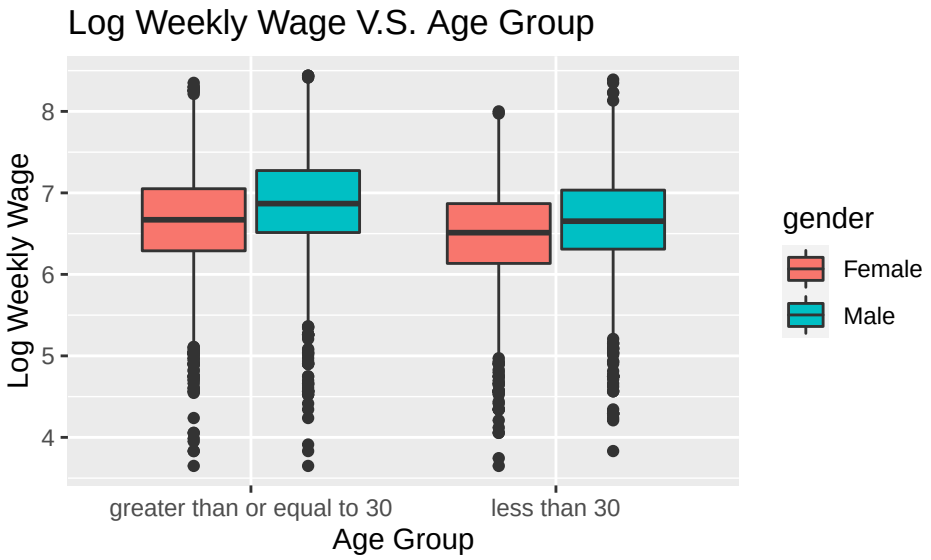
Fontenot, K., Semega, J., & Kollar, M. (2018). Income and Poverty in the United States: 2017. Washington: U.S. Census Bureau.

Jordan B. Peterson. (2018). 12 Rules for Life: An Antidote to Chaos. Random House Canada

Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0

Appendix

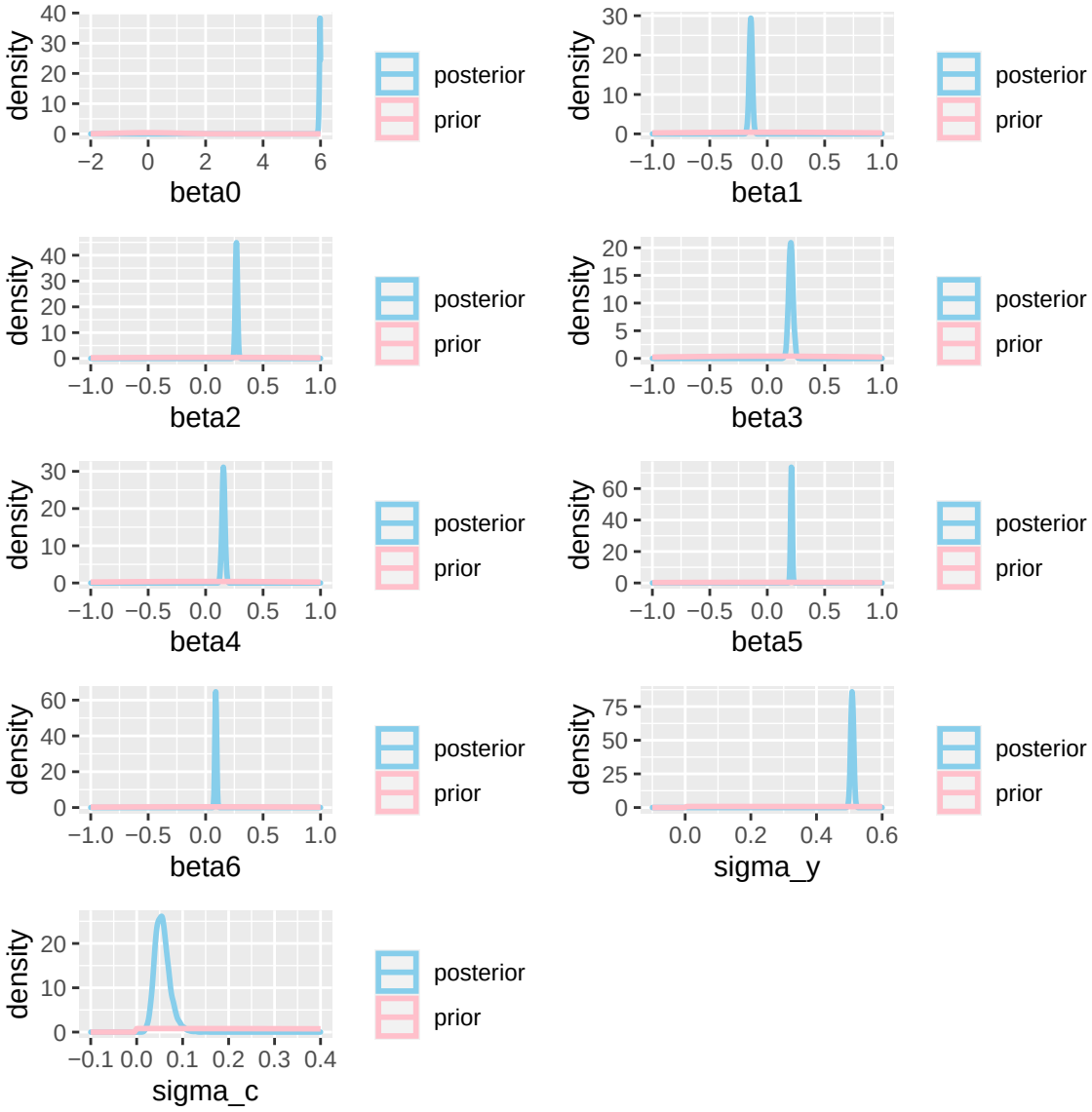
The log weekly wage income against the age group:



The summary of statistics for coefficients in the model:

	mean	sd	2.5%	50%	97.5%	n_eff	Rhat
beta0	5.995	0.02784	5.94	5.995	6.05	1701	1.002
beta1	-0.1435	0.01343	-0.1706	-0.1433	-0.1173	4960	1
beta2	0.2679	0.009192	0.2498	0.2679	0.286	3975	1
beta3	0.205	0.01881	0.1687	0.2049	0.243	3319	1.001
beta4	0.1551	0.01302	0.1294	0.1551	0.1813	4178	1
beta5	0.2108	0.006796	0.1973	0.2109	0.2241	4589	0.9996
beta6	0.08835	0.007435	0.07408	0.08835	0.1027	4297	1
sigma_y	0.5086	0.004495	0.4997	0.5086	0.5174	5212	0.9998
sigma_c	0.05586	0.01613	0.02948	0.05429	0.09296	1078	1.008

The prior and posterior density plots of parameters in this model:



The Leave-one-out expected log pointwise predictive density of this model:

	Estimate	SE
elpd_loo	-4694	83.69
p_loo	24.08	1.001
looic	9388	167.4

	Estimate	SE
elpd_loo	-4695	84.23
p_loo	38.92	1.437
looic	9389	168.5