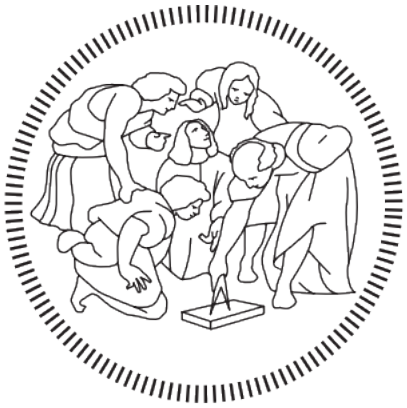

POLITECNICO
MILANO 1863

BAYESIAN PROJECT

AY 2023-2024

Vietri Davide

1 | Presentation of the Dataset

Our dataset is from Kaggle and it presents the Machine Learning Engineer Salary. It could be interesting to study it because probably it will be someone's job in a future. In particular we have a dataset of 16500 observations and the following features:

- **Work year**: It's the year in which the salary data was collected. The considering period goes since 2020 to 2024.
- **Experience level**: We have 4 levels of experience of the employee, i.e. Junior/Entry-Level (EN), Mid-Level (MI), Senior-Level (SE) and Executive-Level (EX).
- **Employment type**: We have 4 types of employment, i.e. Full-Time (FT), Part-Time (PT), Contract (CT) and Freelance (FL).
  Since our dataset is composed of 99 % full-time workers, we will only consider these.
- **Job Title**: We have 155 different Job Titles. The most represented are Data Engineer(21 %) and Data Scientist (20 %).
- **Salary**: It's the salary amount.
- **Salary Currency**: It's the currency in which the salary is denominated. For example GBP for Great Britain Pound. It could be a problem because different currencies entail different values that may be more challenging to compare with each other. Luckily the next features help us to solve this problem.
- **Salary in USD**: The salary amount converted to US Dollars. This allows us to easily compare the values of the salary. We will only consider this variable and disregard the other two.
- **Employee Residence**: It's the country of residence of the employee. We have 88 different countries and the most common are the US (88%).
- **Remote Ratio**: It's the ratio indicating the level of remote work. It goes from 0 (no remote work) to 1.0 (full remote work).
- **Company Location**: It's where the company is headquartered.
- **Company Size**: We have 3 types of size: Small (S), Medium (M) and Large (L). The majority has a medium size (93%).

Our aim is to understand what are the variables that mostly influence our variable target variable, the salary. In order to do this we'll use before a simple Bayesian Linear Regression and later a mixed effects ones.

2 | Exploratory Data Analysis

Firstly we try to clean our dataset. In particular it is important to underline that our dataset doesn't have missed or mismatched values. This is surely an advantage because we can decide by ourselves what remove. As we say before we can take out all the observations which the employment type isn't a Full Time since they are not very representative (they represent less than 80 out of 16494 observations). So we won't have the feature "Employment Type" since we'll consider only one type. As a second thing we can also remove the features "Salary" and "Salary Currency" since we have the variable "Salary in USD", which intrinsically contains the other two. So we went from a dataset of 11 features to one of 8. The observations have decreased slightly.
Before beginning the real exploratory analysis, we have another problem to take care of. Indeed we have 155 different "Job titles" and many are not so representative (like 1 or 2 observations). An idea could be to add a threshold of a minimum of 50-100 observations of a title.
If we put as a ceiling 100 observations we pass from 155 to 18 different "Job titles", which represent the 88,52 % of the total dataset.
As a final preliminary step, we can remove one of the two features, "employee residence" and "company location". In fact, these always coincide, except in very rare cases.
So we have obtained a dataset of 14264 observations and 7 features: work year, experience level, job title, salary in USD, employee residence, remote ratio and company size.
Now we can plot the frequencies of our response variable:

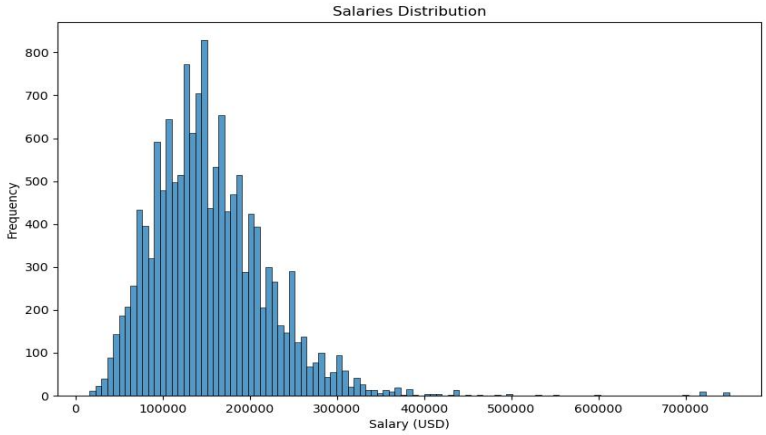

Figure 2.1

The data resemble a Gaussian but we will standardize them, both because this way they can also take on negative values and because many of the variables we are going to consider are dummy variables and so it is certainly better to have similar magnitudes in a regression model.
The second graph represents the average salaries by company over the years:

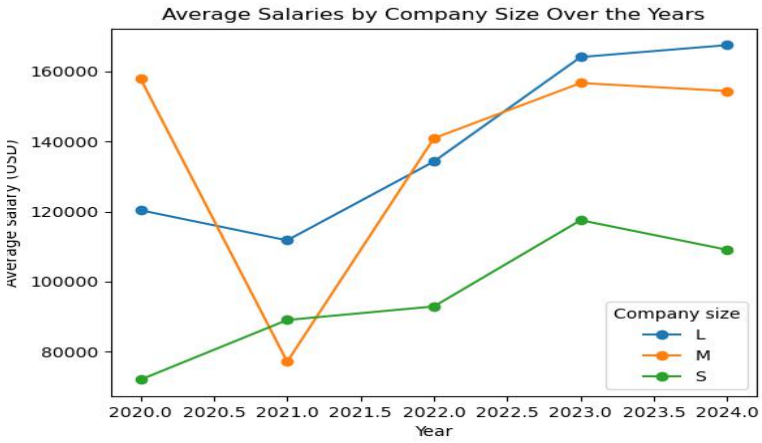

Figure 2.2

From the image, we can observe that during the COVID period (2020-2021), there was a decline in salaries for employees of medium and large companies, while there was an increase for small companies. In the last year, salaries for all types of companies stabilized. Additionally, large companies have consistently paid the highest salaries over the four years considered, while small companies have paid the least.
In particular it could be interesting to understand what is the best type of company for different Experience Levels. A preliminary analysis of this expect can be given by the following graph:

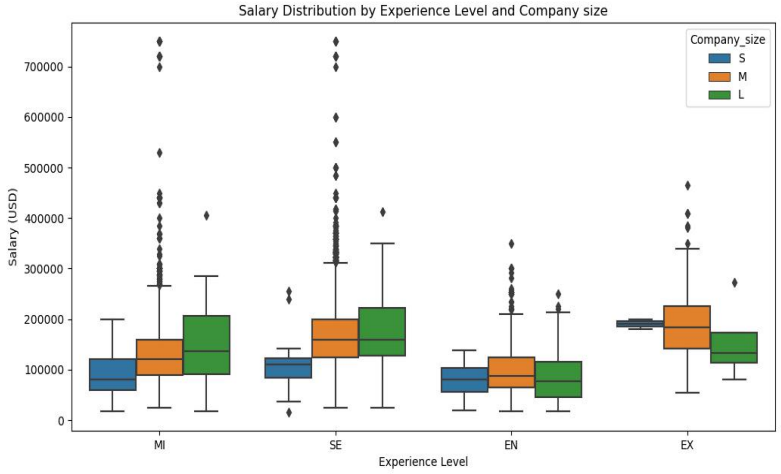

Figure 2.3

We can also compute the median salary that is respectively 86000.0 $ for the Entry-Level (EN), 120000.0 $ for the Mid-Level (MI), 158600.0 $ for the Senior-Level (SE) and 183400.0 $ for the Executive-Level (EX). It is increasing with the prestige of the level and it makes sense.

By the graph we can notice two things: the first one is a strange fact. Indeed we can see that an Executive level generally is paid less than a Senior-level in a large company. The second point is that we don't have so much data for the Small companies (we have a few dozen observations), so we are forced to remove them. Therefore we can consider only Medium companies (12571 observations) and Large ones (714 observations).

The last problem is the amount of outliers we have (as can be seen from Figure 3). One solution may be to remove them by standard deviation thresholding. If we consider a threshold of 1.75 we get the following graph with respect the Company Size (Medium and Large):

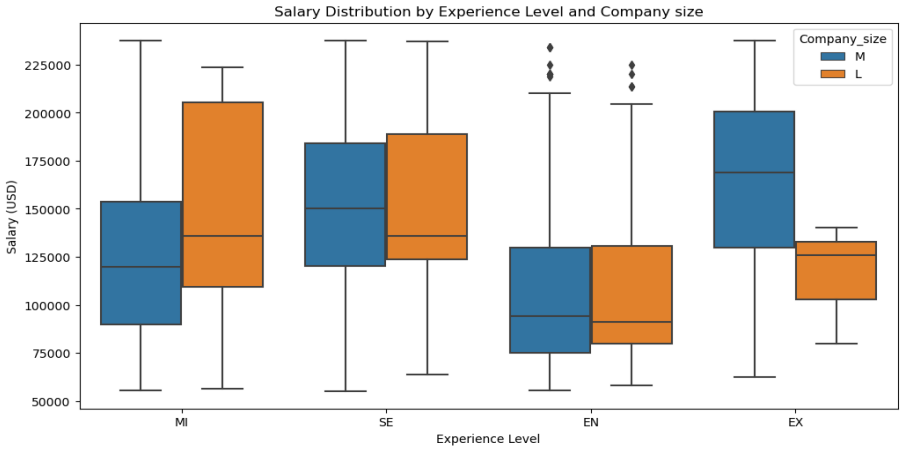

Figure 2.4

In conclusion we have the following dataset:

Work Year	Experience Level	Job Title	Salary in USD	Employee Residence	Remote Ratio	Company Size
2024	MI	Data Science Manager	1.788391	US	0	M
2024	MI	Data Science Manager	0.639839	US	0	M
2024	SE	Business Intelligence Engineer	1.076289	US	0	M
2024	SE	Business Intelligence Engineer	-0.604426	US	0	M
2024	SE	Data Architect	-0.891564	GB	0	M
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021	MI	Data Scientist	-1.408412	US	0	L
2020	SE	Data Science Manager	0.835093	US	1.0	M
2020	MI	Data Scientist	-0.546998	US	1.0	M
2020	MI	Data Scientist	-0.157448	US	1.0	M
2020	MI	Data Engineer	-0.301974	ES	1.0	M

Table 2.1

3 | A simple Bayesian linear model

Firstly, we transform our categorical variables into dummy variables (this transformation will be valid for all the models we will discuss). Omitting for the moment the variables Work Years and Job Title (which we will use later to add mixed effects to our model), thus we obtain the following variables:

- **Continuous Variables**:
  - *Remote Ratio*: It is a number between 0 and 1 and theoretically it could assume every values in this interval on $\mathbb{R}$. In our dataset it assumes only three values 0, 0.5 and 1 (that means respectively all office work, half office work and half home work, and all remote work), so we could consider it as a dummy variable that assumes these three values but after some attempts we can say that it is better to keep it in the form of a continuous variable.
- **Categorical Variables**:
  - *Experience Level*: This categorical variable gives us 4 dummy variables, which corresponds to the 4 groups Entry-Level (EN), Mid-Level (MI), Senior-Level (SE) and Executive-Level (EX).
  - *Employee Residence*: It is equivalent to 6 dummy variables, which are Canada (CA), Germany (DE), Spain (ES), France (FR), Great Britain (GB) and United States (US).
  - *Company Size*: It corresponds to 2 dummy variables, Medium (M) and Large (L).

Now we have to define the design matrix X. We should use all the variables mentioned above, but this could lead to some computational problems since the matrix X is not invertible. In order to avoid this issue we'll consider N-1 dummy variables of each categorical variable. For example in the first group, Experience Level, we'll consider only MI, SE and EX. The impact of the variable EN will be in the intercept of our Bayesian regression model. Moreover we'll add a column of ones in the design matrix to have this intercept. Obviously our response variable is the salary.
Now we can present the first model:

$$
\begin{aligned}
y_i|\beta,\sigma &\overset{\text{iid}}{\sim} \mathcal{N}(x_i^T \cdot \beta, \sigma^2) \\
\beta_j|\sigma_\beta &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta^2) \\
\sigma_\beta &\overset{\text{iid}}{\sim} \text{Inv-Gamma}(1,3) \\
\sigma &\sim \text{Half-Cauchy}(0,1)
\end{aligned}
$$

As we can see, the chains converge:

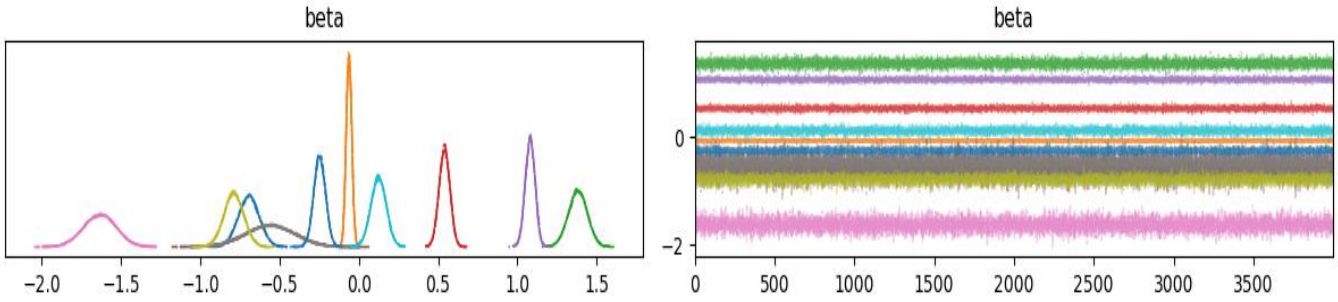

Figure 3.1

In order to visualize better our distribution we can compute the 95% Confidence Region for each Beta:

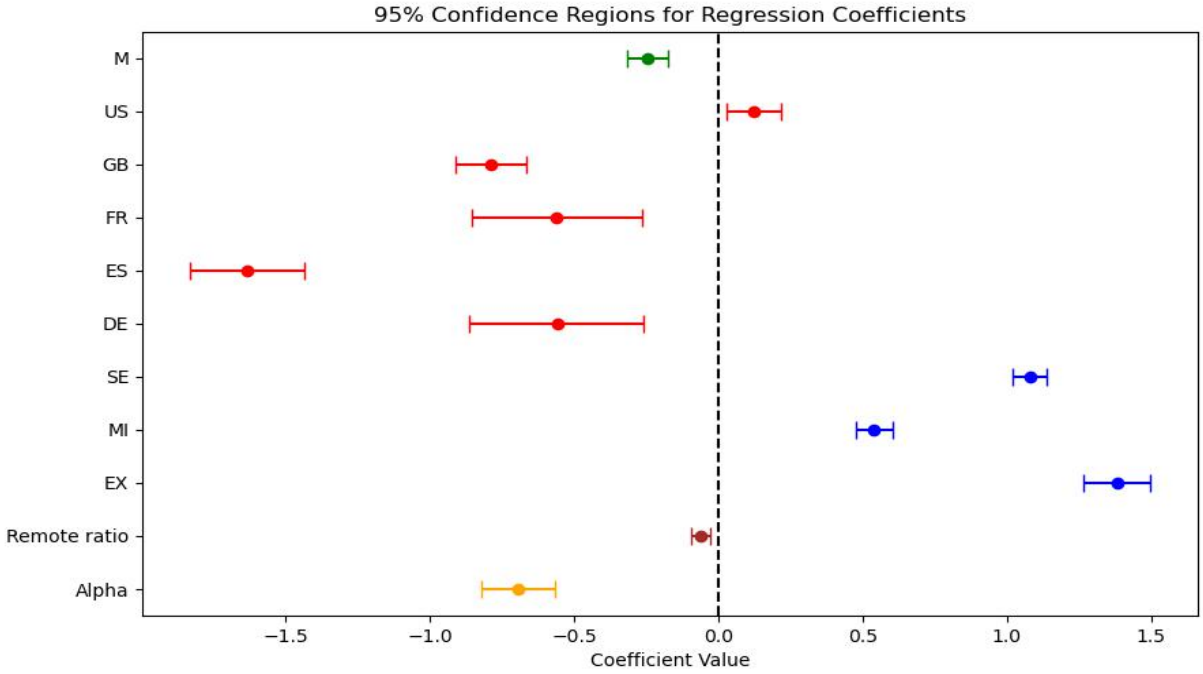

Figure 3.2

Firstly, we observe that all variables are significant at a 95% confidence level, as none of the parameters include zero.

To better understand this graph, we can consider Alpha as a "Base Case," which represents a machine learning worker in a large company (L) who resides in Canada (CA) and holds an entry-level position (EN).
Thus we have that all the other variables represent a difference with respect to this "Base Case". So for example we can say that with a confidence of 95% working in a large company is more financially rewarding compared to working in a medium-sized company.
Regarding the country, we can see that the best location is certainly the United States, while the worst is Spain.
The employment type aligns with logical expectations: the order is Entry-Level, Mid-Level, Senior-Level, and Expert-Level.
Finally, remote work has a slightly negative impact on salary: if one works from home, they will have a slightly lower salary.

4 | A first linear mixed model

We now introduce a slightly more complicated model: we want to consider in our regression also the effect of the year. Indeed, from the Exploratory Data Analysis and, in particular, from the Figure 2.2 we can see as the salary varies over the time. So it could be useful using the work year as a random effect. In order to do this we notice that the first 2 years of the dataset (i.e. 2020 and 2021) has few observations, thus we remove them and we'll consider only 2022,2023 and 2024. Thus, the model is as follows:

$$
\begin{aligned}
&y_{i,t}|\mu_{i,t}, \sigma \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{i,t}, \sigma^2) \\
&\mu_{i,t} = x_i^T \cdot \beta + \lambda_t \\
&\beta_j|\sigma_\beta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta^2) \\
&\sigma_\beta \overset{\text{iid}}{\sim} \text{Inv-Gamma}(1,3) \\
&\sigma \sim \text{Inv-Gamma}(1,3) \\
&\lambda_1, ..., \lambda_3|\lambda_0, s_0 \overset{\text{iid}}{\sim} \mathcal{N}(\lambda_0, s_0) \\
&\lambda_0, s_0 \overset{\text{ind}}{\sim} \mathcal{N}(0,1) \times \text{Inv-Gamma}(1,3)
\end{aligned}
$$

where $y_{i,t}$ makes the dependence of the year explicit (t) and λ_t is a random intercept of the specific of the year. In order to implement it on Python there is a more convenient formulation:

$$
\begin{aligned}
&y_i|\mu_i, \sigma \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2) \\
&\mu_i = x_i^T \cdot \beta + g_i^T \cdot \lambda_t \\
&\beta_j|\sigma_\beta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta^2) \\
&\sigma_\beta \overset{\text{iid}}{\sim} \text{Inv-Gamma}(1,3) \\
&\sigma \sim \text{Inv-Gamma}(1,3) \\
&\lambda_1, ..., \lambda_3|\lambda_0, s_0 \overset{\text{iid}}{\sim} \mathcal{N}(\lambda_0, s_0) \\
&\lambda_0, s_0 \overset{\text{ind}}{\sim} \mathcal{N}(0,1) \times \text{Inv-Gamma}(1,3)
\end{aligned}
$$

where g_i is the indicator vector (i.e. $g_i[s] = 1$ if the observation i belongs to the group s)

Using the same matrix X of the first model (with a few lines down since we removed the lines of 2020 and 2021 as work year), we get the following distribution:

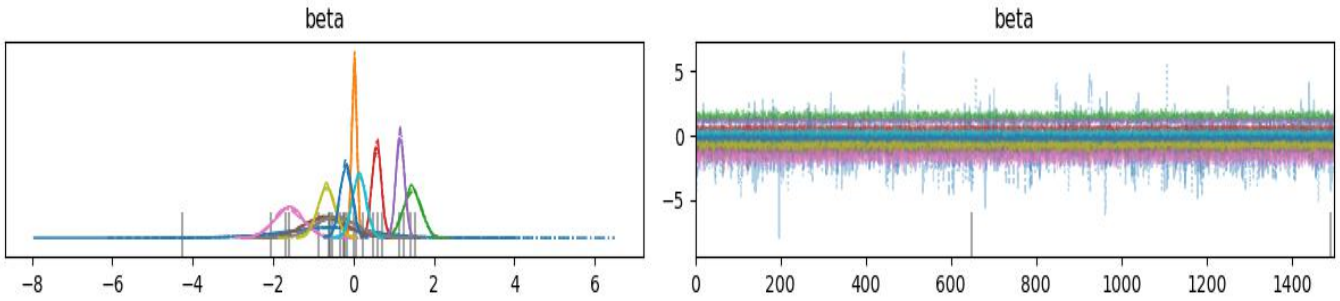

Figure 4.1

As we can see there are two iterations that are divergent, but we can accept this model. The confidence regions for the Beta of this model are the following:

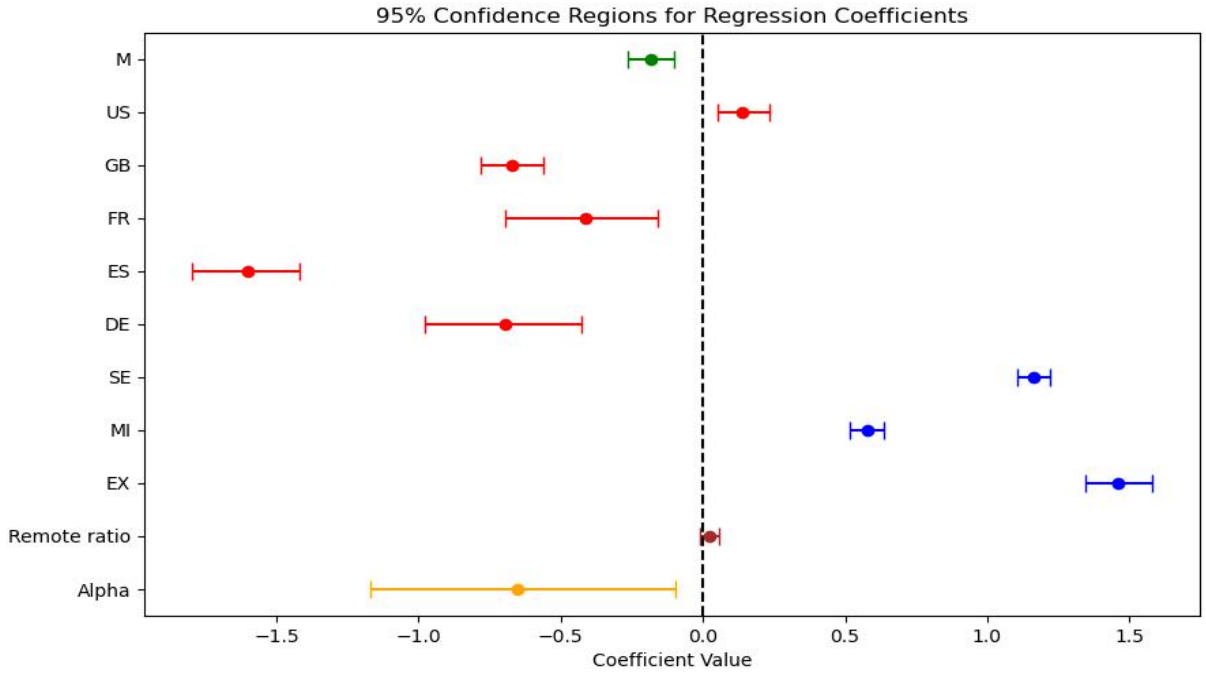

Figure 4.2

As before, we are considering the intercept associated with the baseline case: a Canadian worker, Entry-Level, and employed in a Large company.
What more can we say compared to the simple linear model?
Adding a mixed effect for the years allows us to achieve more or less the same results, except for the remote ratio, where the increase becomes positive. This is certainly due to the impact of Covid, which made remote work increasingly common whereas before it may have been more niche. Indeed, while remote workers were only 18 % in 2022, they

increased to 32% in 2023 and further to 37% in 2024.
For the categorical variables, however, we obtain results very similar to those before.

5 | A second linear mixed model

Now we want to introduce in the model also the "Job Title" as a random effect. We have two possibilities: crossed random effects model and nested random effects model. We choose the first type since there is no real hierarchical relationship between the two random effects.
Before we write the model, we need to address another issue, namely the fact that the categorical variable Job Title contains 18 different types, despite the reduction made in the exploratory analysis. This could lead to computational problems. To resolve this, we consider only three broad groups: Engineers (i.e., those whose Job Title contains the word 'Engineer'), Scientists (i.e., those whose Job Title contains the word 'Scientist'), and Others (i.e., those whose Job Title contains neither word).
Thus the model is the following:

$$
\begin{aligned}
&y_{i,t,s}|\mu_{i,t,s}, \sigma \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{i,t,s}, \sigma^2) \\
&\mu_{i,t,s} = x_i^T \cdot \beta + \lambda_t + \theta_s \\
&\beta_j|\sigma_\beta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta^2) \\
&\sigma_\beta \overset{\text{iid}}{\sim} \text{Inv-Gamma}(1,3) \\
&\sigma \sim \text{Inv-Gamma}(1,3) \\
&\lambda_1, ..., \lambda_3|\lambda_0, s_0 \overset{\text{iid}}{\sim} \mathcal{N}(\lambda_0, s_0) \\
&\theta_1, ..., \theta_3|\lambda_1, s_1 \overset{\text{iid}}{\sim} \mathcal{N}(\lambda_1, s_1) \\
&\lambda_0, s_0 \overset{\text{ind}}{\sim} \mathcal{N}(0,1) \times \text{Inv-Gamma}(1,3) \\
&\lambda_1, s_1 \overset{\text{ind}}{\sim} \mathcal{N}(0,1) \times \text{Inv-Gamma}(1,3)
\end{aligned}
$$

As we did before, we can rewrite the model:

$$
\begin{aligned}
&y_i|\mu_i, \sigma \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2) \\
&\mu_i = x_i^T \cdot \beta + g_i^T \cdot \lambda_t + h_i^T \cdot \theta_s \\
&\beta_j|\sigma_\beta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta^2) \\
&\sigma_\beta \overset{\text{iid}}{\sim} \text{Inv-Gamma}(1, 3) \\
&\sigma \sim \text{Inv-Gamma}(1, 3) \\
&\lambda_1, ..., \lambda_3|\lambda_0, s_0 \overset{\text{iid}}{\sim} \mathcal{N}(\lambda_0, s_0) \\
&\theta_1, ..., \theta_3|\lambda_1, s_1 \overset{\text{iid}}{\sim} \mathcal{N}(\lambda_1, s_1) \\
&\lambda_0, s_0 \overset{\text{ind}}{\sim} \mathcal{N}(0, 1) \times \text{Inv-Gamma}(1, 3) \\
&\lambda_1, s_1 \overset{\text{ind}}{\sim} \mathcal{N}(0, 1) \times \text{Inv-Gamma}(1, 3)
\end{aligned}
$$

The chains are all convergent:

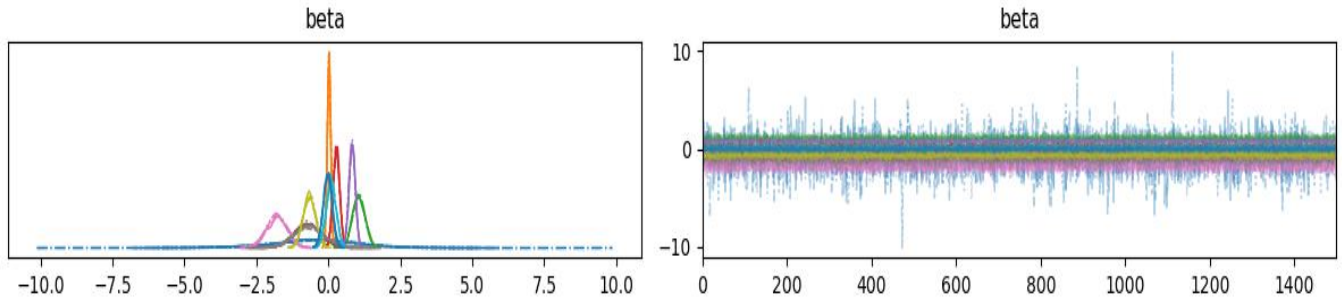

Figure 5.1

Thus the confidence regions for the Beta are the following:

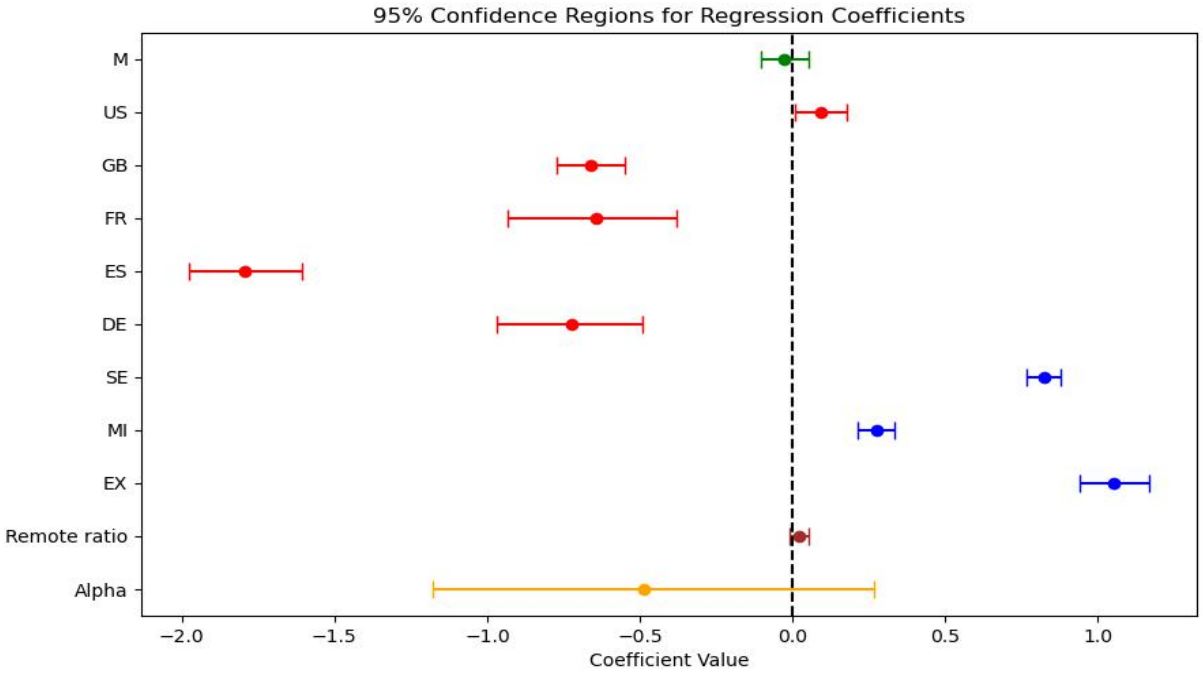

Figure 5.2

This model is more interesting with respect to the previous one, since we have more differences with respect the simple linear model.
In particular, considering our classic "Base Case" (EN, CA and L), we can no longer assert that with a 95% confidence it has a negative impact on the salary. Moreover there is no longer difference between a Medium company and a Large one, since the distribution of the increment of the β_M is centered in 0 and has a small variance.
Furthermore all the increments referred to the Employee residence have decreased or increased.
Finally the increments related to experience level are practically unchanged.

6 | Model Selection

In order to do Model Selection we will use 2 criteria: the WAIC comparison and the LOO comparison.
The WAIC is the "widely applicable information criterion" and is defined as Log-Pseudo Marginal Likelihood (LPML) plus a penalization:

$$WAIC_j = \sum_{j=1}^{N} logm(y_i|\mathbf{y}, M_j) - p_{W_j}$$

where

$$p_{W_j} = \sum_{j=1}^{N} Var_{\theta_j|\mathbf{y}} logf(y_i|\theta_j, M_j)$$

In our case the WAIC comparison is the following:

WAIC Comparison	**Rank**	**elpd_waic**	**p_waic**
Model 1	0	-17500.856617	11.812804
Model 3	1	-28378.590718	6292.072617
Model 2	2	-28558.034727	6181.691967

Table 6.1: WAIC Comparison

It is clear that the model 1 is the best one.
In particular the *elpd_waic* is a measure of expected log pointwise predictive density. Thus the higher it is, the better it predicts the data.
The *p_waic* is a measure of model complexity. Thus the higher it is, the more complex the model is.
So the model 3 has a better predictive measure and is more complex with respect to the model 2.
The LOO is the "Leave-One-Out Cross Validation". In this Validation each datum is removed, the model is recalculated and the prediction for the excluded observation is compared with the observed value.
In our case the LOO comparison is the following:

LOO Comparison	**Rank**	**elpd_loo**	**p_loo**
Model 1	0	-17500.861919	11.818107
Model 3	1	-28948.308330	6861.790229
Model 2	2	-29133.966387	6757.623628

Table 6.2: LOO Comparison

The *elpd_loo* and the *p_waic* have the same meaning as defined in WAIC, but in this case we use the LOO.
We can do the same conclusions as before.

7 | Prediction

In order to do prediction we will use only the first model (the simple linear model) and the third one (the linear mixed effect model with two random effects, the year and the Job Title).
In particular I want to do prediction of my future work: I want to work in a Large Company (since there is greater growth potential), in Spain (since I did the Erasmus here, so I know the language), I'm an Entry Level (since I don't have experience) and I want to do a 50% of Remote Working. Thus, in order to do a prediction with this information we have to consider the following vector:

$$\beta_{pred} = [1, 0.5, 0, 0, 0, 0, 1, 0, 0, 0, 0]$$

where:

- 1 is referred to our Base Case: Entry-Level, Canada and Large company.
- 0.5 is the quantity of remote ratio.
- 1 is the increment to add to get the Spain as Employee Residence.

In particular for the third model we want to do prediction on an existing groups. So for the first group (the work year) I use the 2024, that is the most recent data, while for the second group (the Job Title) I use the "Engineer" group. Thus we get the following predictions:

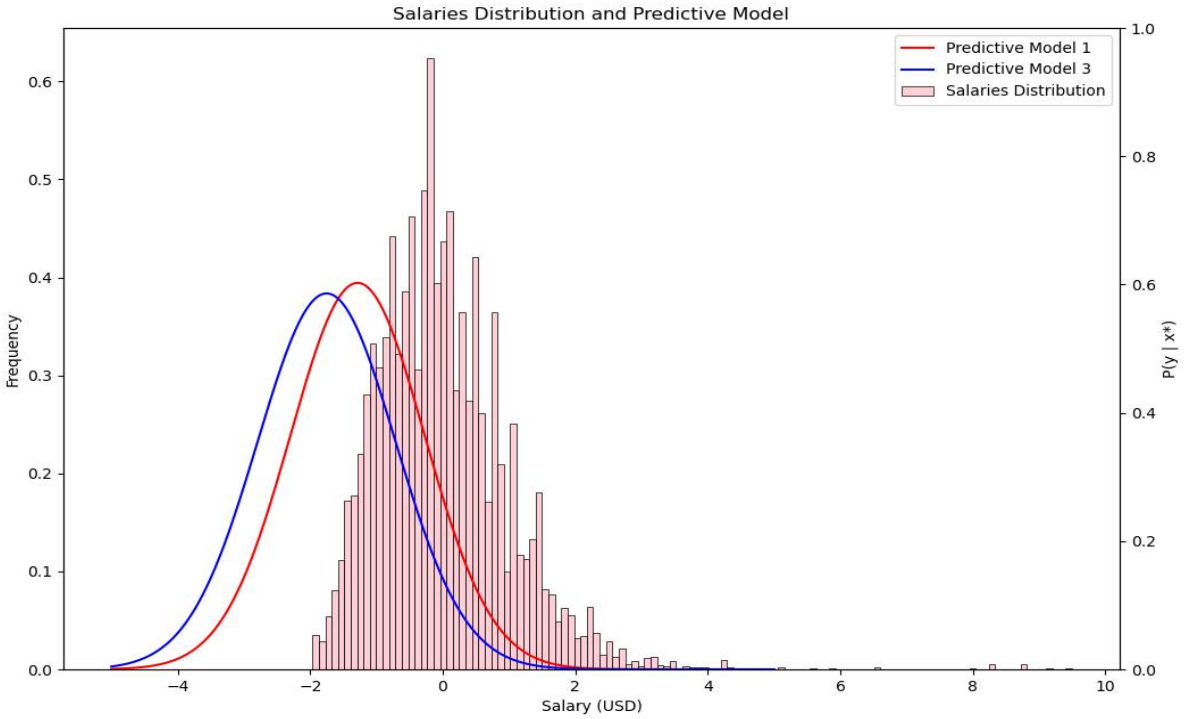

Figure 7.1

As we can see from the plot both results obtained are significantly lower compared to the salary distribution. This is mainly due to the fact that I am at an entry-level position, which is by far the worst among the 4 categories of Experience Level.
Moreover I want to work in Spain that is the worst country among the 6 that we are considering for both models. The slightly difference between the "Predictive Model 1" and the "Predictive Model 3" is given by the fact that the Confidence Interval for the increment of β_{Spain} is slightly worse for the model 3.

8 | Conclusion

It is clear that some results apply to all three models: for example, Spain is always the worst-performing country and the United States is always the best in terms of salary. Furthermore, as logic also suggests, the more experience a person acquires, the higher their earnings, as evidenced by the order in which the confidence intervals of the experience level are arranged.
Considering the two best models instead (the first and the third), there are some things we cannot assert: in particular, we cannot say anything about the impact of the remote ratio. In fact, in the first model, it has a negative impact, meaning the more you work from home, the less you earn. In the third model, however, it has a positive impact, meaning the more you work from home, the more you earn. This is mainly due to the inclusion of a mixed effect related to the year.
Finally, for company size, it is clear that in the first model, it is better to work in a large company than in a medium-sized one, while in the third model, nothing can be asserted since the confidence interval for the increment of β_{Medium} is practically centered at 0.
Regarding predictive ability, the first model is by far the best. The third model, on the other hand, is the most complex.
Finally, the predictions, in accordance with the confidence regions, advise against a person who is entry-level and wants to work in Spain at a medium-sized company (worst case) because their predictive distribution is much lower than the overall distribution. A logical prediction suggests our better case: a person who is expert-level and wants to work in the United States at a large company. In this case, the predictive distribution will be entirely above average.

The code

You can find the code on this link.