Report of MA678 Midterm Project

Zijia Wang

2022-12-04

Abstract

People intuitively associate their wage level with their cost of living. Since the purchasing power of the currency unit varies from country to country, their cost of living also varies. This paper utilizes a commonly accepted linear mixed model with group level `region` to capture the relationship between average monthly salary and factors of cost of living by using almost 5000 cities across the world. We found that the cost of living, such as food, oil and rent, varies significantly depending on the region. But within the same region, there is no big difference between cities, they are all at the same level. Secondly, people's wage level and living standard consumption do significantly correlate. There is also an interaction between the various components of the cost of living.

Keywords: Multilevel model; cost of living; salary.

1. Introduction

According to historical articles indicated in (Grimes, Prime, & Walker, 2019), in the United States, urban leaders in various cities struggle to create conditions that improve job opportunities and raise incomes. And the paper's research proves that urban areas are more productive, which leads to higher wages, despite the higher cost of living One of the most critical variables to consider when assessing wages in different geographic areas is the regional cost of living. In the absence of regional studies on how the cost of living affects wage levels, our paper here offers four contributions: The first is to study a linear model of the cost of living and wage levels under a global scenario. The second is to find out the influences of fixed effects (e.g. basic foods, gasoline, rent and so on) and random effects (region/country). Finally, we examine a linear mixed model of cost of living and wage levels at a regional scale. The remaining parts of this paper present the methodology of the study, a evaluation, the empirical results, a discussion.

2. Data and Methodology

2.1 data wrangling

I found the data set from a public website (https://www.kaggle.com/datasets/mvieira101/global-cost-of-living).

Here is the data dictionary.

Column	Description
city	Name of the city
region	Name of the country

Column	Description
x1	Meal, Inexpensive Restaurant (USD)
x2	Meal for 2 People, Mid-range Restaurant, Three-course (USD)
x3	McMeal at McDonalds (or Equivalent Combo Meal) (USD)
x4	Domestic Beer (0.5 liter draught, in restaurants) (USD)
x5	Imported Beer (0.33 liter bottle, in restaurants) (USD)
x6	Cappuccino (regular, in restaurants) (USD)
x7	Coke/Pepsi (0.33 liter bottle, in restaurants) (USD)
x8	Water (0.33 liter bottle, in restaurants) (USD)
x9	Milk (regular), (1 liter) (USD)
x10	Loaf of Fresh White Bread (500g) (USD)
x11	Rice (white), (1kg) (USD)
x12	Eggs (regular) (12) (USD)
x13	Local Cheese (1kg) (USD)
x14	Chicken Fillets (1kg) (USD)
x15	Beef Round (1kg) (or Equivalent Back Leg Red Meat) (USD)
x16	Apples (1kg) (USD)
x17	Banana (1kg) (USD)
x18	Oranges (1kg) (USD)
x19	Tomato (1kg) (USD)
x20	Potato (1kg) (USD)
x21	Onion (1kg) (USD)
x22	Lettuce (1 head) (USD)
x23	Water (1.5 liter bottle, at the market) (USD)
x24	Bottle of Wine (Mid-Range, at the market) (USD)
x25	Domestic Beer (0.5 liter bottle, at the market) (USD)
x26	Imported Beer (0.33 liter bottle, at the market) (USD)
x27	Cigarettes 20 Pack (Marlboro) (USD)
x28	One-way Ticket (Local Transport) (USD)
x29	Monthly Pass (Regular Price) (USD)
x30	Taxi Start (Normal Tariff) (USD)
x31	Taxi 1km (Normal Tariff) (USD)
x32	Taxi 1hour Waiting (Normal Tariff) (USD)
x33	Gasoline (1 liter) (USD)
x34	Volkswagen Golf 1.4 90 KW Trendline (Or Equivalent New Car) (USD)
x35	Toyota Corolla Sedan 1.6l 97kW Comfort (Or Equivalent New Car) (USD)
x36	Basic (Electricity, Heating, Cooling, Water, Garbage) for 85m2 Apartment (USD)
x37	1 min. of Prepaid Mobile Tariff Local (No Discounts or Plans) (USD)
x38	Internet (60 Mbps or More, Unlimited Data, Cable/ADSL) (USD)
x39	Fitness Club, Monthly Fee for 1 Adult (USD)
x40	Tennis Court Rent (1 Hour on Weekend) (USD)
x41	Cinema, International Release, 1 Seat (USD)
x42	Preschool (or Kindergarten), Full Day, Private, Monthly for 1 Child (USD)
x43	International Primary School, Yearly for 1 Child (USD)
x44	1 Pair of Jeans (Levis 501 Or Similar) (USD)
x45	1 Summer Dress in a Chain Store (Zara, H&M, . . .) (USD)
x46	1 Pair of Nike Running Shoes (Mid-Range) (USD)
x47	1 Pair of Men Leather Business Shoes (USD)

Column	Description
x48	Apartment (1 bedroom) in City Centre (USD)
x49	Apartment (1 bedroom) Outside of Centre (USD)
x50	Apartment (3 bedrooms) in City Centre (USD)
x51	Apartment (3 bedrooms) Outside of Centre (USD)
x52	Price per Square Meter to Buy Apartment in City Centre (USD)
x53	Price per Square Meter to Buy Apartment Outside of Centre (USD)
x54	Average Monthly Net Salary (After Tax) (USD)
salary	Mortgage Interest Rate in Percentages (%), Yearly, for 20 Years Fixed-Rate

Firstly, I download cost of living data and wiped out 'NA' data. Then, I rename country column as region to avoid politic dispute. Then I chose the data that only countries has more than 30 cities, because Associated with the classical between-subjects design ANOVA, we generally require that the sample size of each group should not be too small, at least 30~50 subjects to ensure power. The suitable regions are as follows. *China, Mexico, Brazil, Germany, Spain, United Kingdom, Canada, Italy, Russia, United States, India.* Additionally, after I analyze the statistics of these data, I find that they are not in the same scale, hence, I normalize these data at the scale of 0 to 1.

Secondly, I use Lasso Regression to chose variables that is fit in my linear regression model, according to the outcome of lasso regression, by deleting the variables that doesn't display coefficients.

Hence, the variables x5, x6, x7, x8, x9, x12, x15,x16, x18, x22, x24, x25, x26, x28, x31, x32, x33, x37, x38, x39, x41, x42, x44, x47, x48, x49, x51 x53, and x55 are suitable. However, they are many variables having the same meaning and belongs to one aspect, and I will chose 3 independent variables , **basic food**, **gasoline**, **rent**, because I found it is hard for me to identify random slope and intercept if it I want to analyze 10 variables' interaction and their fixed effects and random effects.

2.2 Exploratory Data Analysis

Figure 1 shows the salary distribution of different country. From this figure, we can find that the United States has a highest based salary among these countries.

Figure 2 shows the linear regression lines between different factors of cost of living and salary. It is easy to find that basic food have varying intercept but with roughly similar slopes. As for gasoline,it has different intercept and slopes. And for rent price factor,it has different intercept but obviously similar slopes.

3. Evaluation

3.1 Model fitting

I use method ANOVA to compare different models I built, because I want to include the variables' interactions effect in my linear model, which also is the level 1 model. According to the results, we find that interactions **basic_food:gasoline** and **basic_food:rent** have significant influence in this model.

Hence, the final linear model I build is:

```r
lm1 <- lm(salary ~ basic_food + gasoline + rent + basic_food:gasoline + basic_food:rent, data = df5)
```

Here is the summary table of my basic model of multilevel model and all variables here are considered as statistically significant at $\alpha = 0.5$ level.

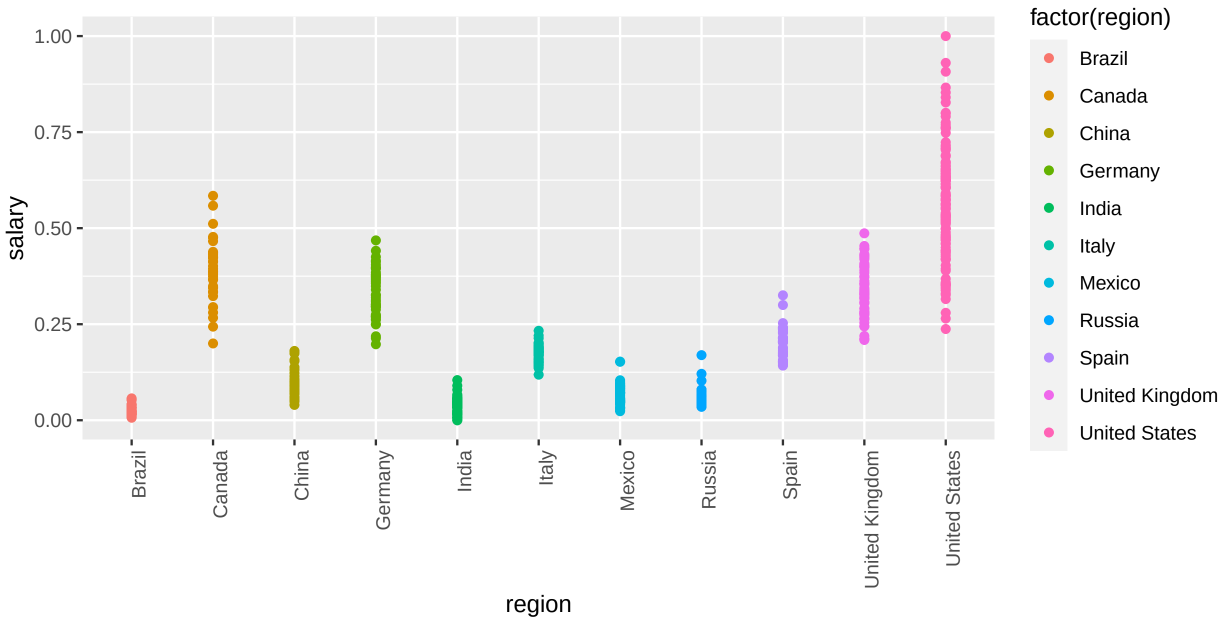

Figure 1: salary distribution of different regions

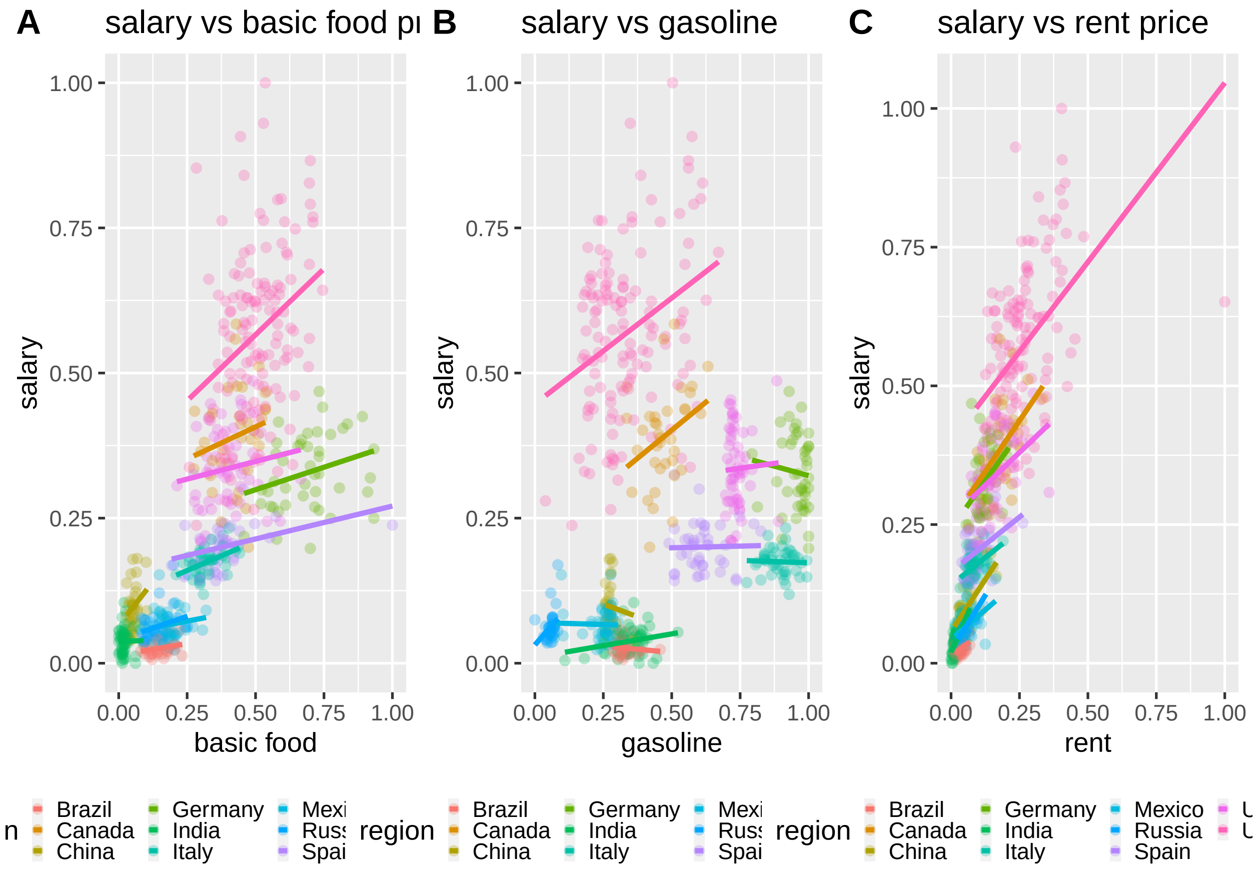

Figure 2: factors of cost of living vs salary

Coefficients:				
	Estimate	**Std.Error**	**t value**	**Pr(>|t|)**
(Intercept)	-0.08202	0.01426	-5.752	1.45e-08 ***
basic_food	0.79679	0.05432	14.668	< 2e-16 ***
gasoline	0.09694	0.03453	2.807	0.005172 **
rent	1.50005	0.14244	10.53	< 2e-16 ***
basic_food:gasoline	-0.54595	0.08013	-6.814	2.46e-11 ***
basic_food:rent	-0.90160	0.25473	-3.539	0.000434 ***

According the EDA graphs we build the multilevel model:

```
lm2 <- lmer(salary ~ basic_food + gasoline + rent + basic_food:gasoline + basic_food:rent + (1 +gasolin
```

```
## boundary (singular) fit: see help('isSingular')
```

Here is the summary of model (fixed effect) but only two variables here are considered as statistically significant at $\alpha = 0.5$ level. To be more clear, a fixed effect parameters are also include in Figure 3. Figure 4 shows a random effect plot for **Region** level.

Fixed effects:					
	Estimate	Std. Error	df	t value	Pr(>|t|
(Intercept)	0.06050	0.04116	30.81786	1.470	0.151690
basic_food	0.29788	0.08831	184.13030	3.373	0.000906 ***
gasoline	0.06223	0.06972	89.53833	0.893	0.374463
rent	0.70422	0.14975	516.28283	4.703	3.3e-06 ***
basic_food:gasoline	-0.17957	0.12038	373.86944	-1.492	0.136640
basic_food:rent	-0.35740	0.26410	502.84945	-1.353	0.176569

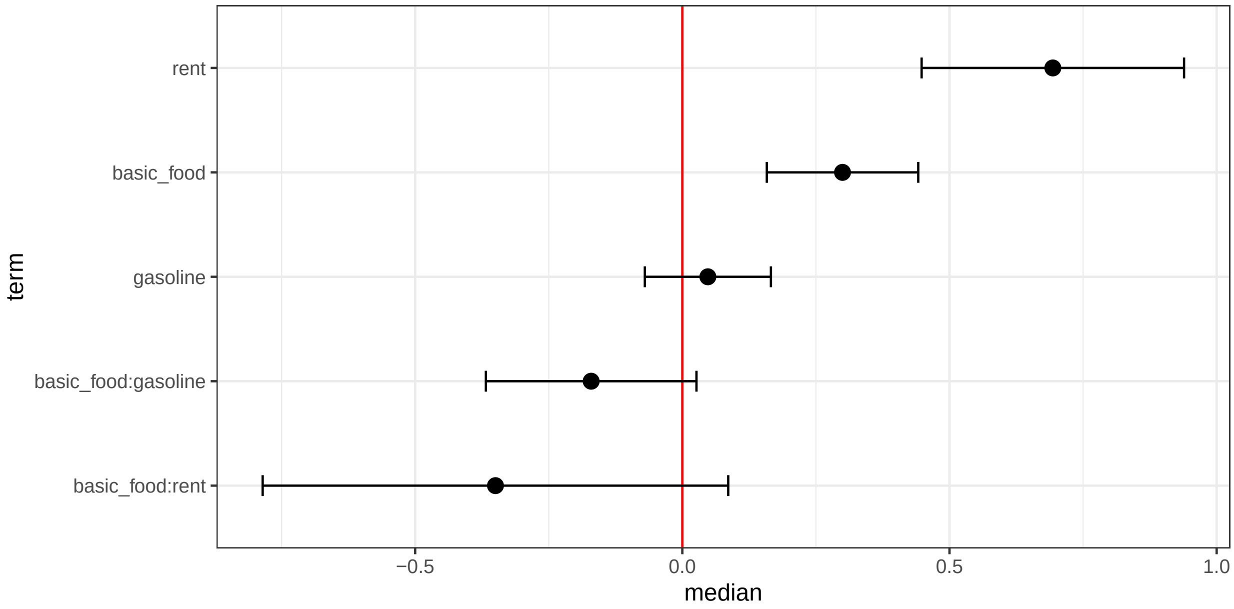

Figure 3: Fixed Effect of cost-living Model

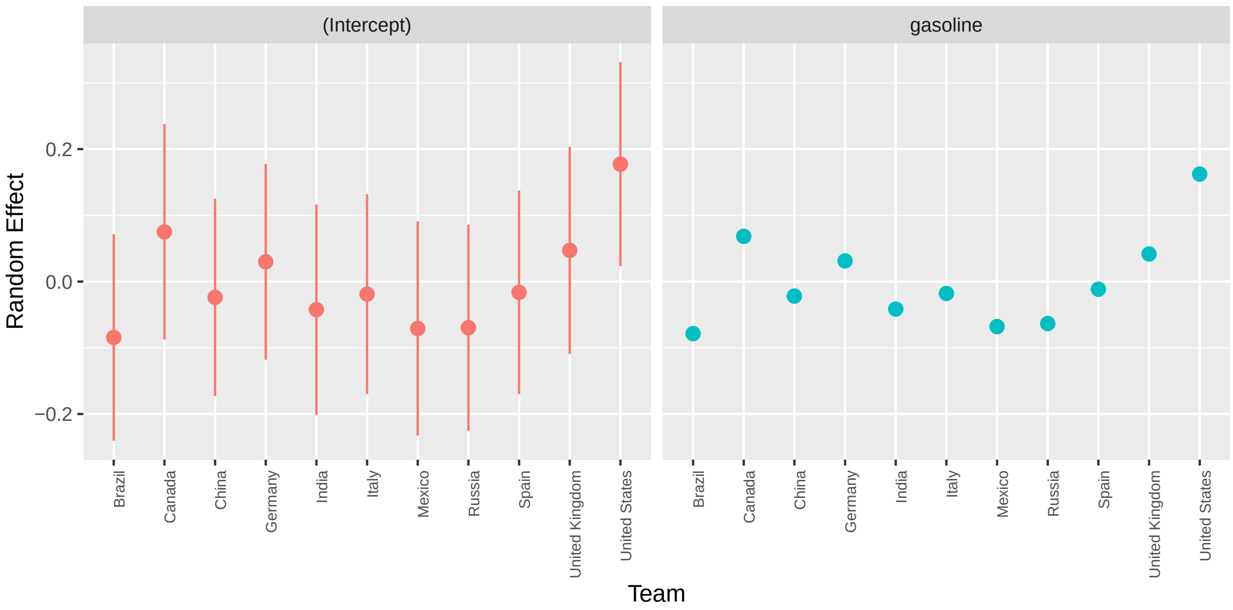

Figure 4: Random Effect of cost-living Model

3.2 Model checking

We check the two model by plotting their residuals. Figure 5 is residual plot and residual Q-Q plot of linear model(lm1). Figure 6 is residual plot and residual Q-Q plot of linear mixed model(lm2). According to these two residual plots, we can easily find that the second residual plot is better, The reasons are as followed. Firstly, it is very symmetrically distributed and tend to cluster in the middle of the plot. Secondly, they cluster around the lower digits of the y-axis (e.g., 0.5 or 1.5). Finally, there is no clear pattern. As for these two Normal Q-Q plots, we find that the second model away from a Normal distribution. This means our data have more extreme values than we expected which may caused by different regions .

4. Result

Firstly, we interpret the model we have. We are able to get the following formula of fixed effect:

$salary = -0.08202+0.79679*basic_food+0.09694*gasoline+1.50005*rent+(-0.54595)*basic_food*gasoline+(-0.9016)*basic_$

Then add the random effect to the intercepts and slopes and get the estimated formula:

$salary = 0.0605+0.29788*basic_food+0.06223*gasoline+0.70422*rent+(-0.17957)*basic_food*gasoline+(-0.35740)*basic_fo$

Then, we predict the multilevel model about each distinct factors and salary. And we only present the graph Figure 7, which is about rent price and salary here. Additionally, the following part is to visualize the information pool of gasoline, because only gasoline factor has not only varying intercept and varying slope. So we can build complete pooling, partial pooling and no pooling. Based on the information pool Figure, we can see gasoline price is centered on the U.S oil price, and China and Germany is out of the pool. It is coincident with the fact in the world nowadays. The U.S. economy is firmly tied to oil and the U.S. dollar, followed by the entire world's monetary system is the dollar standard, we can then conclude that the world's oil prices are inseparable from the U.S. oil prices

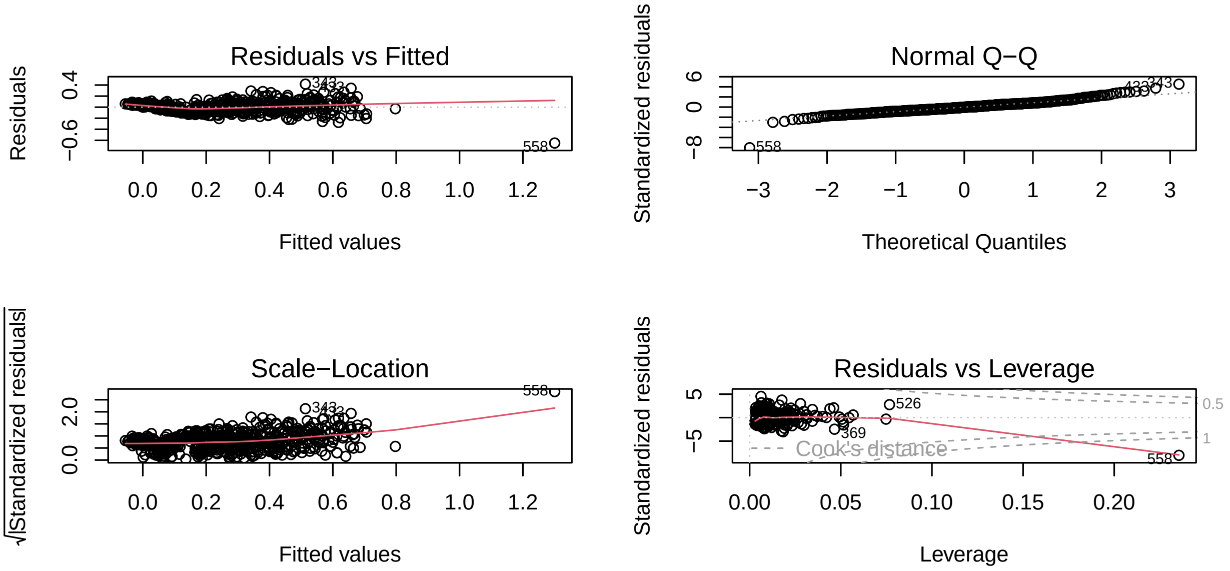

Figure 5: model checking plots of level 1 model

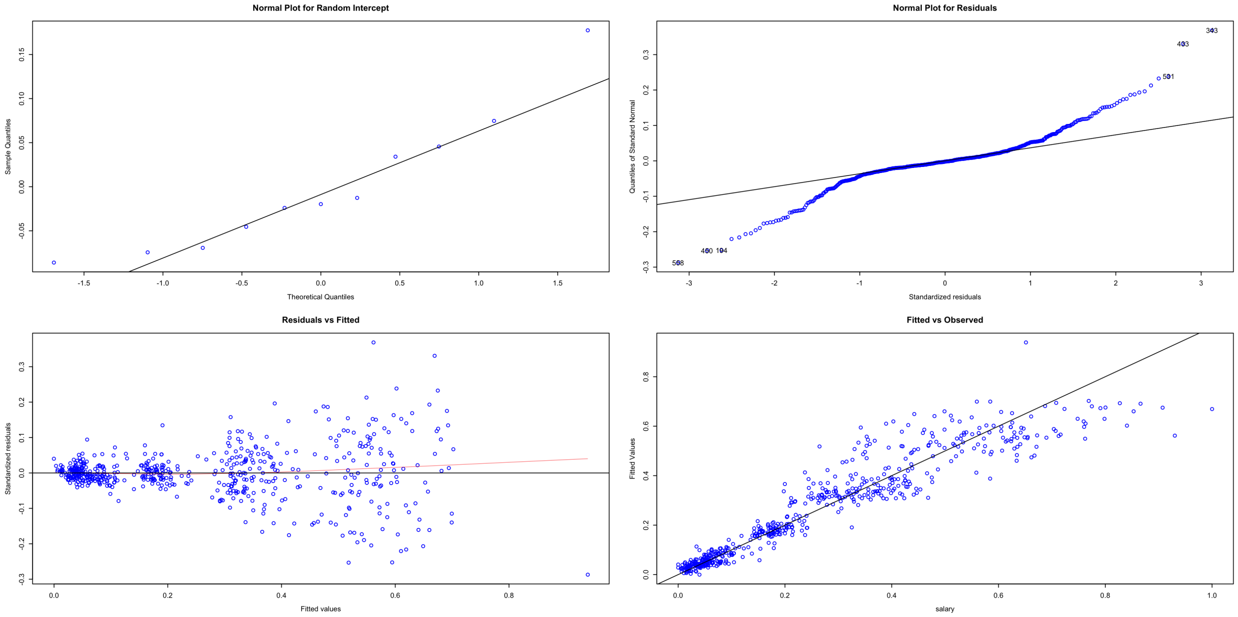

Figure 6: model checking plots of level 2 model

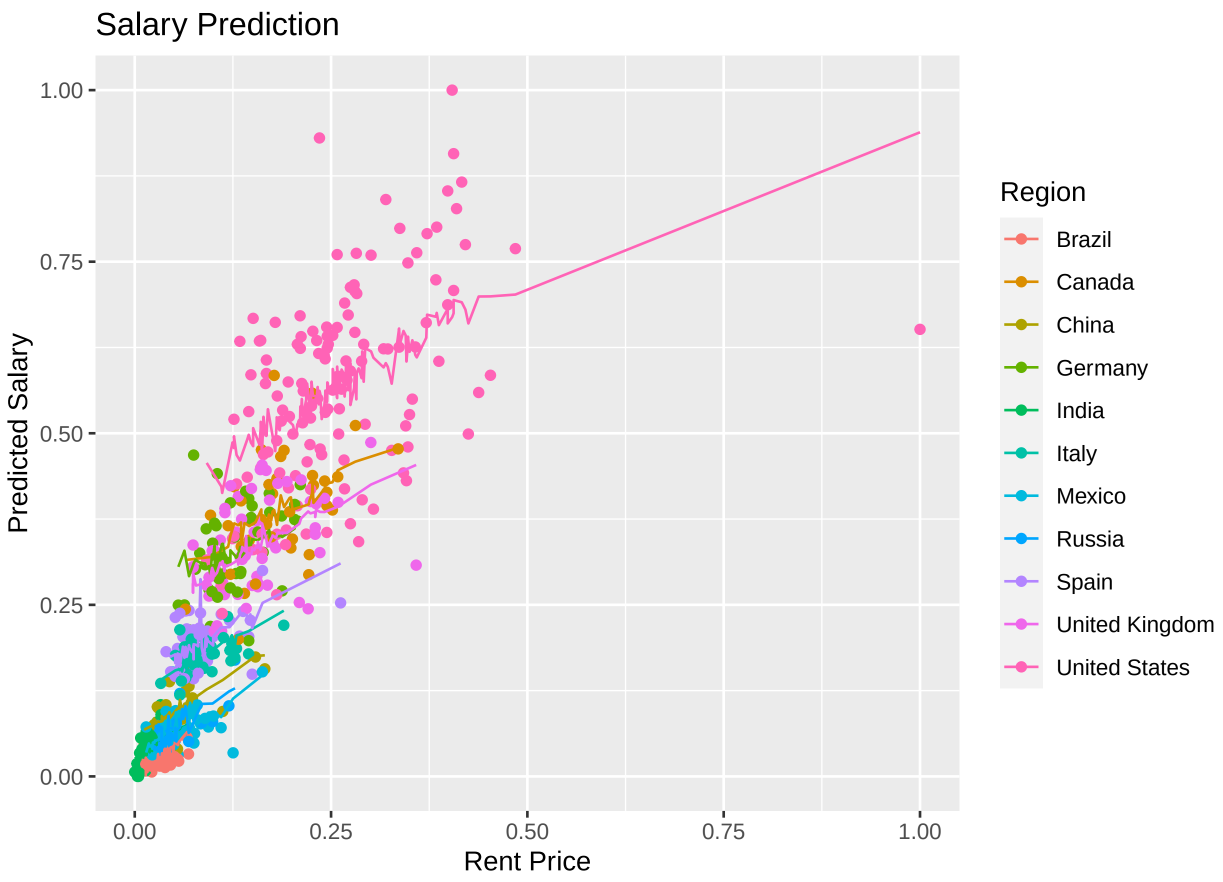

Figure 7: prediction of level 2 model

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```

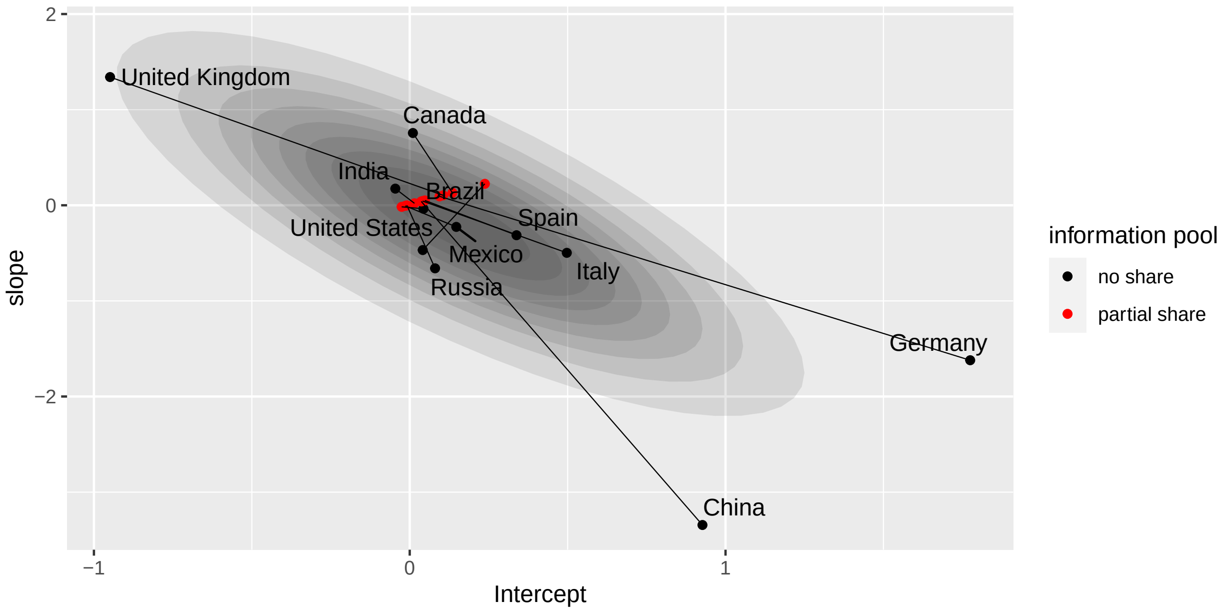

Figure 8: information pool

5. Discussion and Limitation

The first limitation is that data not so large in level 2. Although the data seems large, but it has many groups of level 1 but small data in level 2. Many regions have cities less than 30. In my opinion, it will affect our further analysis. Moreover, the variables with fixed effects of our linear mixed model are not all significant. And this raise me several questions. Is this data not fit the requirements of multilevel model ? Or are we overstate the **region** random effect?

For the future, I would add one more variable related to time in my data, due to it is a quarterly data. Also, I would change a way to analyze the relations between salary and cost of living based on different regions. It is a obvious panel data, so we would have better way that is specific on panel data.

6. Reference

Grimes, D. R., Prime, P. B., & Walker, M. B. (2019). Geographical Variation in Wages of Workers in Low-Wage Service Occupations: A US Metropolitan Area Analysis. Economic Development Quarterly, 33(2), 121-133.

7. Appendix

```
## Warning in par(usr): argument 1 does not name a graphical parameter
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

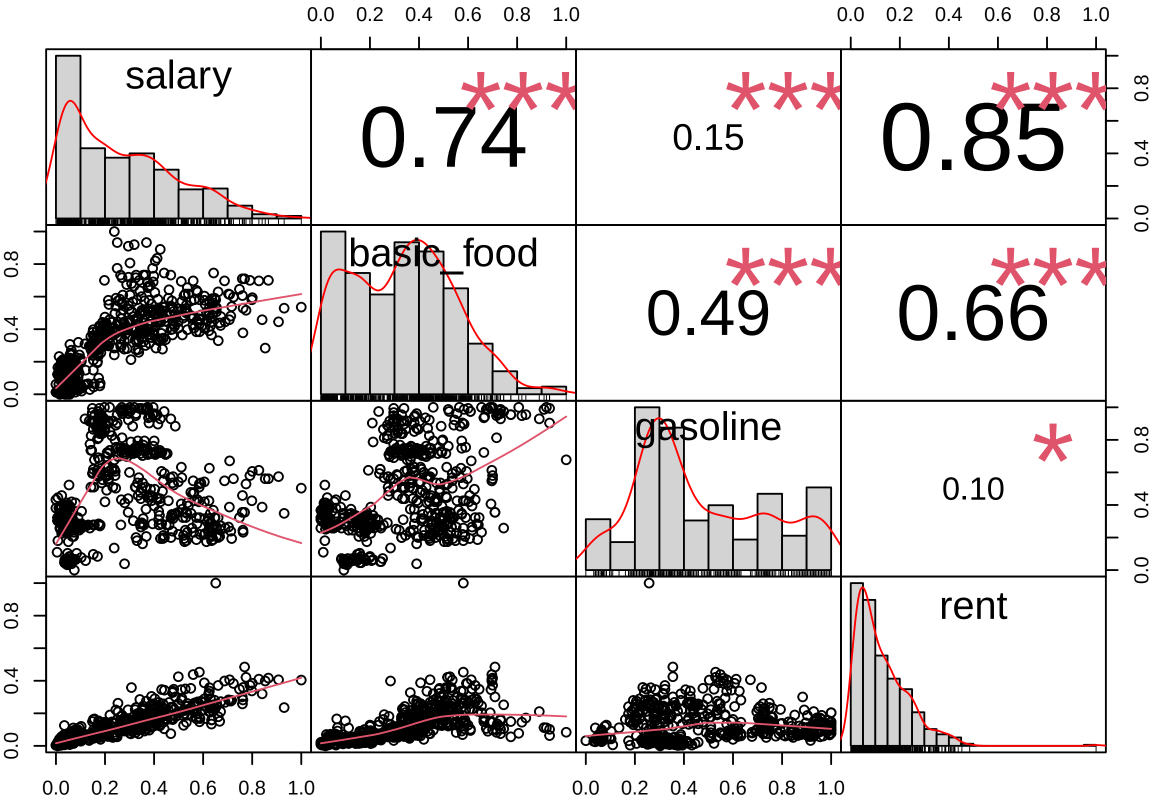

Figure 9: Correlation Matrix

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```

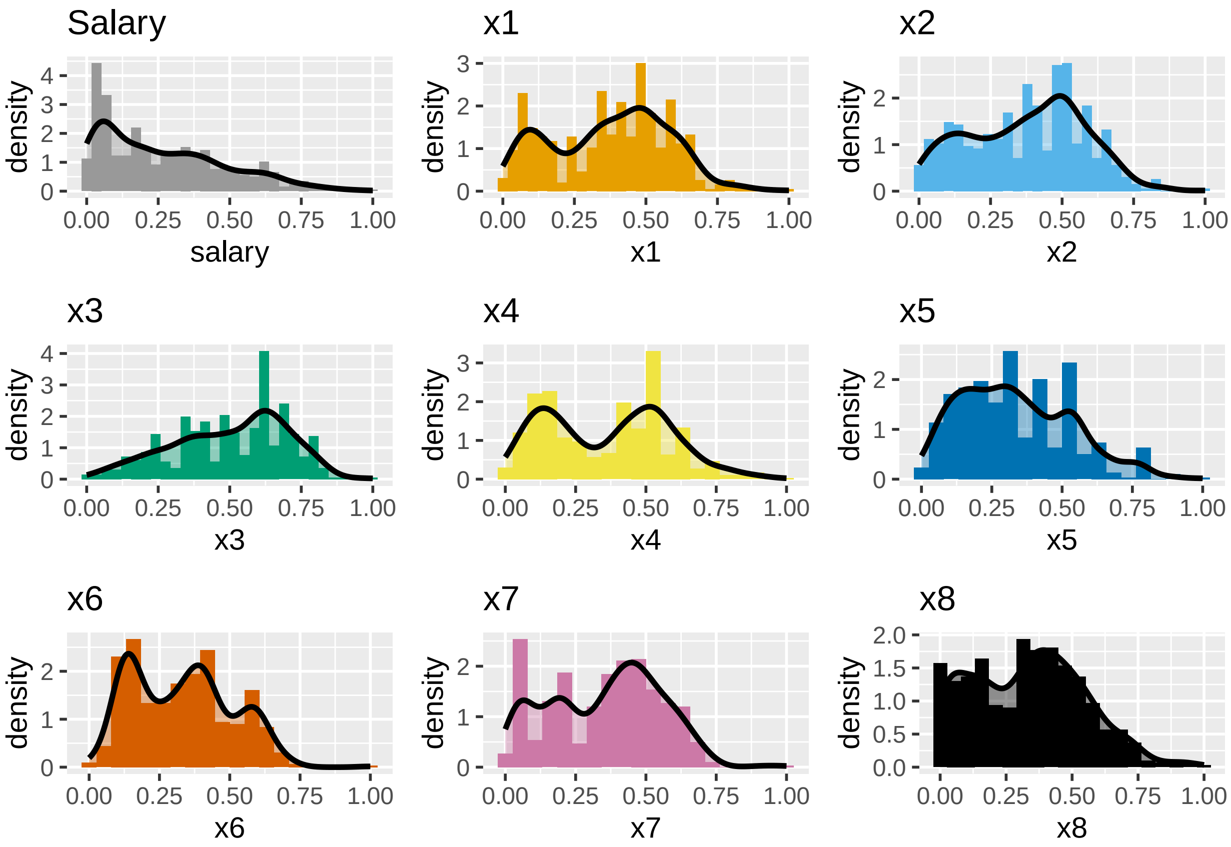

Figure 10: distribution of 25 variables

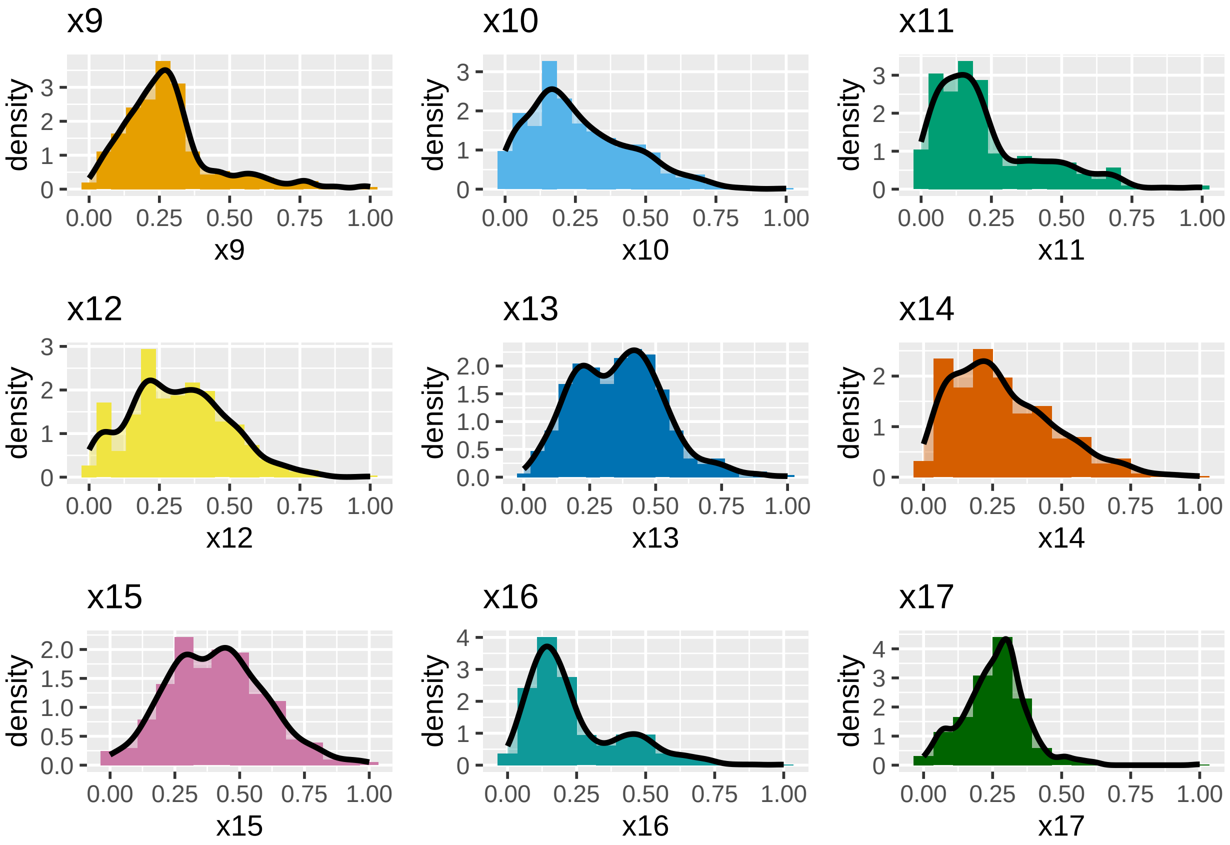

Figure 11: distribution of 25 variables

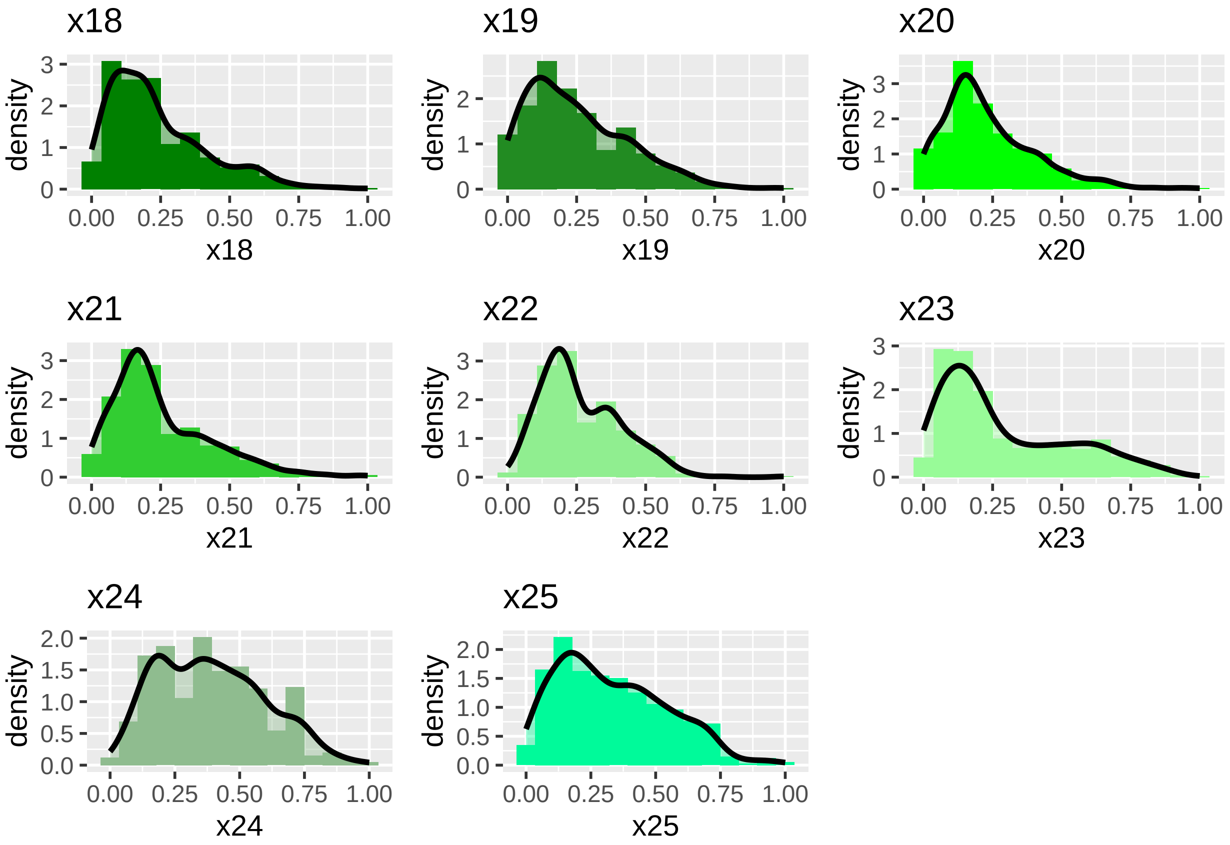

Figure 12: distribution of 25 variables