UNIVERSITÀ DEGLI STUDI DI PALERMO
SCUOLA POLITECNICA
Corso di Laurea Magistrale – Statistica e Data Science
Dipartimento di Scienze Economiche, Aziendali e Statistiche (dSEAS)

November 2022

Analysis of professionals Big Data field salaries

Marta Bongiorno
Francesco Pio Maria Di Carlo
Veronica Milazzo
Elena Varsalona

Prof. Antonino Abbruzzo

ACADEMIC YEAR 2022 - 2023

Introduction

The world is getting increasingly digital, and this means big data is here to stay. In fact, the importance of big data and data analytics is going to continue growing in the coming years. Choosing a career in the field of Big Data and Analytics will be a good choice for your future.

The dataset used for this analysis concerns the salaries of professionals all over the world in Big Data space and the temporal coverage is between 1st January 2020 and 8th September 2022. The variables of interest for this analysis are nine and were collected from 607 people:

- **Working Year:** The year the salary was paid (2020, 2021, 2022).
- **Experience:** The experience level in the job during the year with the following possible values: **EN (**Entry-level / Junior), **MI (**Mid-level / Intermediate), **SE (**Senior-level / Expert) and **EX (**Executive-level / Director).
- **Employment Status:** The type of employement for the role with the following possible values: **PT (**Part-time), **FT (**Full-time), **CT (**Contract) and **FL (**Freelance).
- **Designation:** The role worked in during the year.
- **Salary in Rupees:** The total gross annual salary amount paid.
- **Employee Location:** Employee's primary country of residence in during the work year as an ISO 3166 country code.
- **Company Location:** The country of the employer's main office or contracting branch as an ISO 3166 country code.
- **Company Size:** The average number of people that worked for the company during the year with the following possible values: **S** less than 50 employees (small), **M** 50 to 250 employees (medium) and **L** more than 250 employees (large).
- **Remote Working Ratio:** The overall amount of work done remotely, possible values are as follows: **0** No remote work (less than 20%), **50** Partially remote and **100** Fully remote (more than 80%).

The goal of this analysis is to understand how the salary of professionals in Big Data space depends on the other variables. So, hiring managers, recruiters or people wanting to make a career switch can make better information decisions. The collected data[1] makes salary information publicy available for anyone to use, share and play around with.

This analisys is divided into 6 part:

1. **Data transformation**: to arrange the dataset for further analisys;
2. **Explorative analysis**: to study the variables distributions an the possible relationships among them;
3. **Models and methodology**: a brief explanation of the methods used;
4. **Bayesian analysis**: application to our data and interpretation of the Bayesian models;
5. **Predictions**: use of predictions as an index of goodness of fit of the model and comparison between models;
6. **Comparing and Conlcusion**

[1] The data was collected from https://salaries.ai-jobs.net/download/

1. **Data transformation**

First of all, we dectected the presence of duplicated observations and we saw that there are 42 duplicates, but we did not remove them because we assumed that more people could have the same contract terms and we do not have other personal information (i.e. gender, age, marital status, etc.). To have a better interpretation of results, we made some transformation. Salary in Rupees variable was changed in Euros and we aggregated the levels of Employee Location, Company Location and Designation.

- For Employee and Company Location we aggregated the levels according to the continent they belong to. The possible new values are: **Europe** (AT, BE, BG, CH, CZ, DE, DK, EE, ES, FR, GB, GR, HR, HU, IE, IT, JE, LU, MD, MT, NL, PL, PT, RO, RS, SI, UA), **Asia** (AE, CN HK, IL, IN, IQ, IR, JP, MY, PH, PK, RU, SG, TR, VN), **Africa** (DZ, KE, NG, TN), **Oceania** (AS, AU, NZ), **North America** (CA, HN, PR, US), **South America** (AR, BO, BR, CL, CO, MX).
- For Designation we aggregated the levels according to the role in Big Data field. The possible new values are: **Data Engineer** (who builds systems that collect, manage and convert raw data into useble information), **Data Analyst** (who analyze and report data to take useful information and solve problems), **Data Scientist** (thos role is similar to data analyst but in addition could make forecasts to support decisions), **Machine Learning** (a specialist who combines statistics and computer science in order to develop algorithms) and **Other** (who do not belongs to any of the previous categories).
- For Employment Status we aggregated the two modes "**CT**" and "**FL**" because they are very few and also they have similar contracts terms. The name of the new variable will be "**SE**" (self-employed).

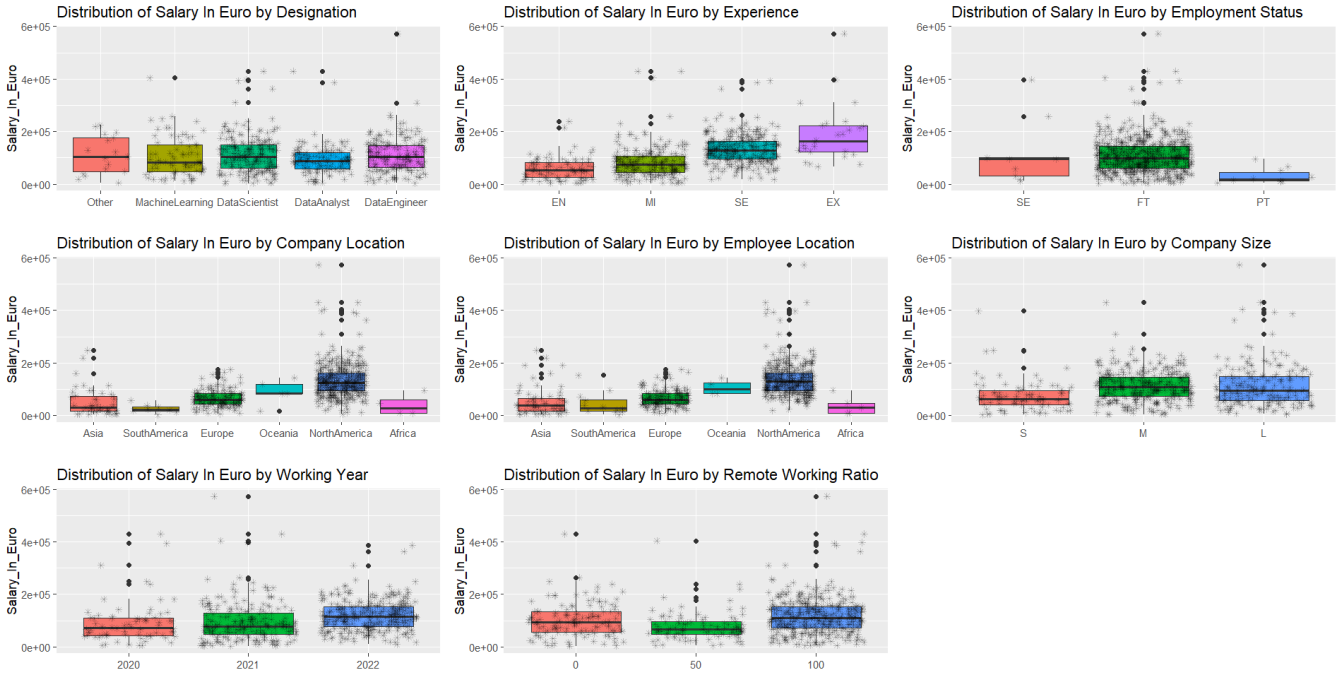
2. Explorative analysis

The first aspect analized is the marginal distribution of each variable (Graph 1).

Graph 1: Marginal distributions of each variable

The Salary distribution is positive asymmetric, so as we know most of the people have a medium salary around €100,000. Only ten people have a very high annual salary greater than €300,000. More than 30% of the people in this dataset is a data scientist follows by data engineer, data analyst and machine learning. Around 45% of professionals have a senior level of experience and only 5% is a director. The large part of employee and companies are located in North America (60%) and 25% are in Europe. Almost all the employees have a full-time job. More than half of companies have a medium size and around 10% have a large size. The greater part of employees work fully remote and the year in which more then half of salaries were paid is 2022.

Then we represented the conditional distribution of salary according to the levels of other variables (Graph 2) using boxplot.

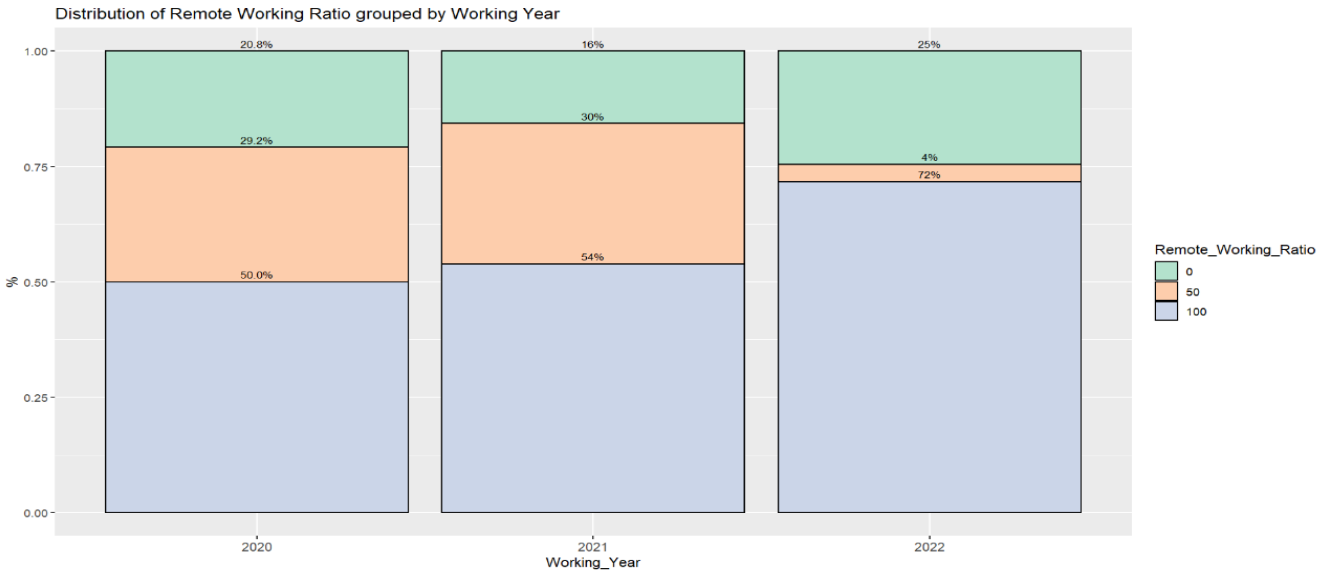


Graph 2: Conditional distributions of Salary in Euros given each other variable

Looking at the distribution of salary by designation we can see that the boxplot are similar among them, only the Data Analyst boxplot has a lower variability. As we expected in according to the experience, the median salary increases as experience increases. In according to the employement status, the employees with SE (self-employed) and FT (full time) have higher salaries, but there are very few employees with SE (self-employed) so the salary for this category is less reliable. The employees in the companies that are located in North America have the highest salaries, follows by the companies located in Oceania and Europe; the same situation occurs for employee location. Probably this is due to higher cost of living in North America rather than the other countries. Eventually the salary is higher for medium and large companies because probably having a higher turnover they are able to guarantee a higher salary. The salary increased from 2020 to 2022 maybe because the Covid-19 pandemic has led to a greater demand for workers in the Big Data field to better respond to the changes that have occurred in recent years and it is lower for those who work partially remote.

Looking at the marginal and conditional distribution of salary we can notice some outliers. We focused on the employee with the highest total gross annual salary amount paid that is equal to € 560,000 to see whether, given his characteristics, the salary received was plausible. He is an american Data Engineer, Executive level/Director, has a full time job for a large company, he works in North America fully remote and the total gross annual salary amount was paid in 2021. So, this employee has features that explain why his gross annual salary is so high.

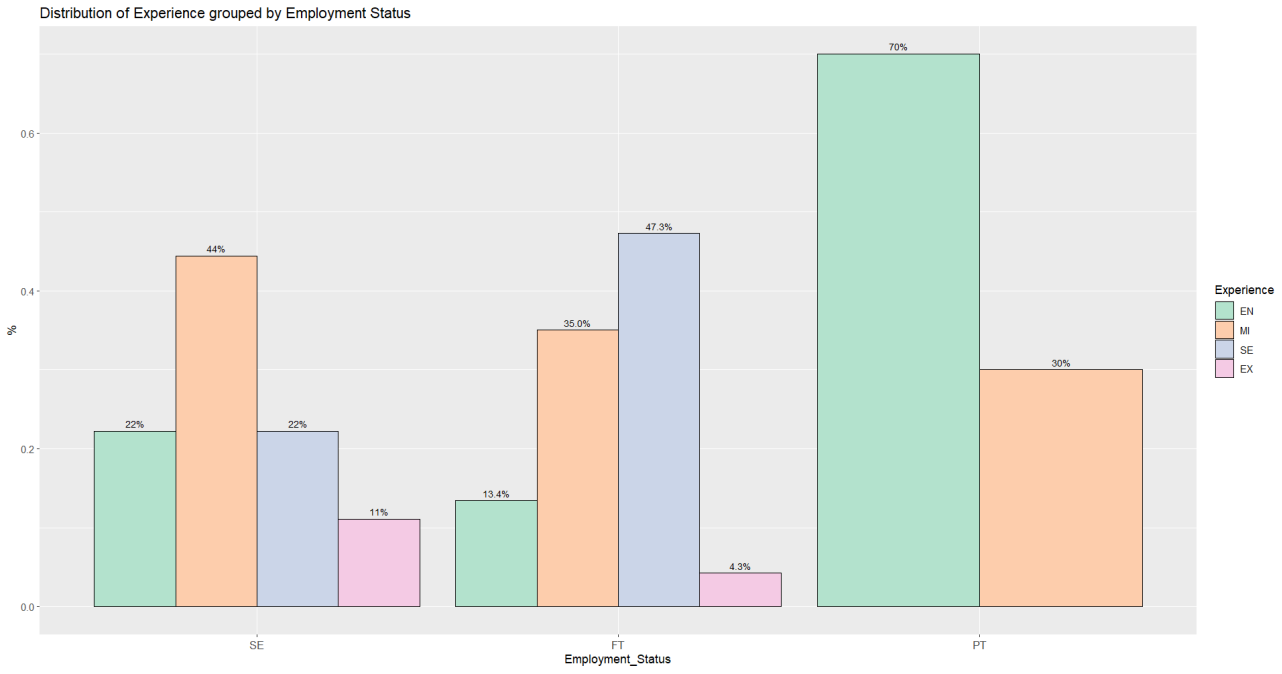
The graph below (Graph 3) shows the distribution of Remote Working Ratio grouped by Working Year. The goal is to understand the employees' working habits during the pandemic years.



Graph 3: Distribution of Remote Working Ratio grouped by Working Year

For all three years, most of the employees works more than 80% remotely, in 2022 we would have expected a greater increase of no remote workers, but probably both employees and companies preferred to work mostly on remote.

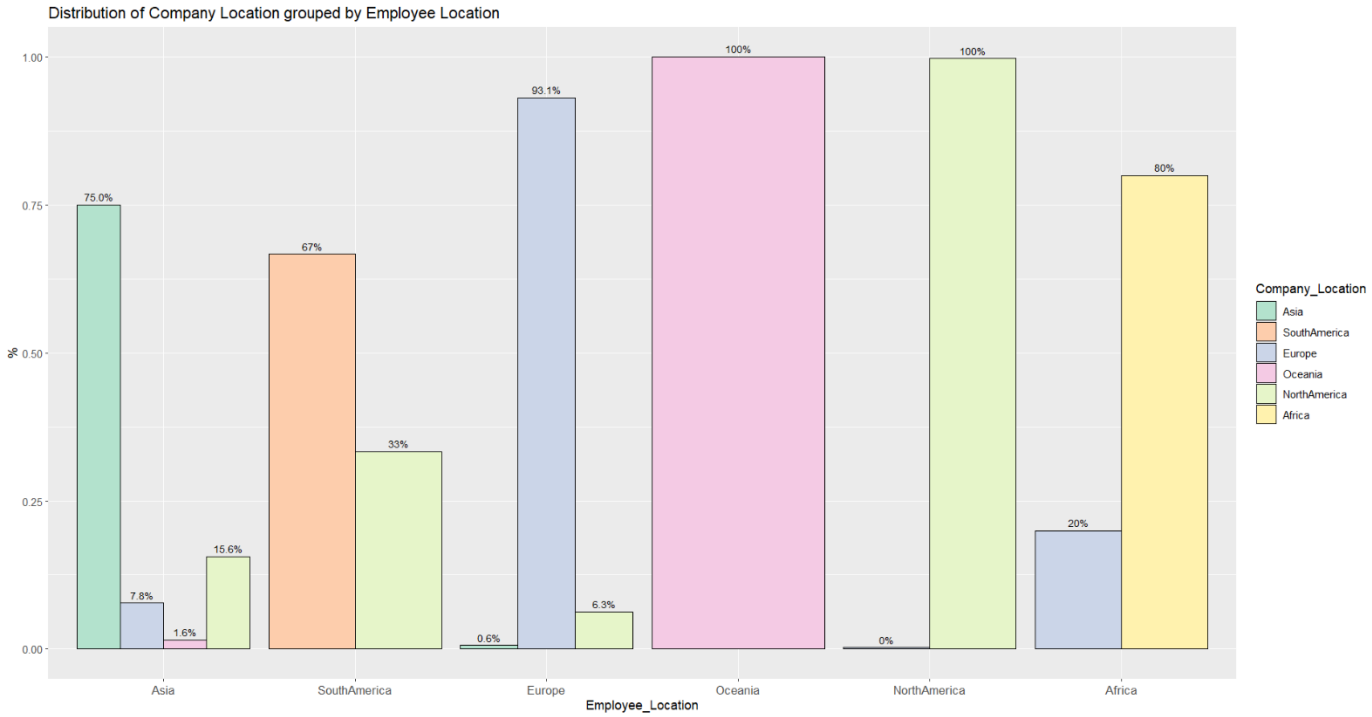
The graph below (Graph 4) shows the distribution of experience grouped by Employment Status.



Graph 4: Distribution of Experience grouped by Employment Status

It is evident that only Entry and Middle experience level professionals have a part-time contract, 70% and 30% respectively. This may be due to the fact that the companies want to test the skills of their employees before hiring them full-time. Among those with full-time, 47.3% have a senior experience level, 35% have a middle experience level and only 4.3% have an executive experience level.

From the conditional distributions of salary by employee location and company location bring us to the same point, so we did a bar plot of company location grouped by employee location (Graph 5) to see the association between this two variables.



Graph 5: Distribution of Company Location grouped by Employee Location

We can notice that those live on one continent work on the same, for this reason we decide to remove Employee Location from the model.

3. Models and methodology

The methodology used is based on the Bayesian regression model. Typically a regression model studies how the response variable depends on the covariates, in the Bayesian case the parameters of interest are not fixed but they are random variables, so we must define prior distributions on these parameters.
Our variable of interest is “Salary in Euro” and takes positive countinuous values, so the distribution chosen could be a gamma or a lognormal. The gamma model is given by:

$$Y_i \sim \text{Gamma}\,(\mu^2\tau, \mu\tau) \quad \text{where } E(Y_i) = \mu_i,\ \text{Var}\,(Y_i) = \tau^{-1}$$
$$\log(\mu_i) = \beta_0 + \beta_1 X_{il} + \ldots + \beta_p X_{ip}$$

The canonical link is $\mu_i^{-1} = \eta_i$, but problems may appear in the MCMC algorithms because μ_i must be positive, moreover, interpretation is not straightforward, therefore we chose the log-link[2].

The lognormal model is given by:

$$Y_i \sim N(\mu, \tau) \quad \text{where } E(Y_i) = \mu_i, \text{ Var } (Y_i) = \tau^{-1}$$
$$\log(\mu_i) = \beta_0 + \beta_1 X_{il} + ... + \beta_p X_{ip}$$

Usually to define the prior distribution's parameter previous data are used, but in this case pre-2020 data are not available, so we decided to use a flat prior for β and the Unit Information Prior for τ. For the flat prior distribution of $\beta \sim N(\beta_0, \Sigma_0)$, we can select:

- β_0 by setting the paramater that refers to the intercept equal to the logarithm of the mean salary and the others equal to zero;
- the variance-covariance matrix Σ_0 as a diagonal matrix with high variances.

A Unit Information Prior is one that contains the same amount of information as that would be contained in only a single observation. This distribution cannot be strictly considered a "real" prior distribution, as it requires knowledge of y to be constructed.
For the prior distribution of $\tau \sim \text{Gamma}(v_0/2, v_0\sigma^2_0/2)$ we can weakly center around $\hat{\sigma}^2_{ols}$ by taking:

$$v_0 = 1 \qquad \sigma^2_0 = \hat{\sigma}^2_{ols}$$

To understand which covariates are important to describe Y we use the Bayesian model selection. This method assigns a probability to each regression model in fact we cannot proceed with classical model selection in which we estimate the significance of each parameter. To interpret the model parameters, the goal is to find the a posterior distribution and summarize it finding a point or interval estimate. For this reason it is often desirable to identify regions of the parameter space that are likely to contain the true value of the parameter. An interval [l(y), u(y)], based on the observed data Y = y, has 95% Bayesian coverage for θ if

$$\Pr(l(y) < \theta < u(y) | Y = y) = .95$$

All points in an HDI (Highest Density Interval) have a higher posterior density than points outside the interval[1].

4. Bayesian analysis

Before starting the Bayesian analysis, outliers were removed from the model to have better predictions and the initial dataset was divided into two parts: training set and test set, containing 447 (75%) and 150 (25%) of observations respectively. The training set was used to define the unit information prior parameters of τ, while the test set was used to make predictions and to compare them with the real values.
The reference profile we chose to interpret the parameters has the following features:

- Working Year = 2020
- Designation = Machine Learning
- Experience = EN (Entry level)
- Employment Status = FT (Full time)

- Company Location = North America
- Company Size = S (Small)
- Remote Working Ratio = 0 (No remote work)

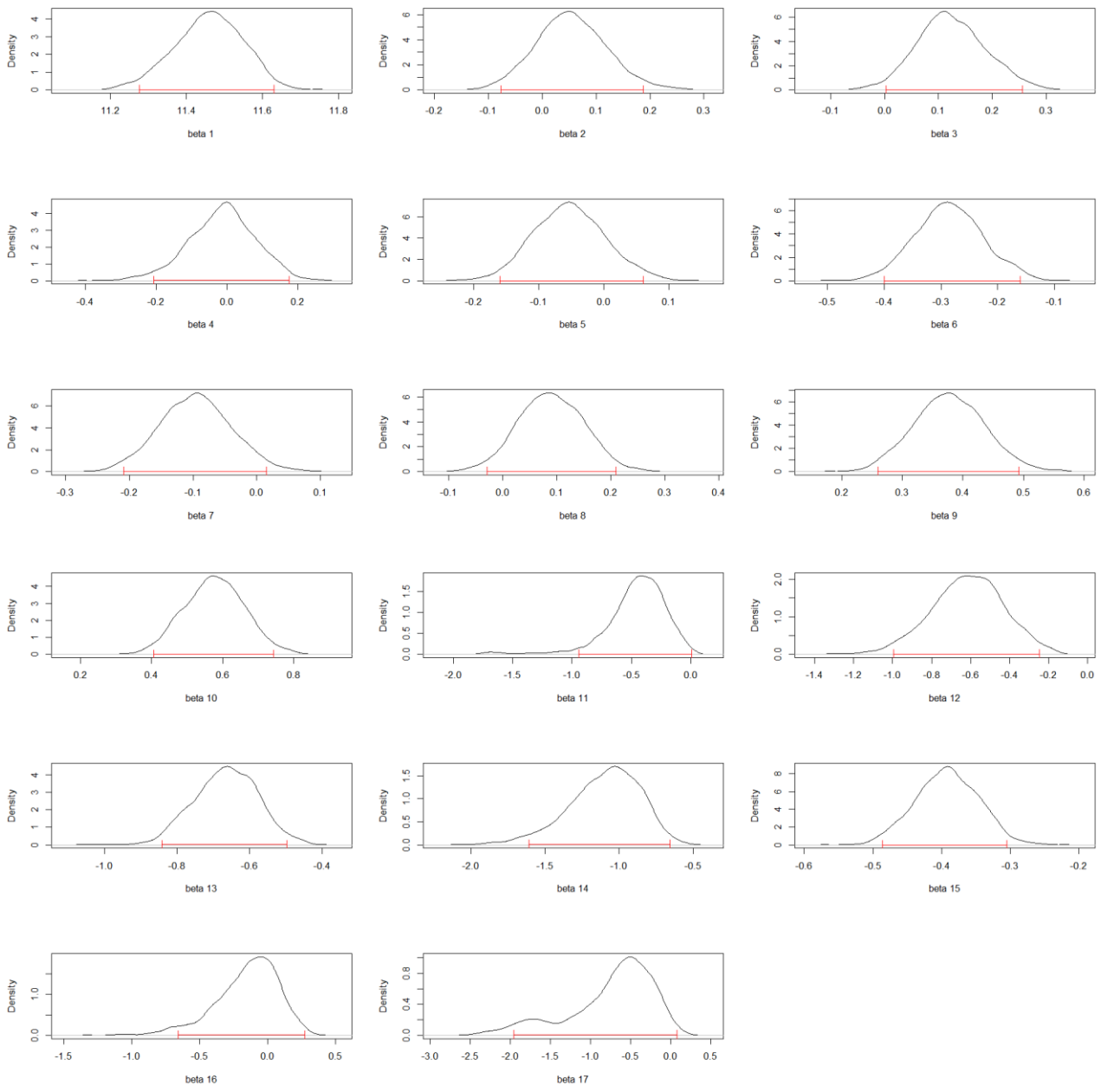
4.1 Gamma model

As we said in the paragraph 3, to each model is assigned a probability and the model with the highest probability equal to 25,85% has the following variables: Working year, Designation, Experience, Employment Status and Company Location.
After applying the algorithm to the best model and estimating the chain we need to check that the parameters are stationary and not autocorrelated. Good practices that allow to reduce the indipendence are:

- Burn-in which consist to discard the first k produced values, where k depends on the speed of convergence of the chain;
- Thinning which consist to keep one draw every d, where d is a fixed number choosen heuristically.
- Multiple chains which consist to find different chains starting from different initial values to avoid sensitivity to the starting point[2].

After making some attempts we did the PACF to find a better value for the thinning and eventually we chose k = 10000, d = 500 and two chains.
In general all parameters are stationary and almost all parameters don't show autocorrelation except for β_{16} and β_{17}. We represented the posterior distributions and the HDI of parameters (Graph 6).
It can be seen that:

- Only one parameter HDI (β_2) related to the variable **Working Year** (β_2 and β_3) contain zero, hence there is no difference in total gross annual salary between 2020 and 2021, while who works during 2022 earns more than who worked during 2020.
- only one parameter HDI (β_6) related to the variable **Designation** (β_4, β_5, β_6, β_7) does not contain zero, hence Data Analyst earns less that Machine Learning.
- two parameters HDI (β_9, β_{10}) related to the variable **Experience** (β_8, β_9, β_{10}) do not contain zero, so as experience increases the salary increases but between middle level and entry level there is any difference in salary.
- parameters HDI related to the variable **Employment Status** (β_{11}, β_{12}) does not contain zero, hence a part time and self employed worker earns less than a full time worker.
- only one parameter HDI (β_{16}) related to the variable **Company Location** (β_{13}, β_{14}, β_{15}, β_{16}, β_{17}) contains zero, so there is no difference in salary between those who work in the companies located in Oceania and North America. In contrast, who work in the companies located on all other continent earn lower. However, the HDI of β_{17} is in a borderline situation that is contrary to the explorative analysis, therefore, regarding the level "Africa" we cannot give a valid interpretation.



Graph 6: Posterior distributions of the gamma model parameters

4.2 Lognormal model

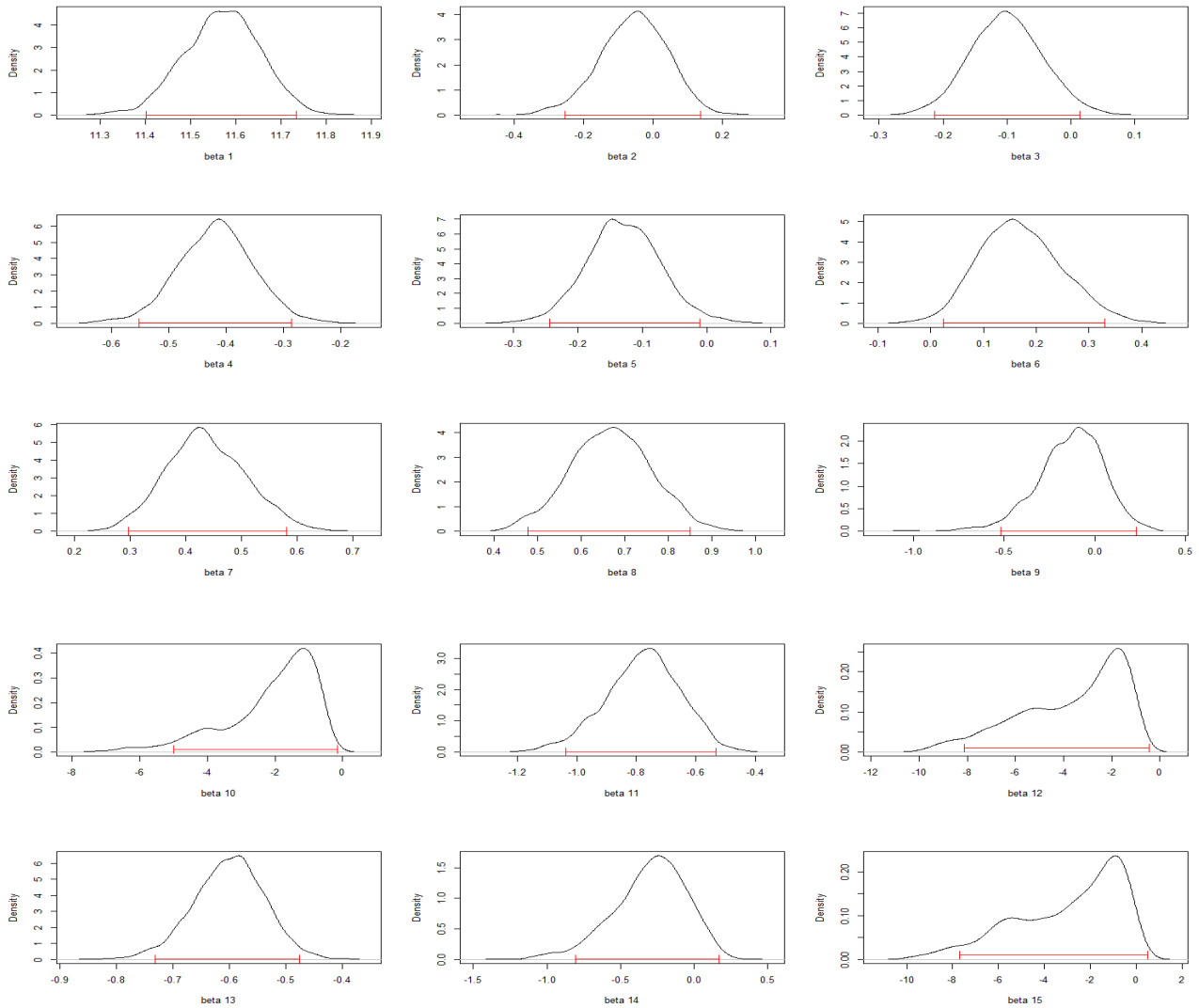
Also in this case we used the same procedure that we followed for the gamma model. The model with the highest probability equal to 24,85% has the following variables: Designation, Experience, Employment Status and Company Location.

After making some attempts we did the PACF to find a better value for the thinning and eventually we chose k = 10000, d = 300 and four chains.

In general all parameters don't show autocorrelation and not stationarity except for β_{10}, β_{12} and β_{15} and only β_9 presents autocorrelation problems. We represented the posterior distributions and the HDI of parameters (Graph 7).

It can be seen that:

- two parameters HDI (β_4, β_5) related to the variable **Designation** (β_2, β_3, β_4, β_5) does not contain zero, hence Data Analyst and Data Engineer earn less that Machine Learning.
- all parameters HDI related to the variable **Experience** (β_6, β_7, β_8) do not contain zero, so as experience increases the salary increases.
- only one parameter HDI (β_{10}) related to the variable **Employment Status** (β_9, β_{10}) does not contain zero, hence a part time worker earns less than a full time worker.
- only one parameter HDI (β_{12}) related to the variable **Company Location** (β_{11}, β_{12}, β_{13}, β_{14}, β_{15}) contains zero, so there is no difference in salary between those who work in the companies located in Oceania and North America. In contrast, who work in the companies located on all other continent earn lower. However, the HDI of β_{15} is in a borderline situation that is contrary to the explorative analysis, therefore, regarding the level "Africa" we cannot give a valid interpretation.

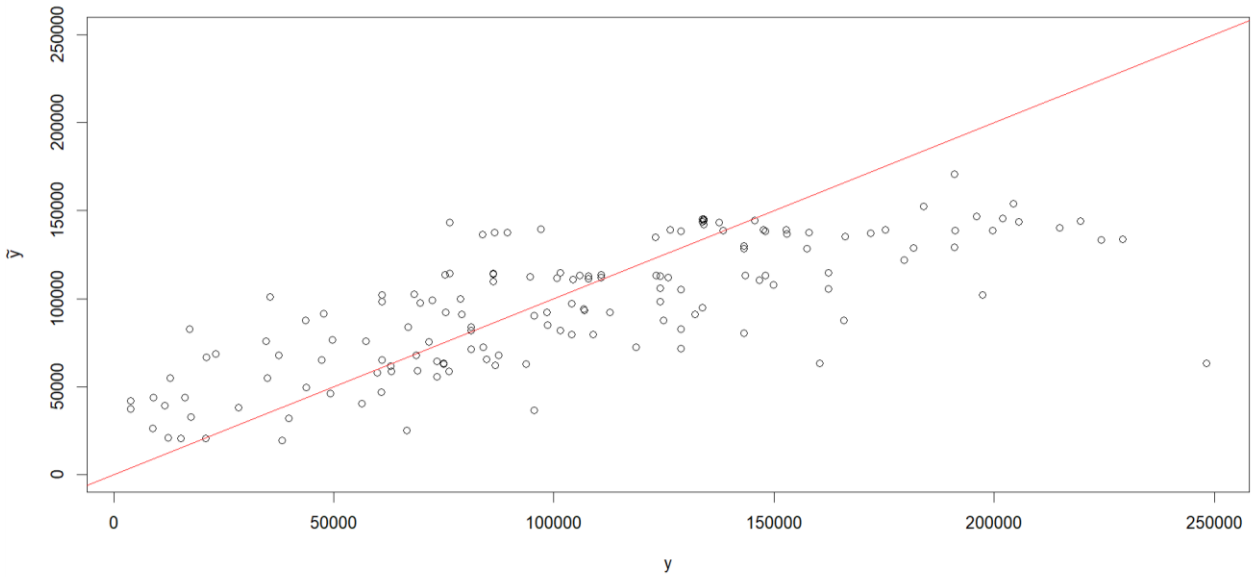


Graph 7: Posterior distributions of the lognormal model parameters

5. Predictions

To test the predictive ability of the model, we used the test set to estimate the new values and compare them to the true values.

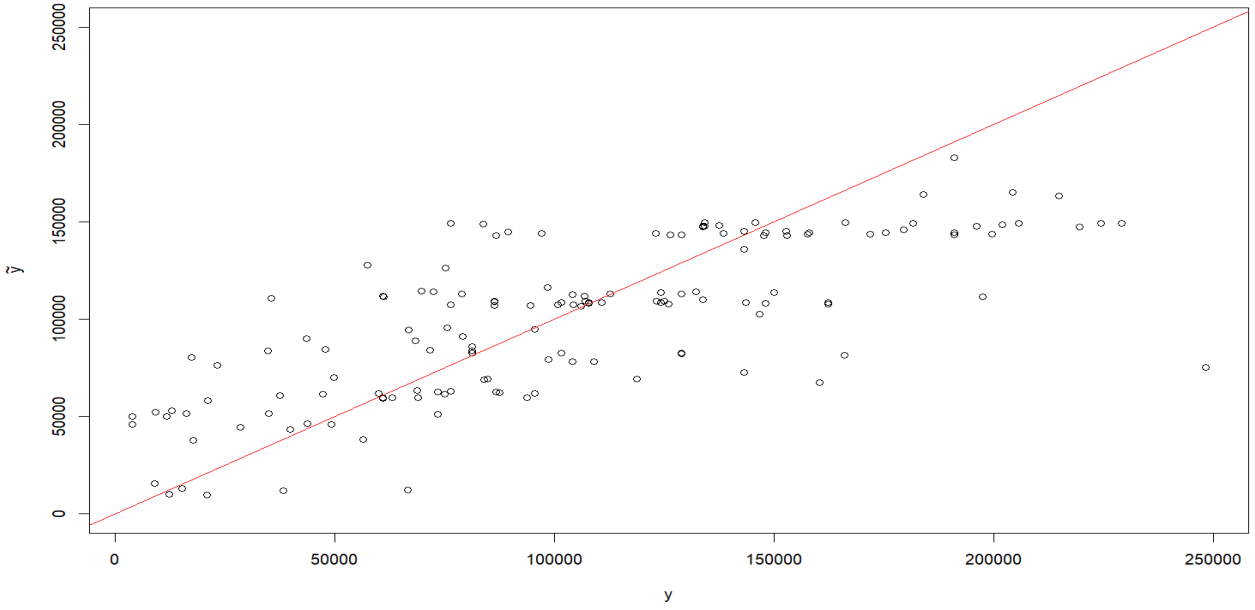
The following graph shows predictions and true values of the gamma model (Graph 8).



Graph 8: Comparison between predicted and observed values in the gamma model

It can be seen that there is a good closeness to the bisector, sign that the model has good predictive ability, although it seems to underestimate high values. In addition, the linear correlation coefficient between predictions and observed values is reasonably high (0.734).

The following graph shows predictions and true values of the lognormal model (Graph 9).



Graph 9: Comparison between predicted and observed values in the lognormal model

Also in this case the model has good predictive ability, and the linear correlation coefficient between predictions and observed values is 0.738.

6. Comparing and Conclusions

We want to conclude the analysis by making some comparisons.

6.1 Gamma model VS Lognormal model

We compared the gamma and lognormal models in terms of significant variables and predictions. We note that the common variables are Designation, Experience, Employment Status and Company Location, while in the gamma model there is also Working Year variable. Looking to the common variables we can see that the parameters of Company Location bring us to the same interpretation instead the parameters of Designation, Experience and Employment Status bring us to slightly different interpretation, respectively:

- for gamma model only the Data Analyst earns less than Machine Learning while for lognormal model also the Data Engineer earns less than Machine Learning.
- for gamma model, between middle level and entry level there is any difference in salary while for lognormal model this difference is there.
- for gamma model both self employed workers and part time workers earn less than full time workers, while for lognormal model only part time workers earn less than full time workers.

As far as the previsions is concerned, we note that the lognormal model leads to slightly better previsions and we also prefer it in terms of parsimony.

6.2 Bayesian models VS Fisherian models

Using the stepwise selection, the best fisherian gamma model contains the following variables: Designation, Experience, Employment Status, Company Location and Company Size, so we note that respect to the bayesian gamma model there isn't Working Year variable but there is Company Size variable.
The linear correlation coefficient between predictions and observed values of fisherian gamma model is 0.725 that is lower than the bayesian one (0.734).

Using the stepwise selection, the best fisherian lognormal model contains all variables except Working Year, so we note that respect to the bayesian lognormal model there are two more variables (i.e Company Size and Remote Working Ratio).
The linear correlation coefficient between predictions and observed values of fisherian lognormal model is 0.736 that is lower than the bayesian one (0.738).

6.3 Conclusions

Of all the estimated models, the one that is slightly better in terms of prediction is the lognormal Bayesian model.

Model	Linear correlation coefficient
Bayesian gamma	0.734
Bayesian lognormal	0.738
Fisher gamma	0.725
Fisher lognormal	0.736

Table 1: Linear correlation coefficient of models

In the end, to answer the objective of the analysis, we can say that those who want to pursue a career in Big Data field can base their decisions mainly on the work experience they have gained over the years, their role in the company, where the company is located and the contract terms. Therefore as these characteristics change, the total gross annual salary received will change.
In addition, hiring managers and recruiters can consider these aspects to make more conscious decisions.

Bibliography

[1] P.D. Hoff, *A first course in Bayesian Statistical Methods*, Springer Science & Business Media, 1st ed. 2009 edition.

[2] Ntzoufras Ioannis, *Bayesian Modelling using WinBUGS*, Wiley, 2009 edition.