Final Project

Zhi Tu

2022-11-18

Read data

Download the data from Kaggle: https://www.kaggle.com/datasets/josephgutstadt/data-jobs.

I got this data from the Kaggle site that contains data science related jobs posting on the Glassdoor website. It gives job descriptions, companies information and the estimated salary range. I would like to find out the reason behind different salaries with different companies. That is to say, I wonder how the location, the company background and the requirement for job applicant affect the predictive salary in data science related jobs. If I can build a model based on the information in this datasets. Individuals then are able to get a sense of the estimated salary paying for their skills and preferred working locations, etc.

Most information is contained in the job descriptions. So I would like to get key words from the job descriptions posted by companies. The first thing I did is to search for required or recommended skills for the job applicants. For data science related works, it is reasonable to assume some kind of programming languages, such as Python, R, and SQL. At the same time, some popular tools such as spark, excel and AWS. These keywords are valued as 1 if it's contained in the job description; otherwise, it's 0. As a result, we got columns consist of 0s and 1s for each keywords, which then can be used as predictors in the regression model.

Data cleaning

We need to clean the Job Estimated Salary column. The original column is a string that contain the range of predicted salary. So we first extract the two ends of the salary range. Then, I planned to take the average on the estimated salary and use it as the outcome of the regression model; and set the other columns as the predictors. Observed that there are -1 values in these columns, I mutate these values to Unknown or Not Applicable by examining the classifications.

EDA

Since our goal is to see if the predictors that have an effect on the median outcome of the data related job post, I would like to get a glimpse at the distribution of the median income. And on the graph below, we can see most jobs are offering around 70k on the Glassdoor. I'd like to find out what predictors influence the income the most.

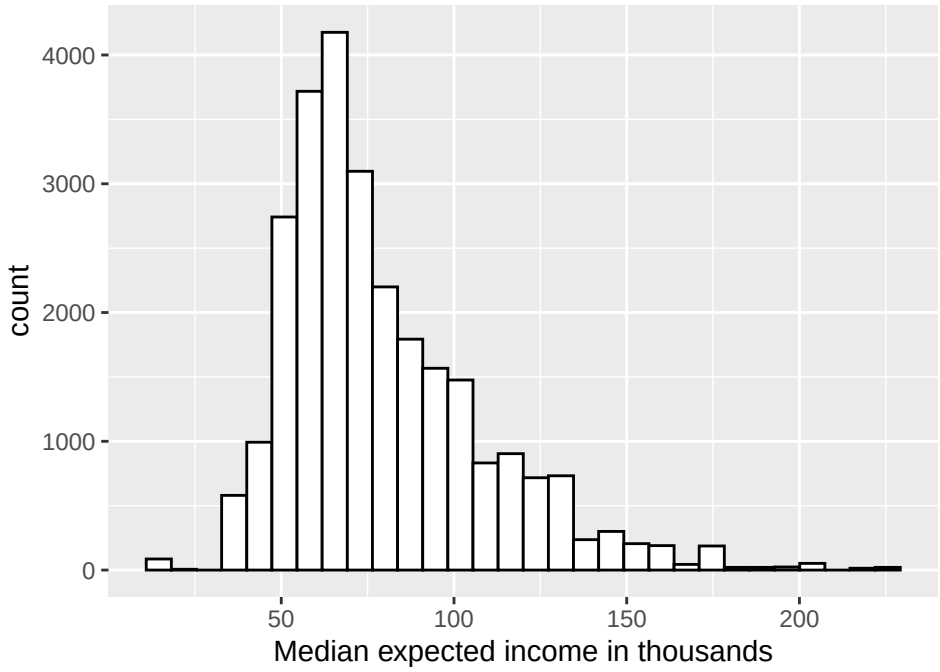


Figure 1: Income Distribution

First of all, I would like to find the relationship between the programming skills, such as R, Python, SQL, etc., and the median income posted by In the following graph, I am showing the skills mentioned in the job description in relation to the predictive median income from companies in different sizes and revenues.

As you can see in the figure 2, these skills are positively correlated with the log income. For example, the job description has the key word: python; unsurprisingly, jobs with Python mentioned in description tend to have a higher median income than those without.

However, in the figure 3, there are also skills that are not as much related to the predictive median income or even has a negative relationship on the income. The SQL skill and Excel skill seem to not be a positive influence on the income.

In figure 4, I would also like to get the relationship between the continuous predictor, rating of the companies, and predictive income. There are multiple classifications on jobs provided by different companies. For example, the location of the job, the sector of the job, and the even companies' sizes and revenues. All of these classifications can be incorporated into the multilevel regression model. In the following graph, I am showing the rating of the company in relation to the predictive median income of the job posts for companies in different sizes and revenues.

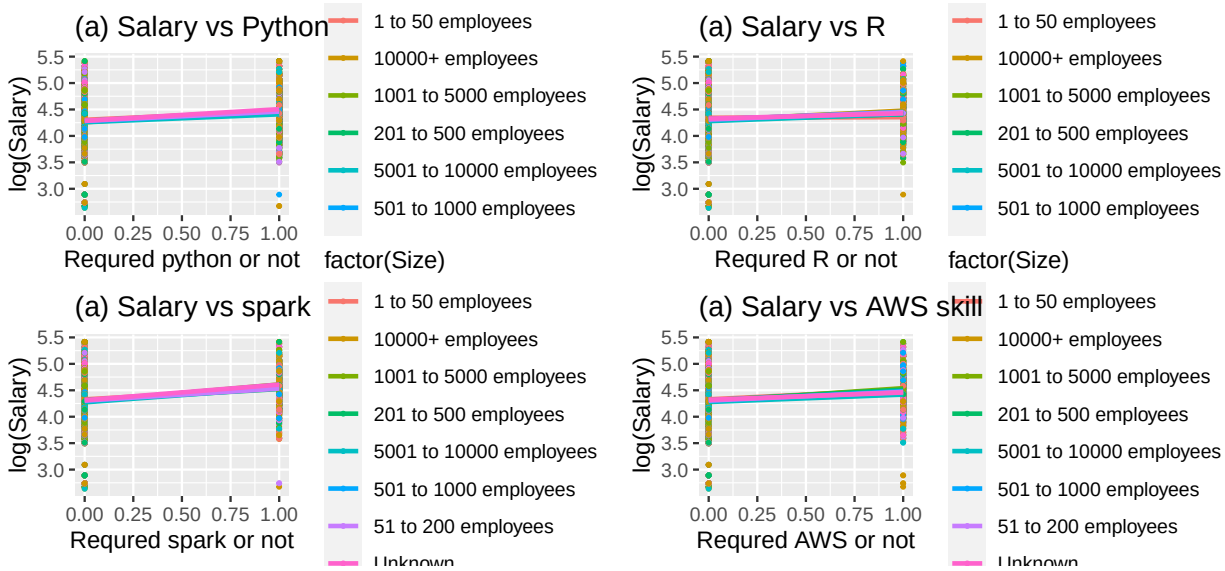


Figure 2: Positive Relation between Log Income and Skills by Size

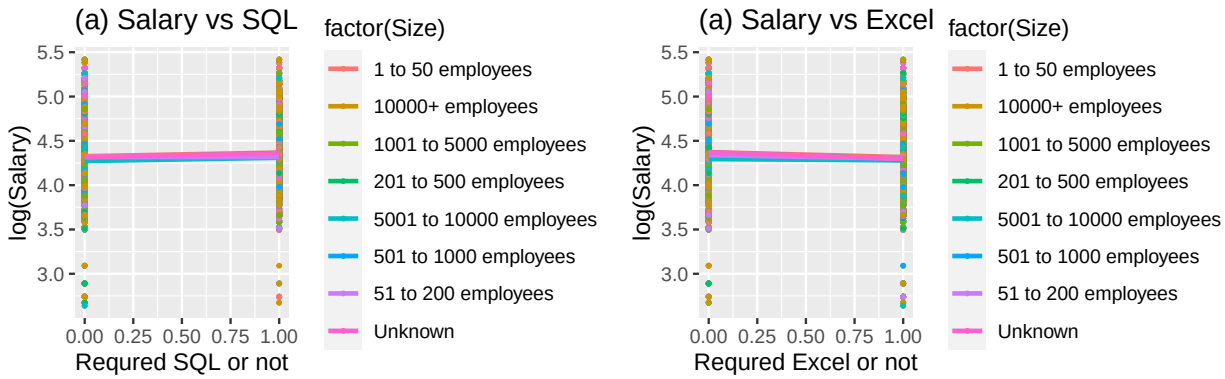


Figure 3: Non Positive Relation between Log Income and Skills

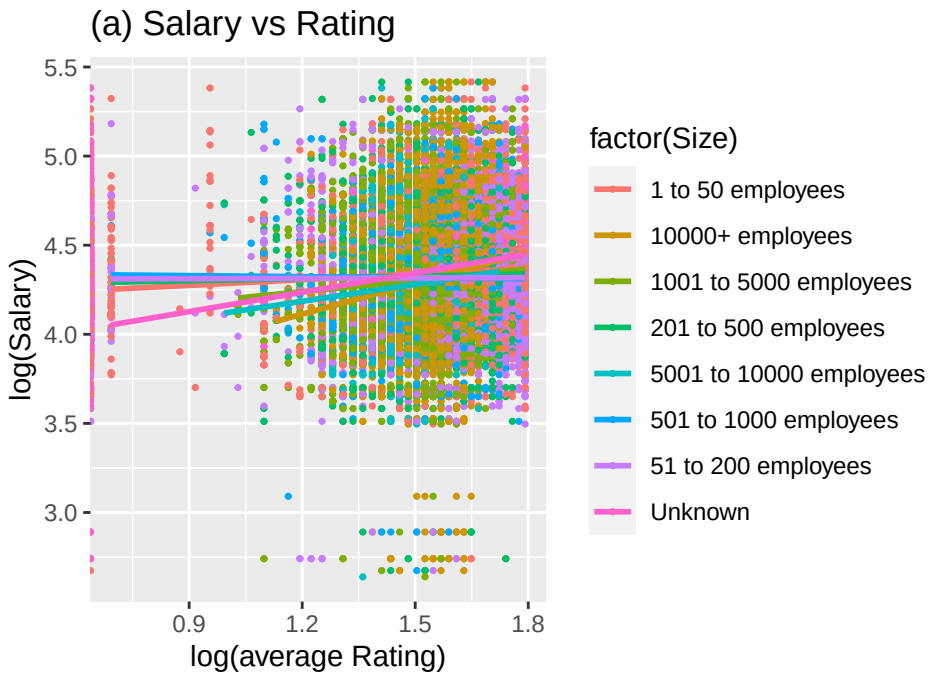


Figure 4: Relation between Log Income and Rating

Model fitting

First of all, by examine the data, the size and the revenue of the company seems to have some impact on the income. So I choose to apply the multilevel model fitting on the data. it is important to take the log on the large values of income and factorize the predictors with multiple categories, which then could be put into the model.

Since many of the predictors are got from job descriptions, those are binary predictors: 1 for True and 0 for False. To fit these predictors into the model, they are fixed. Whereas, the `Location`, `Job.Title`, `Headquarters` , `Size`, and `Revenue` are can be random effects. I first try to fit the mixed model including all the fixed and random effects. However, some of the coefficient are not significant. As a result, I only leave some of the significant random effect with high random intercept variance over the total variance; particularly, I used `Location` and `Job.Title` for random effect.

```
model1 <- lmerTest::lmer(median ~ Rating + R + python + aws + sql + spark + excel +  (1|Location) + (1|
```

From the summary table and graph of fixed effect, we can consider the predictor is statistically significance at $\alpha = 0.05$ level. In this case, the coefficients for `python`, `aws`, `sql`, `spark`, and `excel` all have p-values less than 0.05, indicating that they are significant predictors in the model. The coefficient for `Rating` and `R` does not have a significant p-value, indicating that it is not a significant predictor in this model.

From figure 5, we can see the significance of fixed effect on all predictors. As you can see, there are 2 predictors are not as significant as others: the `R` and `Rating`.

As for the random effect, as you can see, both `Job.Title` and `Location` has a high proportion relative to the residual. So the ICC score will be high.

From the figure 6, we can see a pretty close posterior predictive curve compared to the actual ones.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.3102815	0.0105013	1907.326	410.4514006	0.0000000
Rating	0.0023851	0.0034828	25842.296	0.6848112	0.4934692
R	0.0084729	0.0087941	26669.893	0.9634775	0.3353167
python	0.0752486	0.0057684	26489.621	13.0450603	0.0000000
aws	0.0162875	0.0080109	26435.373	2.0331611	0.0420462
sql	-0.0085631	0.0044045	26488.982	-1.9441765	0.0518847
spark	0.0849569	0.0086213	26568.084	9.8542619	0.0000000
excel	-0.0171688	0.0038822	26506.880	-4.4224226	0.0000098

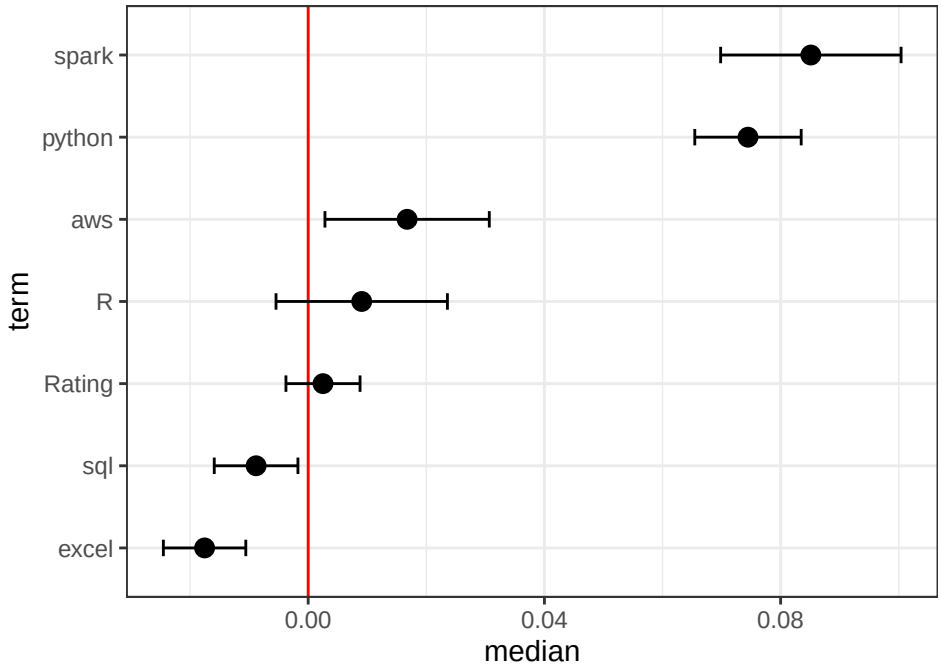


Figure 5: Fixed Effect of Median Income Model

grp	var1	var2	vcov	sdcor
Job.Title	(Intercept)	NA	0.0432938	0.2080715
Location	(Intercept)	NA	0.0348976	0.1868091
Residual	NA	NA	0.0552172	0.2349834

	(Intercept)
Aberdeen Proving Ground, MD	0.17
Addison, IL	-0.06
Addison, TX	0.03
Alachua, FL	-0.18
Alameda, CA	0.00
Albuquerque, NM	-0.15

	(Intercept)
_Data Analyst - III	-0.01
: Sr. Data Engineer	0.13
!!!100% Remote!!! Sr. Data and Integration Engineer	0.17
. Java Engineer	0.05
.Net Analyst with L1 Support Experience	-0.42
.Net Business Analyst	-0.01

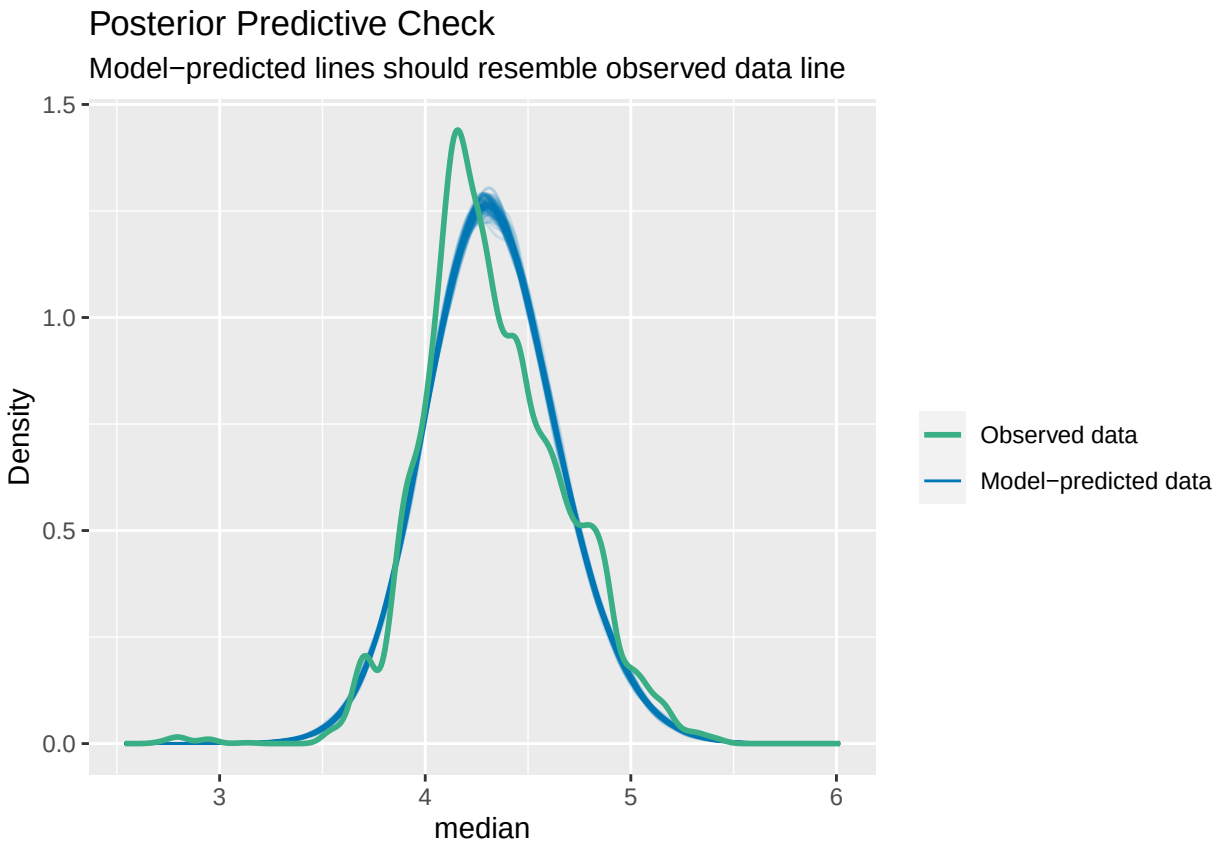


Figure 6: Posterior Predictive Check

Results

Interpretation

First, we can fit the formula with only the fixed effect:

$$log(median + 1) = 4.31 + 0.08 \times python + 0.02 \times aws - 0.01 \times sql + 0.08 \times spark - 0.02 \times excel$$

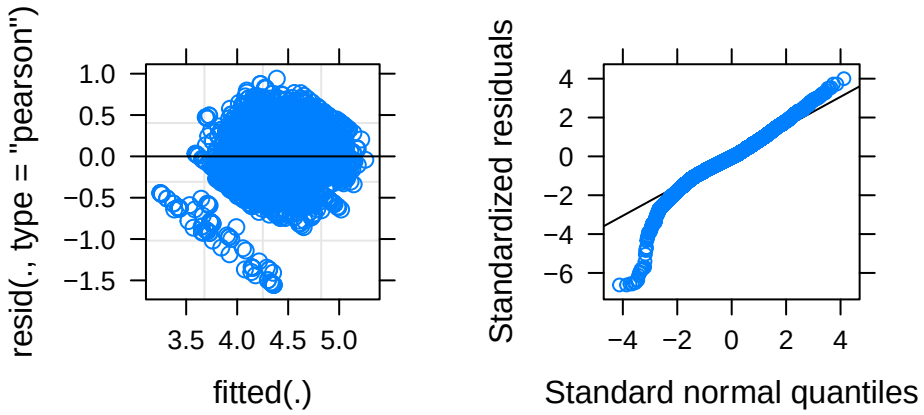
Different Job Title and Working location will also have an impact on the median income. And in the formula, the main change is on the interceptions. If we take the Sr. Data Engineer as the job title and the working location is in California, the formula will become:

$$log(median + 1) = 4.37 + 0.08 \times python + 0.02 \times aws - 0.01 \times sql + 0.08 \times spark - 0.02 \times excel$$

In both formula, Rating and R are dropped due to its insignificant p-valus. The formula can be interpret as for a Sr. Data Engineer worked in California, the median income is $e^{4.31} - 1 = 73.44$ thousands. The coefficient for `python` has a weight of 0.08, which means that if `python` is addressed in the job description, the averaged median income is $e^{0.08} = 1.08$ times higher than those isn't. The coefficient for `spark` is also 0.08 and so it can be interpreted the same as `python. aws` has a coefficient of 0.02, which means that if `aws` is addressed in the job description, the averaged median income is $e^{0.02} = 1.02$ times higher. The coefficients for `sql` and `excel` are negative, which means that job descriptions with SQL and excel mentioned on average have lower median income than those aren't. Basically, jobs with SQL will have 1% lower median income than those aren't and those with excel are 2% lower.

Model Check

On the left is the residual plot and right is the residual Q-Q plot. On the residual plot, we can see the points are kept in the -2 to 2 range and the distribution of points are relatively random. On the residual Q-Q plot, we can see the residual is roughly randomly distributed except for the tails. So we can accept the normal distribution assumption.



Discussion

In this project, I used the multilevel model to find out the elements that have an impact on the predictive median income for the data science related jobs. And the model take account in two group effects: the job's title and job's location. In short, the fixed effects aren't all as significant as I thought at the beginning and I even find negative correlation between the predictors and outcome which I've never expected before. So the

Rating of the company and the R programming skills are not statistically significant to the income. While python, aws, and spark is positively correlated to the income, SQL and excel has negatively correlation to income. As for the random effects, they are reasonably identify the difference between jobs in different titles and locations by changing the intercepts of the model.

However, there are some limitations in my project as well. First and foremost, most of my predictors are extract from the job descriptions consist of a big chunk of text. Although it will be precise for some distinctive words such as python and spark, getting the R programming skill out of the text is hard and imprecise. So I might miss some predictor which leads to an underestimate of the significance.

References

[1] https://quantdev.ssri.psu.edu/sites/qdev/files/RBootcamp_MLMInteractions_2019_0820_Final2.html

[2] https://github.com/BU-Franky/MA678_midterm

Apendix

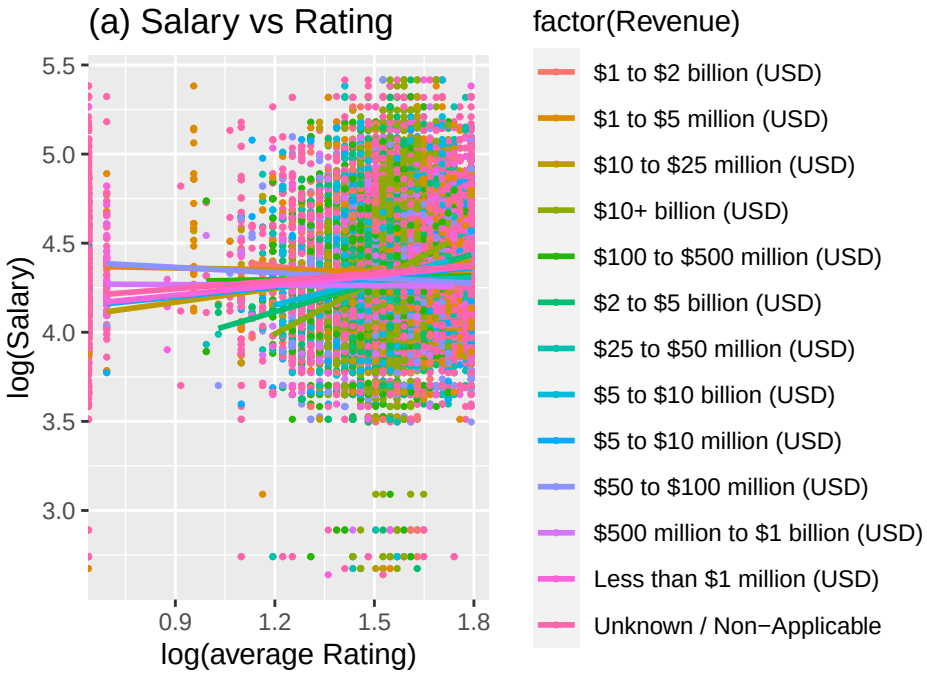
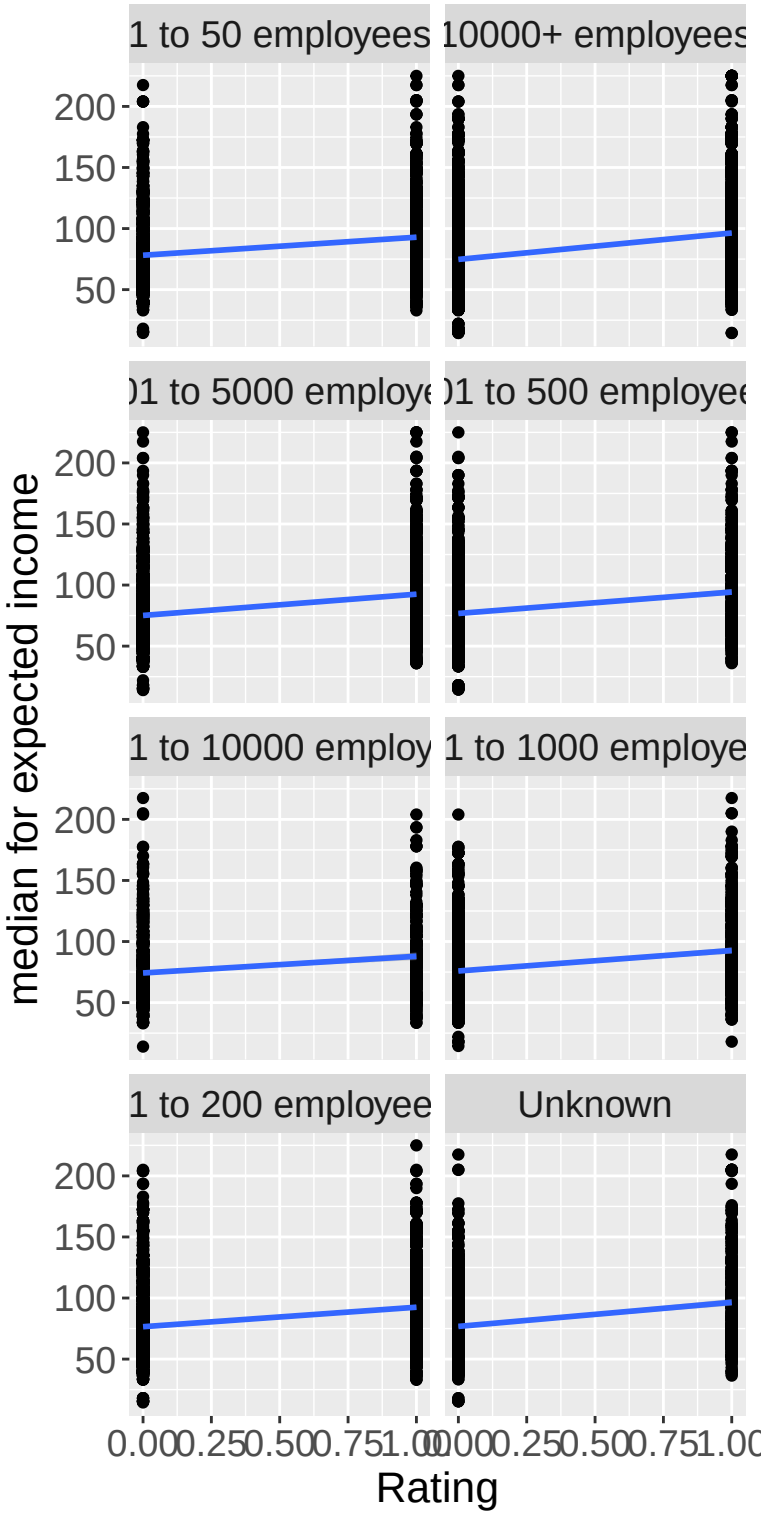


Figure 7: Relation between Log Income and Rating by Revenue

```
## `geom_smooth()` using formula = 'y ~ x'
```


1 to 50 employees
10000+ employees
)1 to 5000 employe
)1 to 500 employee
1 to 10000 employ
1 to 1000 employe
1 to 200 employee
Unknown
median for expected income
200
150
100
50
0.00
0.25
0.50
0.75
1.00
Rating