Data Analysis Project

By Dongliang (Larry) Yi

Summary

This project aims to find key factors affecting clerical employees' overall attitude rating from 30 departments. Both linear regression and Bayesian regression are used and compared in this project. The one factor model with complaints handling seems best among all models.

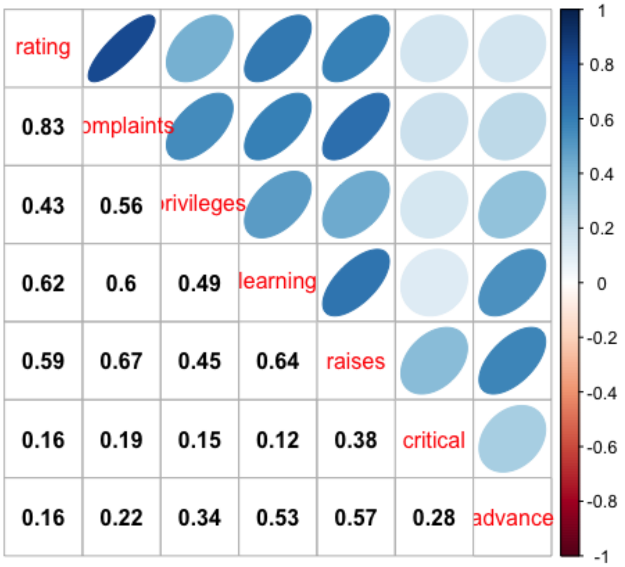
Introduction

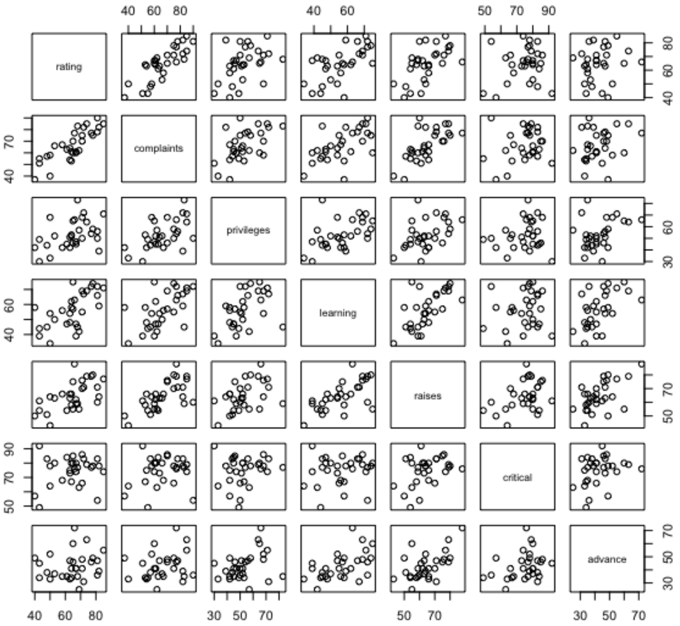
The overall employee attitude towards employers is affected by many factors. The factors may include salary, bonus, promotion, benefits, etc. In this project, I used The Chatterjee-Price Attitude Data ("attitude" in R), and analyzed key influencer on employees' overall rating towards their employers.

Data

The data used here is The Chatterjee-Price Attitude Data ("attitude") from R database. The response data is employees' overall rating("rating"), and there are seven explanatory variables, including handling of employee complaints("complaints"), special privileges("privileges"), learning opportunity("learning"), raises("raises"), too critical("critical") and advancement opportunity("advance"). There are totally 30 observations from different departments, and no missing value is found in the data.

I first plot the correlations among variables. It seems complaints, learnings and raises have high correlations with rating. The other factors do not seem to have a strong relationship with the rating.





Model

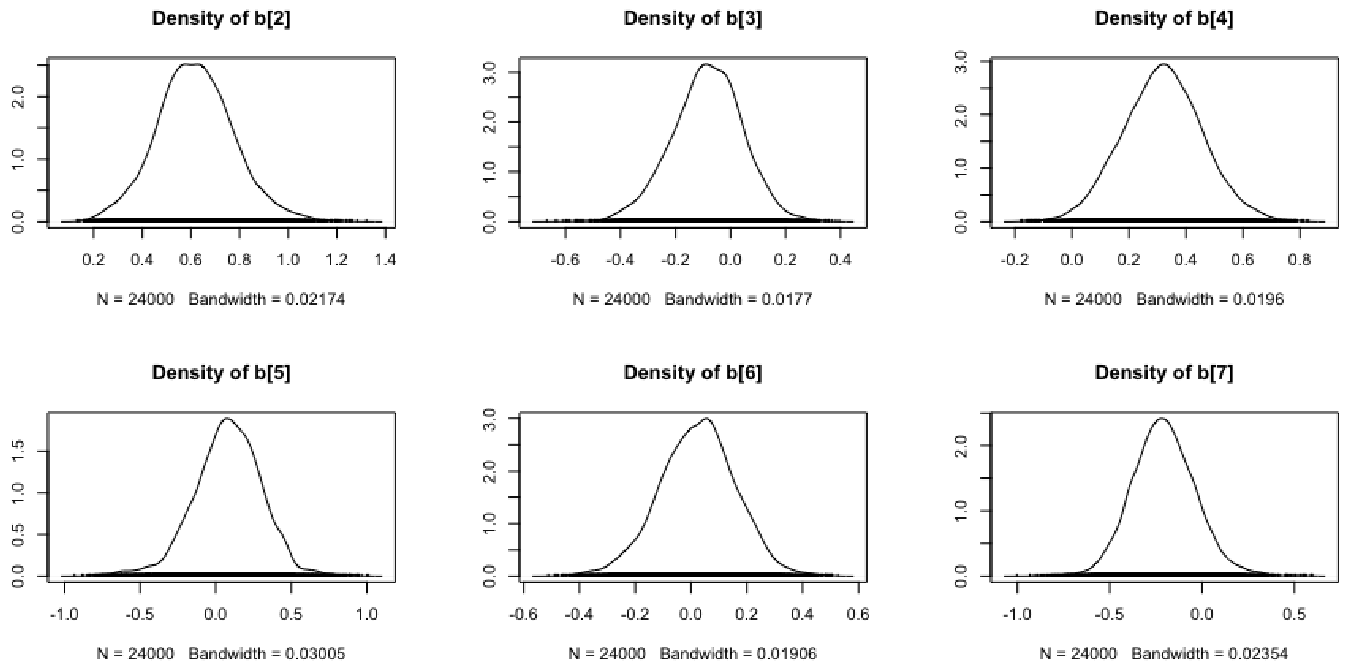
I first fit a linear regression model. Only “complaints” is significant under 5% significance level, but “learning” is the other significant factor if we release significance level to 10%. The adjusted R-square is 0.66.

Later I fit a JAGS model with all factors. Here are the hierarchical specifications:

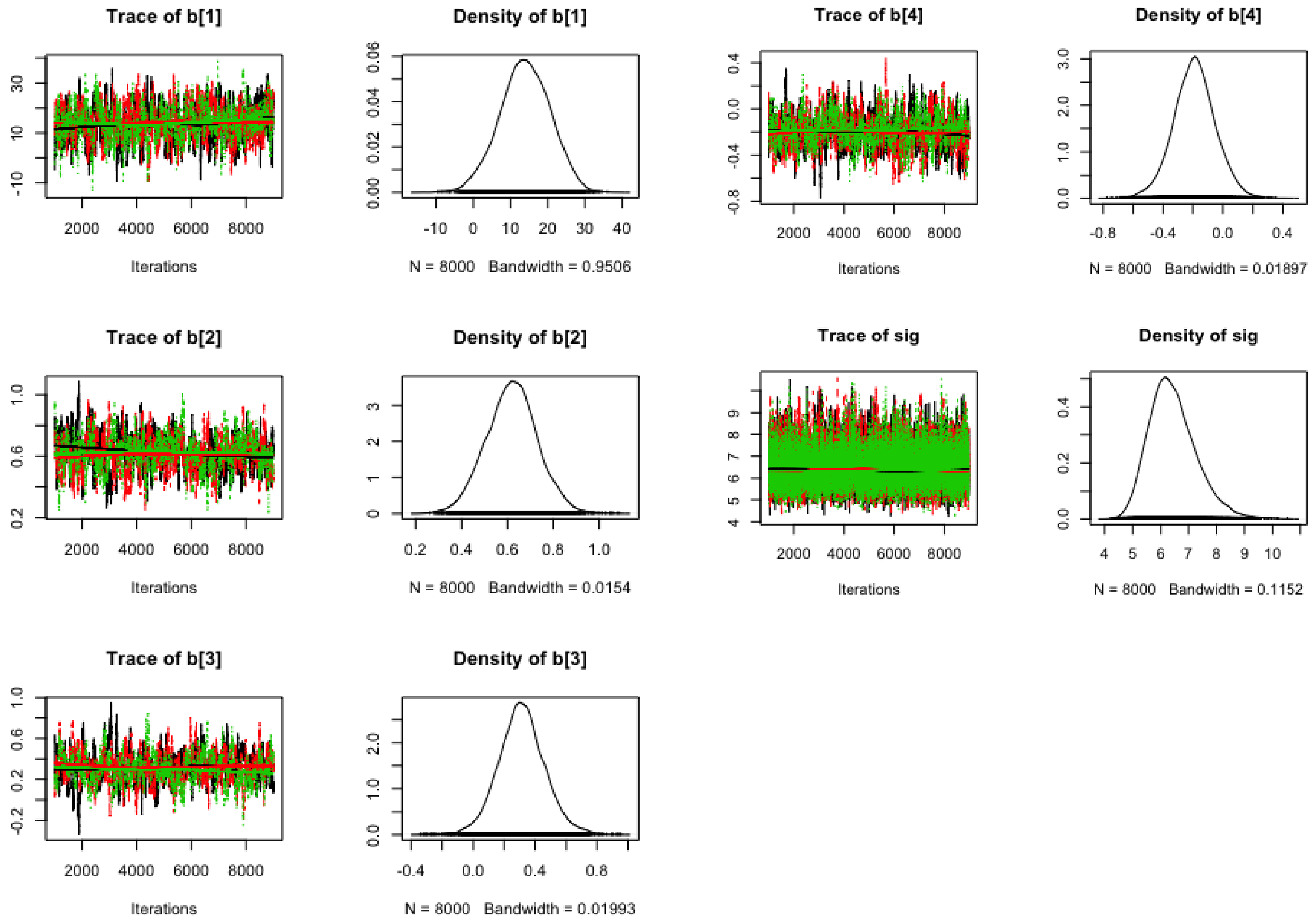
$$rating_i = b[1] + b[2] * complaints_i + b[3] * privileges_i + b[4] * learning_i + b[5] * raises_i + b[6] * critical_i$$
$$+ b[7] * advance_i + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \dots, 30$$

$$rating_i | x_i, b, \sigma^2 \overset{iid}{\sim} N(b * X, \sigma^2);\ x_i \ and\ X\ indicate\ observed\ explanitory\ variable\ and\ vector$$
$$b\ with\ prior\ normal\ distributions; \sigma^2 with\ prior\ gamma\ distribution$$

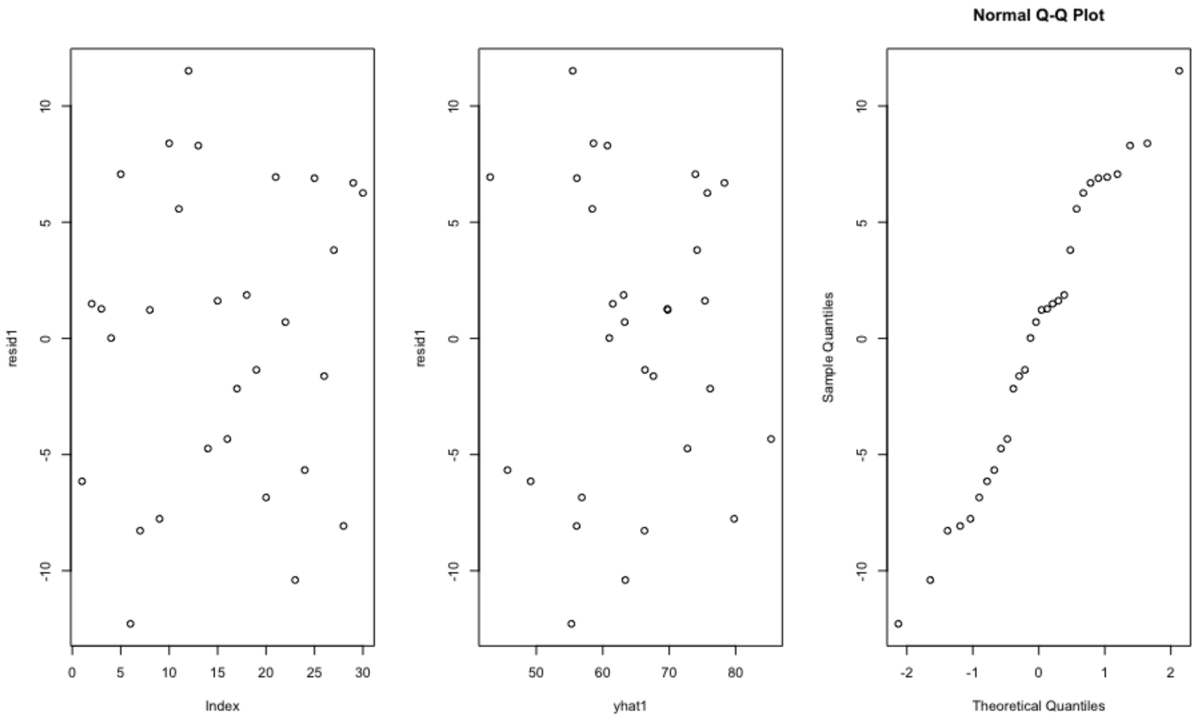
The density distribution of coefficients is shown as below. It is clear that the coefficient for variable “complaints”, “learning” and “advance” are not 0. The posterior distribution for the “privileges”, “raises” and “critical” are close to the prior which is a normal distribution with mean equal to 0.



The next model I fit is a JAGS model with intercept, “complaint”, “learning” and “advance”. I use normal distribution prior for coefficients and gamma distribution for variance. Coefficients show strong autocorrelation, but the variance looks fine. I choose to use 8000 iterators with 3 chains. The effective sizes of coefficients are 379, 312, 258 and 467 which are much smaller than variance’s 5936. So, it is better to increase iteration number if we want a confidence interval of those coefficients. According to Gelman and Rubin’s convergence diagnostics, the point estimates are smaller than the upper confidence interval, so MCMC converges.



The residual analysis is shown below. The variance seems constant and QQ plot also fits good.



Since “complaints” is the only variable which is significant in the 5% level. I fit a JAGS model with only intercept and “complaints”. As a result, this model has a DIC score of 206.036 which is slightly smaller that original three factor model’s 206.059. It shows the additional two factors “learning” and “advance” do not provide strong predictive power on the overall rating.

Result

For the JAGS model with intercept and “complaints”. “complaints” has a mean coefficient of 0.75. It means each point increase in handling complaints will expect to result in 0.75 increase in overall rating. In the JAGS model with three explanatory factors, “complaints” has a mean coefficient of 0.62. The difference can be explained by effects from additional two variables “learning” and “advance”.

Conclusion

From both linear regression and Bayesian regression, “complaints” is both the most important factor affecting the overall rating. “learning” and “advance” also have some limited influence on overall rating. “privileges” and “critical” have insignificant influence on the rating.

It is surprising that “raises” is also insignificant for predicting the overall rating considering it has a correlation 0.59 with “rating”. I think it may be caused by high correlations with “complaints” (0.67) and “learning” (0.64), so “raises” does not provide any additional prediction power after accounting “complaints” and “learning”.