Employee Performance: An Adventure of Missing Data

Matt Oehler & Josh Meyers

September 27, 2018

Abstract

Businesses and other organizations are interested in the performance of their employees. When employees are happy and motivated to work their organization is more likely to reach its business goals. On the other hand, when employees disengage from their work the company is likely to suffer. This paper looks at the effect that employee well being and job satisfaction have on job performance. The dataset of interest is missing a substantial amount of data, thus this paper will also examine the use of stochastic regression imputation in estimating missing values.

1 Introduction

The success of any business or organization is largely determined by the effectiveness of its employees. Motivated and productive employees help their organization reach its goals and earn higher profits. Disengaged employees, on the other hand, can lead to a damaged reputation and decreased earnings. Of interest to employers is the effect that an employee's happiness has on job performance. The goal of this analysis is to look at how an employee's well being and job satisfaction are related to their job performance.

2 Data

The *employee* dataset, which contains data gathered on 480 employees at a large university, is used in this analysis. The variables, and their descriptions, included in this dataset are shown in table 1. The pairs plot in figure 1 shows that the 'WellBeing' and 'IQ' variables both have a semi-strong positive relationship with JobPerf. The densities in figure 1 also show that the distribution of each variable appears to be approximately normal.

Variable	Desciption
ID	Randomly assigned employee ID
Age	Employee age
Tenure	Number of years employed at the university
WellBeing	Measure of employee's happiness at the university
JobSat	Measure of employee's satisfaction at the university
JobPerf	Measure of employee's performance at the university
IQ	Measure of employee IQ

Table 1: Description of *employee* dataset.

The problem with this dataset is that 73% of the employees are missing at least one data point. A visual and numeric summary of the missing data is given in figure 2. Note that data is only missing from the WellBeing, JobSat, and JobPerf variables. The WellBeing and JobSat variables are both missing the same number of value, but no single employee is missing both of these values. The JobPerf value is missing for 64 employees and 35 of these employees are also missing the WellBeing value. Section 2.1 will describe the method used to deal with this missing data problem.

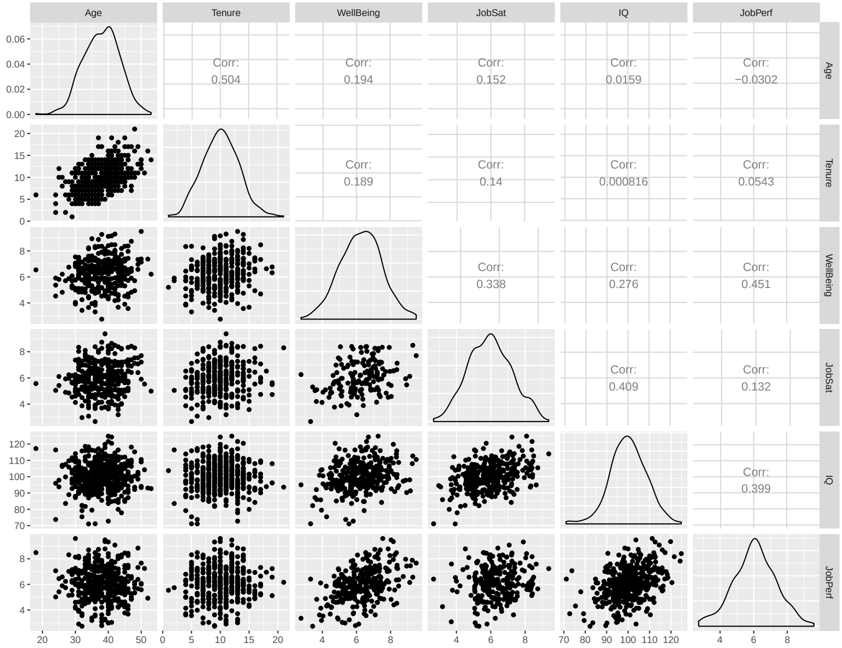
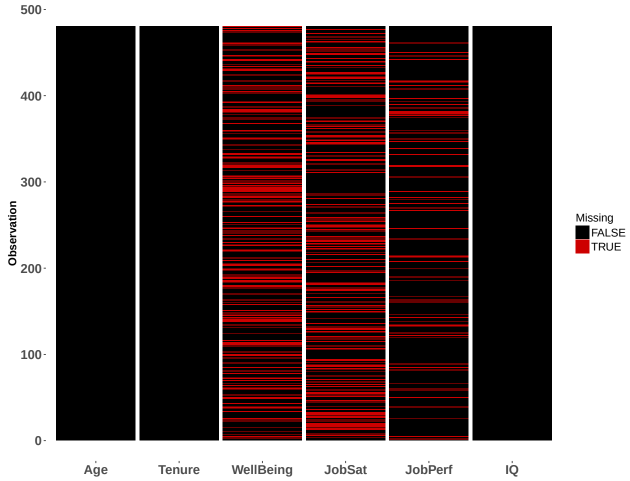


Figure 1: Summary of Data



Variable	Number Missing
Well Being	160
Job Satisfaction	160
Job Performance	64
JobPerf and WellBeing	35

Figure 2: Summary of Missing Data

2.1 Imputing the Missing Data

The missing data were estimated using stochastic regression imputation based off of a multivariate normal (MVN) distribution. This method makes the assumption that all of the variables (both explanatory and response) in this model are jointly distributed MVN. This means that each variable, their joint densities, and their conditional densities are all distributed normally. Missing values are imputed by taking draws from the conditional MVN distribution. Figure 1 shows that each variable appears to be normally distributed and that the relationship between each of variables appears to be linear. These two observations validate the assumption that this dataset is distributed as a MVN. We chose to use multiple imputation for a few reasons. Rather than throwing away all the information in the partial observations or just fill in missing observations with the mean of each variable, this method it allows us to use the data that we have to better fill in the missing values and potentially produce unbiased estimates. We do this however, at the expense of underestimating the standard error of the estimates. A summary of the algorithm used to impute data and estimate model parameters is given as follows:

1. Note which values of the original dataset are missing

2. Calculate $\boldsymbol{\mu}$, the overall mean for each variable, and $\boldsymbol{\Sigma}$, the covariance matrix for the data (where each variables variance is on the diagonal and the covariances are on the off diagonal), using all of the complete data
3. Repeat the following process M times
    - Impute each missing value by taking a random draw from the conditional distribution

      $$Y_1|Y_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$$

      Where

      $$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_2^{-1}(Y_2 - \mu_2)$$
      $$\Sigma_{1|2} = \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}$$

      and where the subscript 1 indicates which values are missing and the subscript 2 indicates the values which are not missing in the original dataset.
    - Fit the model and record the estimated parameters, their standard errors, and the r-squared
    - Re-estimate μ and Σ using the "new" complete data set
4. Pool the estimates calculated in step 3 to create final estimates.

3 Model and Methods

3.1 Model Description

To determining the effect that employee wellbeing and job satisfaction have on job performance we will use the stated multiple linear regression model:

$$\text{JobPerf} = \beta_0 + \text{Age}*\beta_1 + \text{Tenure}*\beta_2 + \text{WellBeing}*\beta_3 + \text{JobSat}*\beta_4 + \text{IQ}*\beta_5 + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where β_0 represents the intercept of the model and where $\beta_1 \ldots \beta_5$ represent the estimated effect that each variable has on JobPerf. Note that we included all of the available variables by request of the party of interest. This model gives an estimate, along with a measure of uncertainty, of the effect that each WellBeing and JobSat (along with with other variables) have on job performance. Along with these estimates we are also able to test whether or not the effect is significant in explaining JobPerf. This model works well for our imputed data because the errors of our imputed values are, by construction, distributed normally. The following assumptions must be made to properly use this model:

- The relationship between job performance and the explanatory variables must be linear.
- Each employee's job performance must be independent of another employee's performance.
- The residual terms must be normally distributed
- The residuals must be homoskedastic (meaning the they have equal variance across the entire regression line)

These validity of these assumptions will be proven below in section 3.2.

3.2 Model Diagnostics

We verified the aforementioned model assumptions using the complete observations. First, we verified the linearity assumption by creating added-variable plots (see figure 3.2) to show that each covariate has a linear relationship with job performance, after accounting for all other covariates.

Next, we verified the assumptions of normality and homoskedasticity. We did this using the plots shown in figures 4 and 5. The histogram of residuals in figure 4 confirms that the residuals are approximately normally distributed. The residual plot in figure 5 shows that the variance throughout the data set is approximately equal. Lastly we verified the assumption of independence in the residuals of job performance with critical thinking. In small companies, it might seem plausible that the performance of one employee might have an effect on the performance of another since there is more frequent interaction. However, since these data are gathered from employees that work at a large university, we determined that their job performance would be independent of one another.

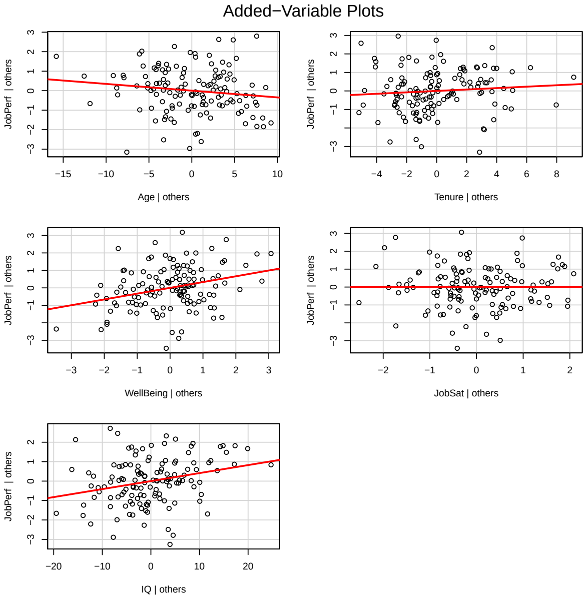


Figure 3: Added-Variable Plots for Verifying Linearity Assumption

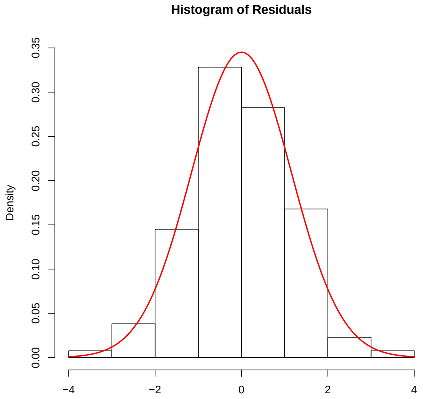


Figure 4: Histogram of Residuals

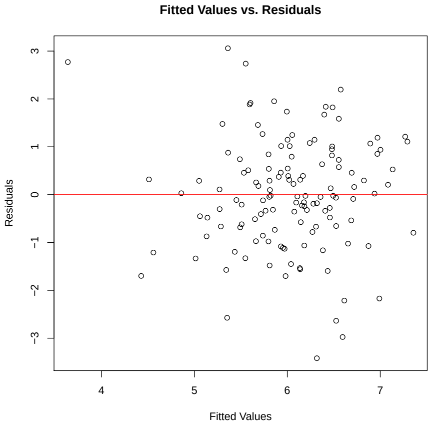


Figure 5: Residual Plot

4 Results

Using multiple stochastic imputation method described in 2.1, we imputed data for 5,500 iterations, but we threw out the first 500 iterations as burn-in values leaving us with 5,000 sample draws of the means of the imputed variables as well as the model coefficients for the models that were fit to each of the imputed datasets. To ensure that the mean estimates reached a steady-state, we created trace plots for Well Being, Job Satisfaction, and Job Performance (the variables that had missing values) which are displayed in figure 4.

Once we verified that the our sample draws reached a steady state, we calculated the pooled mean and variances of the coefficient estimates for each iteration to come up with an overall estimate and 95% confidence interval for each coefficient. In context of the previously defined model, we interpret the estimated effect of a given coefficient in the following manner (we will use Well Being as an example): Holding all other covariates constant, for each 1 unit increase on the Well Being scale, an employee's Job Performance will increase by 0.42 on average. Additionally, we are 95% confident that the true effect of Well Being on Job Performance (holding all else constant) is between 0.31 and 0.52. All of the other estimates and intervals are interpreted in the same manner.

The main effects that we are interested in are Well Being and Job Satisfaction. Well Being, as just discussed, has a positive effect on Job Performance. Job Satisfaction, however, has an

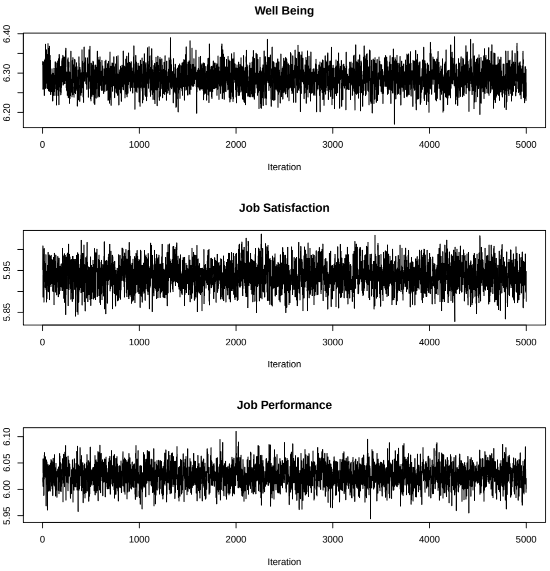


Figure 6: Trace plots of draws from conditional distribution for missing values

estimated effect that is slightly negative. This is irrelevant though since we conclude with a p-value of 0.82 that Job Satisfaction is not significant. The calculated 95% confidence interval for Job Satisfaction contains zero, which confirms that it doesn't have a significant effect on one's Job Performance. This means that companies will get the biggest return on investment in terms of Job Performance by focusing on ways to improve employee Well Being rather than attempting to improve other things such as their Job Satisfaction or IQ.

	Estimate	2.5%	95%	P-value
Intercept	-0.10	-1.55	1.34	0.55
Age	-0.03	-0.05	-0.01	1.00
Tenure	0.03	-0.01	0.07	0.09
WellBeing	0.42	0.31	0.52	0.00
JobSat	-0.05	-0.17	0.06	0.82
IQ	0.05	0.03	0.06	0.00

Table 2: Coefficient Estimates, 95% Confidence Intervals, and P-values

In terms of the general fit of the multiple linear regression model, we took the mean R^2 value for each iterated model, which resulted in an estimated R^2 of 0.30. This means that 30% of the variability in Job Performance is explained by Age, Tenure, Well Being, Job Satisfaction, and IQ (all of the variables in the model).

5 Conclusion

The goal of this analysis was to assess the relationship between an employee's hapiness and their overall performance at work. With an incomplete data set of 480 university employees, we showed that by using multiple stochastic imputation and multivariate normal regression, we could estimate coefficients with a given level of uncertainty that represent the relationship between employee happiness and job performance. By using the coefficients to better understand that relationship, companies are better able to enact policies that will target an increase in employee performance which will help them to maximize profits.

Although this analysis yielded fruitful results, it could still be improved and expanded upon. Having a larger and complete data set of the metrics we looked at could yield more accurate results by eliminating the need to impute missing data which decreased the standard error of our estimates. In addition, other variables of employees, specifically more detailed metrics on the things that contribute to an employee's overall well being could be further examined to help companies directly target factors that will maximize the performance of their employees.