Employee Performance

David A. Lowe

February 9, 2017

Abstract

LOWE, D. A. Understanding the relationship between employee happiness and employee job performance. **Purpose:** To a) identify happiness related variables that have a significant effect on job performance, and b) find other characteristics of employees that have a significant effect on job performance. **Methods:** 480 observations were collected from a university with information about individual employees. These data have a large proportion of missing values. Because the data is assumed to follow the multivariate normal distribution, multiple imputation is used to impute missing values and retain variability in the estimates. Multiple linear regression is used to estimate the effects of each variable on job performance, and an aggregate value of each iteration is used to provide overall estimates from the model. **Results:** The employee's wellbeing, age, and IQ are found to have significant positive relationships with job performance, while job satisfaction is found to have a negative influence on job performance. The R^2 from the overall model is 0.304, indicating that further variables could help explain the variability in job performance. **Conclusion:** An employee's job performance depends on wellbeing, age, IQ, and job satisfaction, though further information is required to understand what factors go into job performance.

1 Introduction

Companies are constantly concerned with getting better performance out of their employees. As the ability and efficiency of employees increase, so does the company's success and profits. Optimizing job performance is a task that has been hypothesized over for centuries. One of the many varying philosophies is that job performance is based on the happiness of the employees. Happiness can come from a variety of sources, including the employee's health, intelligence, stage of life, job security, and fulfillment in their job. From this analysis a company would hope to understand the relationship of happiness with job performance. It can also influence hiring decisions so as to acquire employees who most likely foster success. This type of study can be used to develop company culture, work-life balance, and compensation plans, and will hopefully lead to an increase in the company's revenue and value. With these goals in mind, data were collected on 480 employees at a university.

1.1 Data

To collect relevant data, there are variables observed from various sources with the purpose of painting a more complete picture of what goes into an employee's job performance. Various assumed sources of happiness were measured to see this effect. Each variable can be argued to add to or subtract from an employee's happiness as a collective. Employees at the university rated their own wellbeing and job satisfaction on a scale from 1 to 10. Also included in the dataset is the number of years the employee has worked at the university, their age, and their IQ. These will all be used as explanatory variables for job performance, which is the variables we are trying to measure and understand from this study. Job performance, a variable assumed to be given by the employer, is a rating between 1 and 10 as well. The model will use each of these variables to understand what type of employee performs best at work. Figure 1 plots each of these variables against each other. As can be seen, there are no obvious collinearity problems, and the data seem to have a normal relationship between job performance and the others. One slight issue with the data is that job performance only ranges between 0 and 10. This stretches the assumption that the data are distributed normally, however there is evidence in the data to suggest that job performance follows a normal shape (see Figure 2), and we will continue through the analysis under the assumption of normality.

Since we are most interested in the relationship between job performance and the previously mentioned variables, the model will focus mostly on producing statistically significant estimates of the effects of wellbeing, job satisfaction, and the other aforementioned variables on job performance. Multiple linear regression (MLR) is a useful analysis technique that provides estimates of the effect of each variable on a clearly defined response variable. This can be an appropriate tool for measuring how wellbeing, job satisfaction, IQ, age, and tenure actually affect the job performance of employees. Before moving into the model development, however, the data must be found suitable for analysis.

One major concern in fitting a multiple linear regression model is the amount of data available and missing data values. While there are 480 total observations in the dataset, there are only 131 rows of the dataset that have a value for every variable. Table 1 shows combinations of missing values, and the frequency of the missing combination in the dataset. In the table, a 1 represents the

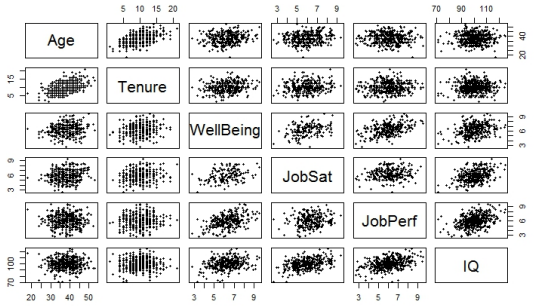

Figure 1: Plot Matrix of All Variables

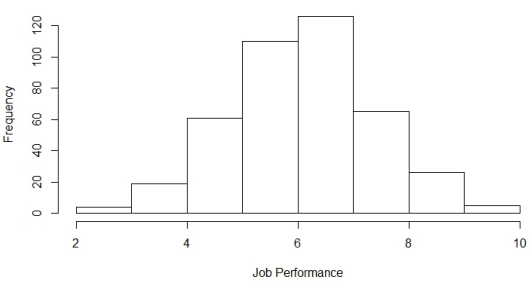

Figure 2: Histogram of Job Performance

Table 1: Missing Observation Combinations

	Age	Tenure	IQ	JobPerf	WellBeing	JobSat
131	1	1	1	1	1	1
125	1	1	1	1	0	1
160	1	1	1	1	1	0
29	1	1	1	0	1	1
35	1	1	1	0	0	1
0	0	0	64	160	160	384

observation is present, and a 0 represents a missing value. If we were to take out all observations with missing data we would greatly reduce the amount of usable data. We may miss out on important information given by missing data, such as a specific reason the data was not given. Notably, employees may have chosen to leave job satisfaction and wellbeing unanswered if under a certain threshold, possibly hoping not to notify their employer of their dislike for their position. These types of concerns make it important to use some type of data imputation to complete the fragmented dataset.

2 Methods & Model

2.1 Data Imputation

Data imputation provides a means by which missing data can be estimated. Some methods are better than others for our given dataset. Because we are not sure of the relationship between each variable, it can be hard to impute missing values. Mean imputation relies on the mean of the variable, missing values excluded, and then fills in missing values with the mean only. This method can bias the model, and stretches the assumption of linearity in the data. It also reduces the variability of the data, and the correlations between the explanatory variables and the response. These attributes of mean imputation disadvantage this approach as an appropriate method for this variable.

Stochastic regression imputation, a second approach to data imputation, allows us to add some variability to the imputed values. It is done by using regression on the observations we do have to estimate the missing values, adding a source of error from the standard errors. This takes away the bias that is caused by mean imputation. This method, however, may underestimate the true standard errors, or variability in the data.

A third, and more appropriate method presents a way in which the variability and expected value can be estimated based on all of the available observations. Multiple imputation assumes the data is jointly distributed as a multivariate normal. The joint multivariate normal distribution assumes that each variable interacts with the others at the same time, which allows prediction of any variable(s) based on the others. To do this we find the conditional distribution of each missing value combination. From Table 1 we see that there are four different combinations of missing values. Each of these combinations has a conditional distribution from which we can draw the missing values. This meets our previously defined goal of understanding how happiness affects job performance. As we understand the relationship of this multivariate normal distribution, we are able to more completely measure the effect of each variable on job performance.

The assumption that the data come from a multivariate normal must be verified before continuing with the multiple imputation method. In Figure 3, the conditional distributions are shown to be distributed normally. These plots map the residuals of each conditional model (variable given all other variables) on the theoretical quantiles of the normal distribution. They follow the plotted red line fairly closely, indicating that they are normally distributed, and therefore the data is assumed to be from a multivariate normal distribution. The assumption now being met, we can draw random values from the conditional distribution of each corresponding missing variable combination with the multivariate normal, or the univariate normal, depending on how many missing values there are in the observation.

Once all of the missing data is imputed, we can run a multiple linear regression on the observed and imputed data to get estimates of the effect of age, tenure, wellbeing, job satisfaction, and IQ on job performance. After saving the estimate values, or coefficient values, the standard error of

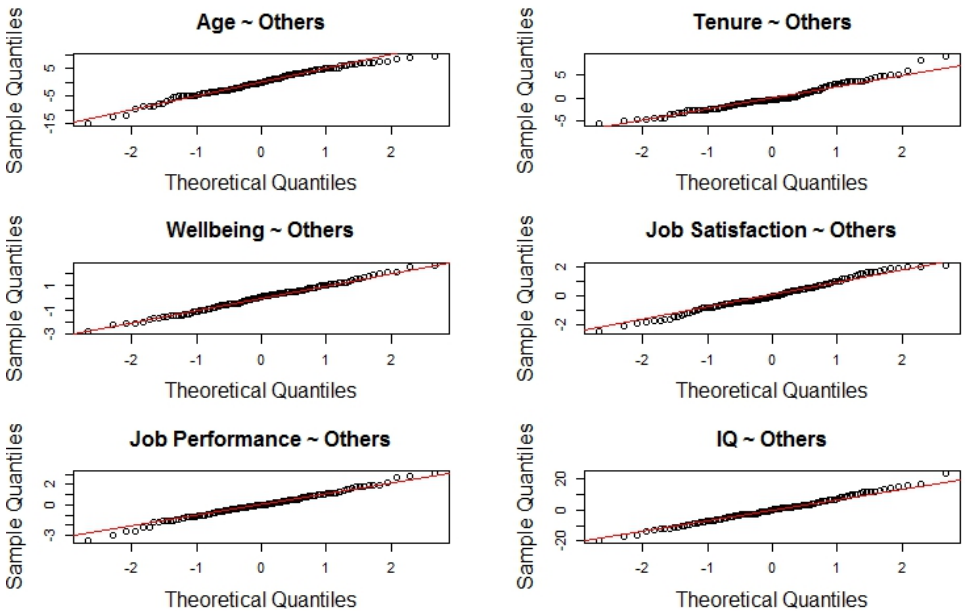

Figure 3: Q-Q Plots of Each Variable

the coefficients, and the R^2 values produced from the model, the newly imputed dataset is used to update the conditional mean and variance for the missing values. To ensure we capture the true range of values for each missing value, this process is repeated 10,000 times. From this iterative process, we obtain a large sample of estimates for the effect of our explanatory variables that can give us insight on the relationship between happiness and job performance. Each observation, along with the uncertainty of the true value, is used in the model to obtain the estimates.

2.2 Model

The goal of this study is to understand the relationship of each of the given variables on an employee's job performance. Each explanatory variable can represent the happiness of an employee in some way, so no variable selection will be done to reduce the model. The model can be represented in the following form:

$$\begin{aligned} JobPerformance_i = \beta_0 + \beta_1(Age_i) + \beta_2(Tenure_i) + \beta_3(Wellbeing_i) \\ +\beta_4(JobSatisfaction_i) + \beta_5(IQ_i) + \varepsilon_i \end{aligned} \quad (1)$$

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

This model represents the i^{th} employee's job performance, given the employee's age, tenure, wellbeing, job satisfaction, and IQ. Each β represents the estimated effect, or coefficient, of that variable on job performance. The ε_i represents the error, or distance between the i^{th} employee's job performance and the estimated value produced from the β estimates and the given values of the explanatory variables. Because we are not interested in prediction, we will focus mostly on the significance and magnitude of the estimates, β, for each variable. From these estimates we can decide which of the variables account for an increase in job performance.

The β estimates found in equation 1 are found by calculating the mean of the estimates from each of the 10,000 iterations of running the model with the imputed data. Each has their own uncertainty, which we will discuss later. To make sure our estimates of the employee characteristics and ratings are correct, the estimated values of each variable must converge. Figure 4 shows the estimate value from each multiple linear model run over the 10,000 iterations. These plots show that the estimates explored the possible values of the variable's estimated effect, incorporating the variability we acquired from using the multivariate distribution to model the data.

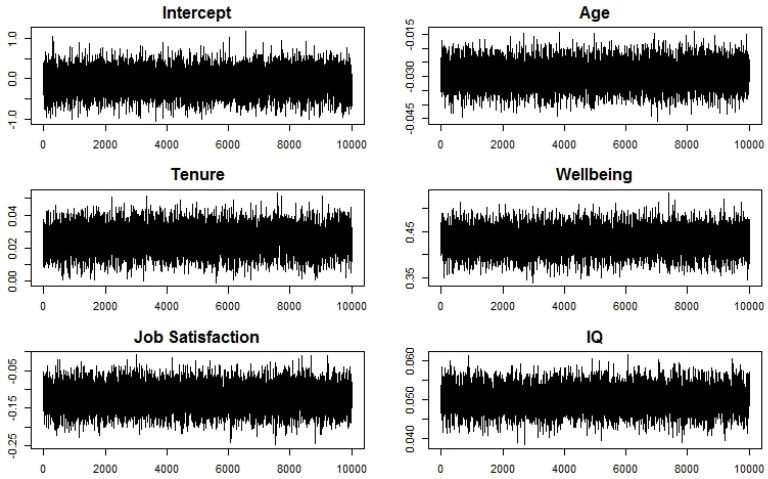

Figure 4: Trace Plots of β Estimates

Our initial goal of finding the estimates using every possible data point is met. We are also adding the necessary variability in the missing data to account for the uncertainty in the true values of the missing observations. By assuming the data follows a multivariate normal distribution, we account for the initial problem of missing observations by both including each data point in the dataset, and accounting for the range of values each of these missing observations can take. Because these coefficient estimates each converge on a finite value, we can use the estimates to get an overall estimate of all iterations. With these iterations of β we can calculate an overall estimate for each variable's coefficient, or estimate. Each overall estimate helps us understand the strength and significance of the variable on explaining job performance. We will further discuss this model's ability to represent job performance through these variables.

3 Model Justification & Performance Evaluation

This model's purpose is to explain how happiness affects job performance. While job satisfaction and wellbeing seem like more obvious indicators of happiness, there are other factors that go into happiness, such as job security, the ability to learn, and maturity. These are demonstrated through the variables age, tenure, and IQ. Even if not considered direct indicators of happiness, these variables can give us insight on other factors in an employee's life that will positively affect job performance. This is one of the reasons I have left each available variable in the model. Variable selection also becomes more complicated when imputing data. Another reason to keep all variables is our assumption that the data is distributed jointly as a multivariate distribution. From this assumption we can say that each variable affects the values of all the rest, and therefore all should be included in the model.

3.1 Assumptions

Model assumptions allow us to have confidence in the results of the model. Because this model uses multiple imputation instead of a complete dataset, some of the assumptions are not as clearly represented as with models that do use a complete dataset. To test the assumptions of this model I will use the original 131 observations with no missing values. As the assumptions are met in this limited dataset, they can be extended to the imputed dataset by the multivariate normal assumptions made earlier.

Linearity is the first assumption needed to verify a model's validity. Added variable plots demonstrate the linearity of age, tenure, wellbeing, job satisfaction, and IQ with job performance. All show similar linearity patterns after adding all other variables (see Figure 5). These plots show that each variable has a linear relationship with job performance. Linearity helps to justify the validity of the linear model, which allows us to draw inference from the estimates of each variable, as is our goal.

The next assumption is independence in each employee's job performance. Intuitively, the

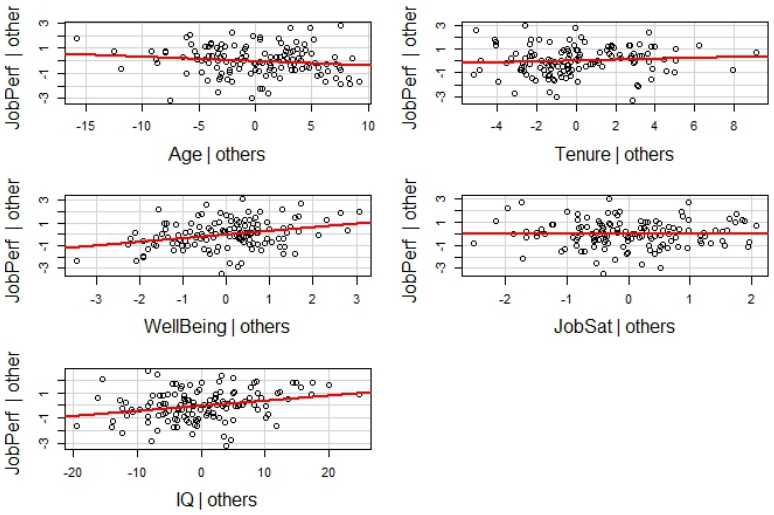

Figure 5: Added-Variable Plots for Each Variable

job performance of one employee can potentially affect the job performance of other employee's, however we will assume that those who collected the data picked 480 random employee's that unlikely perform based on the other employees' job performance. A fitted vs. residuals plot solidifies this claim. In Figure 6 the residuals show that there is no clear pattern in the data, signifying that the job performance of one employee does not affect the performance of another. Independence comes with many useful properties, one of which is assurance that there is not an added source of variability in the model that we are not accounting for.

Another assumption necessary for the model's efficacy is normally distributed residuals, or distances between the actual values of job performance and the model's expected values. This can be shown with a normal q-q plot (see Figure 7). As explained earlier, normality is assumed when the fitted residuals fit the theoretical quantiles of the normal distribution fairly closely. Figure 7 shows that this assumption is not unreasonable. Normality allows us to make valuable assumptions about the estimates found through the model. As the number of observations increase, the estimates converge on the true value of the variables effect on job performance.

The last assumption needed for a linear model is equal variance in the residuals. Figure 6 demonstrates that the residuals are heteroskedastic. Besides these main assumptions, outliers and influential observations can affect a model's estimates. This dataset does not seem to contain any troubling outliers. Now that the assumptions are verified, we will discuss how well the model fulfills our goal to understand how happiness affects job performance.

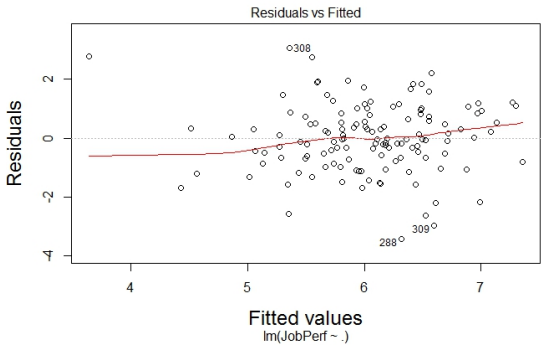

Figure 6: Fitted Values vs. Residuals

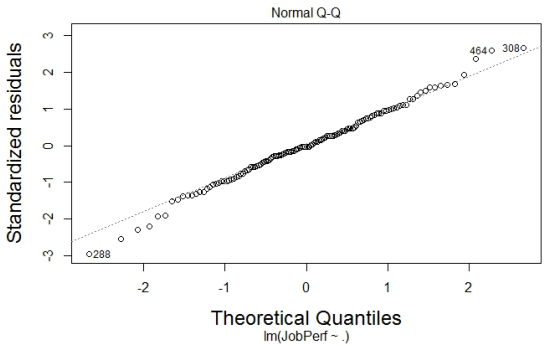

Figure 7: Q-Q Plot of Residuals

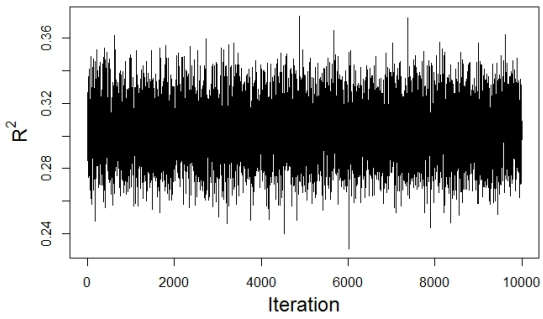

Figure 8: Trace Plot of R^2

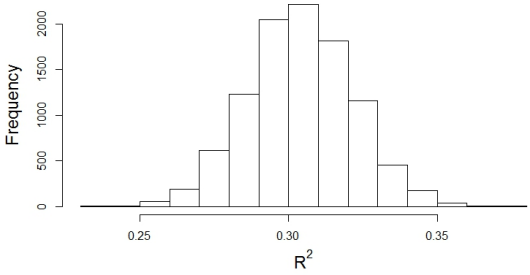

Figure 9: Histogram of R^2

3.2 Model Performance

In this scenario, a company would be most interested in what variables positively affect an employee's job performance. This purpose focuses on inference, or understanding the what drives job performance, and how well the model explains the variation in job performance. The significance of each β estimate will tell whether a variable has a significant effect on job performance (see Table 2). From this table we see that age, wellbeing, job satisfaction, and IQ have a significant effect on job performance, though not all positively. These estimates will be discussed and interpreted further at a later part of the report.

The second, and most important overall indicator of the model's ability to represent job performance is the R^2 value. The R^2 value was also calculated for each iteration in the multiple imputation process. As seen in Figure 8, the value of R^2 converges to a finite value as well. The histogram of R^2 in Figure 9 shows the range of values of R^2 that were produced through the iterative process. The mean R^2 value for all 10,000 iterations is 0.304. This value indicates that 30.4% of the variation in job performance can be explained by the included covariates. This R^2 is fairly low, suggesting that there are other variables not included in the model that carry the majority of the explanation of the variability in job performance. While the model has some power to explain job performance, it may be too little information to base a company culture shift on, or make other meaningful decisions to help improve job performance.

4 Results

The original goal of the study was to understand if employee happiness improves job performance. Job satisfaction and well being are primarily used to describe happiness, though happiness is hard to define clearly. For this reason each available variable was included to help explain what factors into job performance. The multiple linear regression model, using the original and imputed data, provides us with estimates of each variable's effect on job performance. Each variable's model coefficient (Equation 1) and estimated effect is shown in Table 2, along with a 95% confidence interval, and the p-value associated with the estimate. The confidence intervals are based on the pooled standard error from the multiple imputation method used to fill the missing data, and ultimately get our estimates from. Pooled standard error accounts for the standard error found

Table 2: Estimates with 95% Confidence Intervals

Coefficient	Estimate	Lower 2.5%	Upper 97.5%	p-value (>.05)
$\beta_{Intercept}$	-0.07	-1.51	1.36	0.92
β_{Age}	-0.03	-0.05	-0.01	0.01
β_{Tenure}	0.02	-0.01	0.06	0.21
$\beta_{Wellbeing}$	0.43	0.33	0.53	0.00
$\beta_{JobSatisfaction}$	-0.12	-0.23	-0.01	0.03
β_{IQ}	0.05	0.04	0.06	0.00

within each individual model (of each of the 10,000 calculated models), and from the standard error of each individual estimate against the aggregate average we use for the final estimate of each variable.

As seen in Table 2, age, wellbeing, job satisfaction and IQ are significant. The other variables have intervals that cross zero, indicating that they could have no effect on job performance. Those variables that are not found significant may not come as a surprise, seeing as the model only accounts for 30.4% of the variability in job performance. The significant, positive estimate of wellbeing indicates that at least one obvious measure of happiness has an effect on an employee's efficiency in their job. Increased age and intelligence (not surprisingly) also have a positive effect on an employee's job performance. One effect that may be surprising is the negative effect that job satisfaction has on job performance. This estimate alludes to a scenario where the more satisfied an employee is with their job, the less effective they are at work.

To bring context to the model I will interpret the estimated effect for wellbeing. If all other variables are held constant, and the employee's rating for wellbeing increases by 1, we expect their job performance rating to increase by 0.43, with 95% confidence that the true estimate of wellbeing's effect on job performance is somewhere in between 0.33 and 0.53. The coefficient estimates for age, IQ, and job satisfaction can be interpreted in a similar manner. Since tenure doesn't have a significant relationship with job performance, it does not have a useful interpretation for the desired goals of the study. From this model a company can design a corporate structure to foster the wellbeing of their employees, since this has the largest effect in this study. Though, as mentioned previously, there are other variables that were not included in this study that will further explain job performance.

5 Conclusion

Reviewing the goals of the study, we can evaluate the efficacy of this model. The interest in this model was to understand if happiness and other employee characteristics affect job performance. From this analysis a company would hope to make more informed decisions in their hiring, company culture, and corporate structure. These goals were met by creating a multiple linear regression model to understand the effects of these variables on job performance. The model was created using imputed data, as there were many observations that had missing values. In doing so we account for the variability in the missing data points, and used all available observations in the analysis. The pooled estimates provided from running the model at each step of the multiple imputation process give estimates for each variable that can be used to see their relationship with job performance. Though the model does not explain a large part of the variability of job performance, it shows that wellbeing and IQ do have a significant positive effect on job performance. This answers the goal of understanding which variables affect job performance. Companies can use such results to make desired changes in their company structure.

While the model accomplishes some of the goals, it does so in a lackluster manner. Only 30.4% of the data were explained by the variables included in the model, indicating that there are other variables that should be brought into the study to improve the model's ability to explain job performance. Characteristics such as marital status, education level, income, health indicators, and other variables could factor into the happiness of an employee, and possibly tell more of the story of what goes into job performance. Another shortcoming of the data is the source from which the data was taken. Universities are unique institutions, generally staffed with intellectual individuals with different work than than the majority of workplaces. An improvement could be made in how, and from where the observations are sampled. In doing so, companies can put more stock in the results of the model, instead of feeling that the only applicability of the model is for other universities, and their employees. In making such changes, this model can become a powerful tool in assisting a company's efforts to create the optimal work environment for success.