Case Study for Employee Data

Kaylee Hodgson

Brigham Young University

Abstract. The analysis in this paper uses multiple linear regression (MLR) to create a model for the purpose of determining the productivity level of employees based on their job satisfaction and wellbeing. I overcome the obstacle of the large amounts of missing data by using multivariate normal imputation to fill in the missing values. The MLR analysis finds that, controlling for a couple of other factors, employees' wellbeing (or happiness level) is significant in determining productivity levels. These findings are significant because they provide employers information on how to increase employee productivity and therefore increase their profits, and provide incentive for employers to focus on improving the work conditions of their employees.

1 Introduction and Motivation

Employee productivity is a large determinant of company profit levels. Intuitively, when employees are more productive, the company's profit levels increase. Because of this, companies would benefit from understanding what factors increase the productivity of their employees. I evaluate some of the factors that I hypothesize are important determinants of an employee's job performance by analyzing data on the employees from a large university. I predict that there are specific variables that impact the employee's job performance at the university, specifically their job satisfaction and wellbeing.

The purpose of this analysis is to provide inferential conclusions to inform employers on methods to increase employee productivity. This evaluation and discussion will use multiple linear regression (MLR), a useful tool for inferential analysis, to inform employers, and specifically universities, on the best ways to improve the performance of their employees, which will likely bring in higher profits. An additional benefit of this analysis is that if, for example, happiness and job satisfaction can be tied to a company's profit, then even purely profit-motivated companies would be incentivized to ensure the wellbeing of their employees.

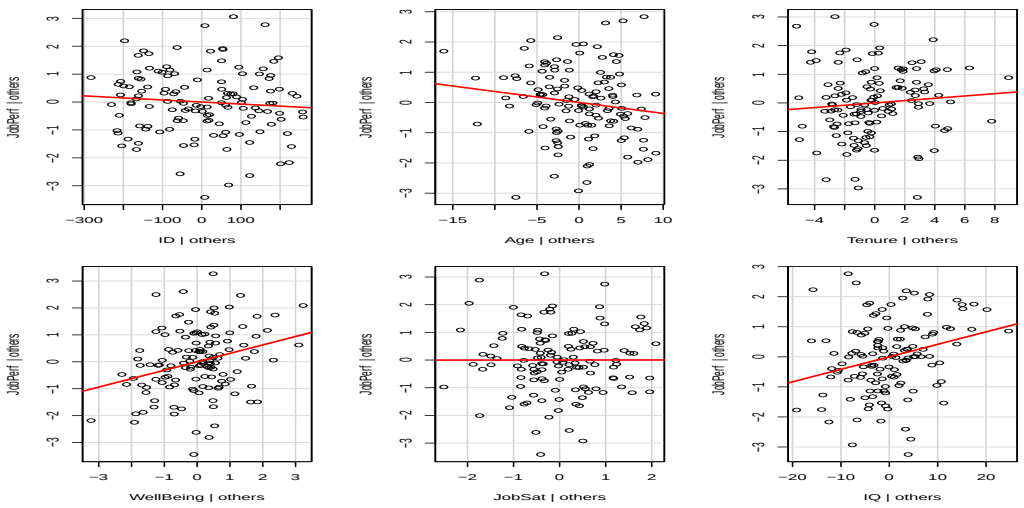
2 Data Summary and Diagnostics

This dataset includes information on 480 employees. As mentioned above, the dataset used includes the information for employees at a large university. Specifically, I look at the employees' wellbeing at the university and their satisfaction at the university, controlling for age, IQ, and tenure. The response variable of interest is the job performance of the employees. All of the variables are kept in the model because there are not many to begin with, and all are assumed to be important to the analysis. The most concerning issue that appears in this dataset is the large amounts of missing data, specifically in the particular variables of interest: wellbeing, job satisfaction, and job performance. The following section will employ an imputation method to fill in this missing data before performing the MLR analysis.

However, before addressing the missing data issue, I first test the data for the assumptions of the linear model, including testing whether there is a linear relationship between job satisfaction and each of the other variables, whether the standardized residuals of job satisfaction are normally distributed, whether the job performance measures are each independent of each other, and whether there is equal variance of the observed y values around the fitted values of the model. To test the linearity of the data, I provide added variable plots for all of the variables in the model in Figure 1. Each of the variables appear to be linear when controlled for every other variable and plotted against job performance. I do not find any specific patterns in the data that are a concern.

I next test to ensure that the standardized residuals of the job performance variable are approximately normally distributed. The first plot in Figure 2 displays the histogram of the standardized residuals. The standardized residuals closely follow the standardized normal density curve, which is plotted over the histogram,

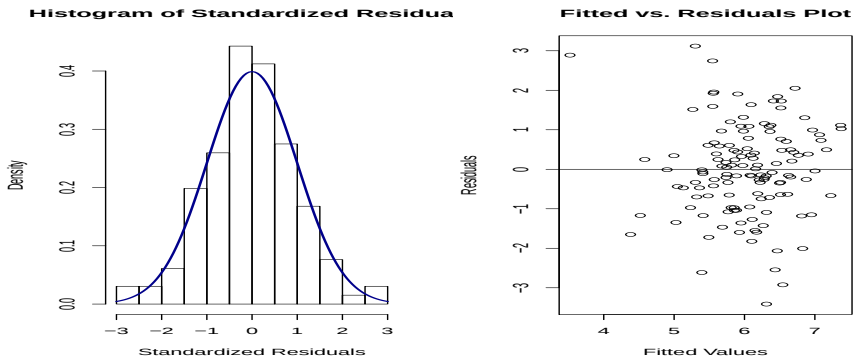
Fig. 1. Added Variable Plots



so I conclude that the data meets the assumption of normal distribution. I lastly evaluate the independence and equal variance (homoskedasticity) of the model with a fitted values vs. residuals plot. Because there are no apparent patterns in this second plot in Figure 2, I conclude that the data is independent. Equal variance could potentially be a concern because the data is clumped in a certain area on the table, with a couple of points on the tails. However, the difference in variance is not dramatic enough to cause concern, especially because this may be caused by the missing data, and there only appears to be one observation on the lower end that extends the plot. This outlier is not removed from the dataset because it provides unique information to the model, and is used in the function to impute missing values. I conclude that this dataset meets the assumptions of a multiple linear model, and determine to proceed with a MLR analysis.

The last diagnostic test I run on the data is a test for multi-collinearity, which would mean that the variables are too highly correlated with each other. If any of the variables are too highly correlated with each other, this can mess with the results of the MLR analysis. I run a variance inflation factor (VIF) test and find that none of the variables' correlations are a concern in this analysis (all VIF's are less than 10). While the data appear to closely follow the assumptions of the linear model and are not collinear, there still remains the issue of the large amounts of missing data. This will need to be resolved before an analysis can be accurately evaluated.

Fig. 2. Tests of Linear Assumptions

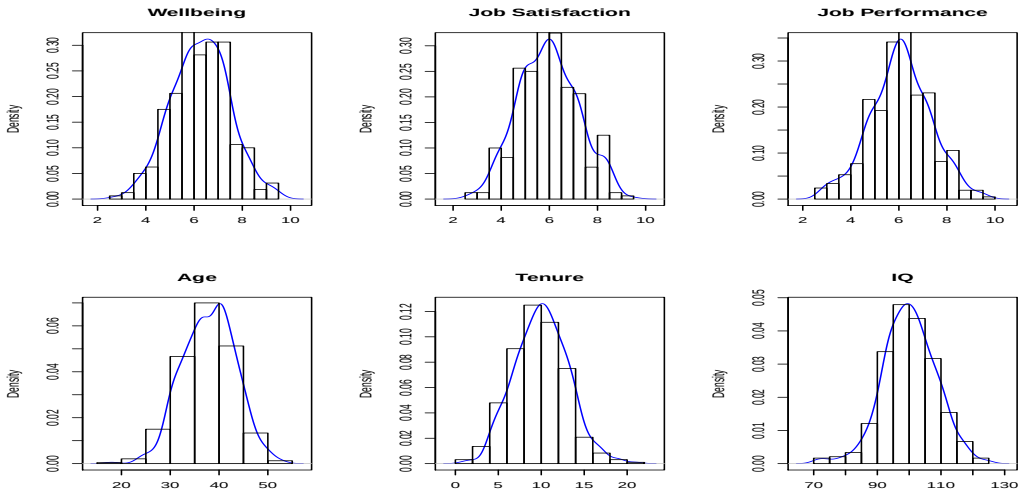


3 Imputation of Missing Data

There are high percentages of missing data in this dataset, with 33% missing from job statistfaction, 33% missing from wellbeing, and 13% missing from job performance. Because of the large amount of missing information, I choose not to throw out any of the subjects, and to instead impute the values using a multivariate normal imputation approach.

Before the missing data values can be imputed, another assumption has to be verified. Since I employ a multivariate normal imputation method, all of the data in each of the variables has to be normally distributed. Figure 3 shows the histogram and density curves of each of the variables in the dataset, and they all appear to closely follow a (symmetric and unimodal) normal density curve, so I can proceed with this imputation method.

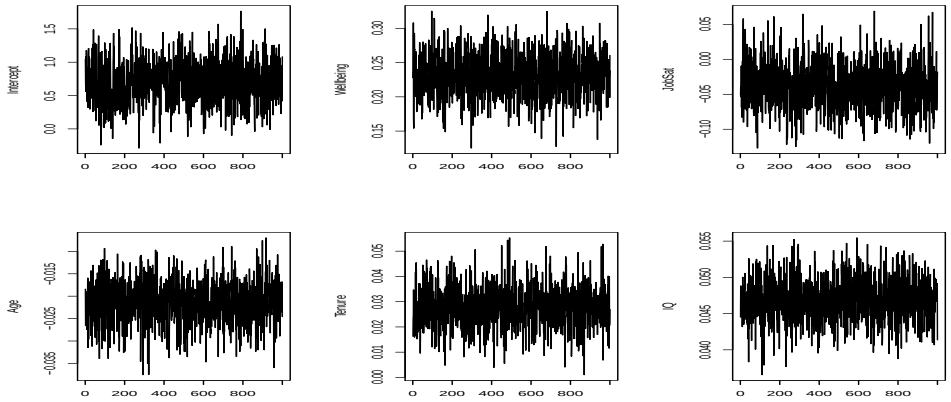
Fig. 3. Tests of Multivariate Normal Assumption: Density Plots



Because the values in the variables are each normally distributed, I impute the data under the assumption that the missing values are also normally distributed. When the variables are combined into a joint distribution, because they are individually normally distributed, they will be approximately multivariate normally distributed. The process to impute the missing data is as follows:

1. Initial values for the mean vector and the variance-covariance matrix were estimated from the original dataset. The mean vector contains the average values for each of the variables and the variance-covariance matrix contains the variance of the individual variables on the diagonal and the covariances between variables off-diagonal.
2. For every missing data point, values were randomly drawn from a multivariate normal distribution with the mean and variance-covariance values calculated in step 1.
3. Coefficients (measures of the average effects of all of the other variables on job performance) and their standard errors were estimated from the new dataset with the imputed values.
4. A new mean vector and variance-covariance matrix were estimated from the new imputed dataset
5. Steps 2-4 were repeated M=1000 times
6. Convergence of the coefficient values was verified by evaluating the trace plots found in Figure 4, which all showed uniform patterns. This indicates the the values of the coefficients were estimating in similar ranges, centered around the same area, throughout the 1000 iterations.

Fig. 4. Coefficient Estimation Trace Plots



4 Model and Methods

Once the iterations to impute the data were completed, the coefficients and standard errors were pooled using the following equations:

$$\tilde{\beta}_i = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{im} \tag{1}$$

$$SE(\tilde{\beta}_i) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}SE^2(\hat{\beta}_{im}) + \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_{im}-\tilde{\beta}_i)^2 + \frac{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_{im}-\tilde{\beta}_i)^2}{M}} \tag{2}$$

These pooled values are the estimates for the multiple linear regression model. The model is written as:

$$\text{Job Performance} = \beta_0 + \text{Wellbeing}^*\beta_{Wellbeing} + \text{JobSat}^*\beta_{JobSat} + \text{Age}^*\beta_{Age} + \text{Tenure}^*\beta_{Tenure} + \text{IQ}^*\beta_{IQ} + \epsilon,\ \epsilon_i \sim N(0,\sigma^2)$$

The β values in the model above are the MLR coefficients. The coefficients are estimates of how changes in wellbeing and job satisfaction, along with the other control variables, impact job performance on average. β_0 is the intercept of the multiple linear regression line, which indicates the average value of job performance when all other variables are at initial (0) values. The values of the $\beta_{Wellbeing}$ and β_{JobSat} coefficients are slopes, which measure how much, on average, job performance scores change as wellbeing or job satisfaction change by one unit, all other variables held constant. The other coefficients describe the effects of the variables that the model is controlling for. This variable allows inferential conclusions, accounting and controlling for other variables that may affect job performance.

Using the pooled values of the coefficients and their standard errors, t statistics, p values, and confidence intervals are calculated for each of the coefficients. The degrees of freedom for each of the coefficients is calculated differently than other models because of the imputation process. The new function for degrees of freedom is as follows:

$$df_i = (M-1)(\tfrac{1}{FMI^2}), \text{ where } FMI = \frac{SE_i^2 - \frac{1}{M}\sum_{m=1}^{M}SE^2(\hat{\beta}_{im})}{SE_i^2}$$

The t statistics, p values, and confidence intervals are used to find the significance of each of the variables on job satisfaction. Because there are multiple tests performed on the model, a Bonferroni correction is used to adjust the $\alpha = 0.05$ for these tests. Six coefficients are tested in the model for significance, so the significance level to reject the null hypothesis is adjusted to $\alpha* = \alpha/6 = 0.008$, which is a conservative adjustment. The p-values are compared to, and confidence intervals are calculated using, this adjusted significance value.

5 Results

The model summary can be found in Table 1. Coefficients with p-values lower than 0.008 are considered significant. Specifically, wellbeing is a significant determinant of job performance. The estimate of the coefficient tells us that, on average, job performance scales increase by 0.231 as wellbeing levels (measured by the happiness scale) increase by 1 point. The confidence interval bounds indicate that we can be $(1-\alpha*) = 99.2\%$ confident that the change in job satisfaction as wellbeing changes is between 0.097 and 0.365. Because the slope 0 is within the confidence interval for job satisfaction, which would mean that job satisfaction has no effect on job performance, I cannot conclude that job satisfaction has a significant variable to the model. These conclusions are clear in the added variable plots back in Figure 1, which show a positive relationship between wellbeing and job performance, and no clear relationship between job satisfaction and their performance.

Because of the unique way this data was analyzed, with all of the imputed values, the R-squared value was also pooled for the model. At each iteration, the R-squared for the model was saved, and the average value and quantiles of the distribution of the R-squared values are recorded. The mean of the R-squared values is 0.17, with a 95% confidence interval of (0.14,0.20), which indicates that 95% of the R-squared values were within this interval. This mean R-squared value indicates that, on average, 17% of the variation and change in an employee's job performance can be explained by this model. One of the shortcomings of the model is this low R-squared value, which indicates that the variables included do not explain very much regarding how job performance changes.

Table 1. Model Summary

Coefficients	Estimate	Lower Bound	Upper Bound	Standard Error	t value	p-value
β_0 (Intercept)	0.668	-1.273	2.610	0.811	0.824	0.205
$\beta_{Wellbeing}$	0.231	0.097	0.365	0.056	4.131	0.000*
β_{JobSat}	-0.037	-0.171	0.097	0.056	-0.657	0.745
β_{Age}	-0.021	-0.051	0.008	0.012	-1.731	0.958
β_{Tenure}	0.028	-0.023	0.079	0.021	1.298	0.097
β_{IQ}	0.047	0.030	0.064	0.007	6.461	0.000*

Despite this, we find significant results regarding how happiness effects job performance in this MLR model that provide important information for companies.

6 Conclusion

The results of the MLR model indicate that, while job satisfaction has no significant effect on job performance, wellbeing does signicantly impact job performance. These results provide crucial information for companies looking for ways to increase their profits. Companies looking to increase their employees' job performance in order to bring in higher profits should consider the results of this study, and work to improve the wellbeing of their employees. These findings are also significant in providing incentive to help improve working evironments for employees. The results provide profit-movitivated encouragement for employers to invest time and money into bettering the work conditions and improving the happiness levels of the companies' employees.

Employers and future researchers should perform further evaluation to find the most efficient and effective ways to increase the wellbeing of employees, and should also explore models that explain more of the variation in job performance (models with higher R-squared values).