Employee Performance Analysis

Jackson Curtis

February 3, 2018

Abstract

This project analyzes the effect of well-being and job satisfaction on job performance. The data provided has many missing values, so imputation using the multivariate normal distribution is performed. Uncertainty from the imputation is included in the estimates.

1 Introduction

Employee happiness is increasingly recognized as a key contributor to productivity in the workplace. Modern theories insist that job performance (and thus profit) is directly affected by the well-being of employees, so investing in employee happiness and job satisfaction is of primary importance. A university would like to test this theory with data collected from their employees. Specifically, their goal is to understand the relationship between an employee's job satisfaction and well-being and their job performance. To do this, they quantified these three variables onto a ten-point scale, as well as collected data on three potentially confounding variables: age, tenure, and IQ.

In order to answer the question I will perform multiple linear regression and then perform inference on the coefficients to estimate the relationship between our variables of interest and the response. From that, we will be able to quantify the effects of job satisfaction and well-being on job performance. Prior to doing the regression, I will first account for the large amount of missing data within the dataset in a way that preserves the original relationships of the variables.

2 Data

The biggest concern with our data is that 72.7% of the 480 rows in our dataset are missing at least one data point. Table 1 summarizes how frequently each variable is missing. Discarding this data would significantly weaken the conclusions we can draw about our population, especially because we weren't told why the data is missing, so there might be systemic reasons that the data is missing. Figure 2 suggests this by demonstrating that complete rows have very few low job satisfaction responses, so people with low satisfaction may be hesitant to report it. Instead we would like to take advantage of the correlations between the variables we do have in order to impute the values we are missing in a way that preserves the unbiasedness of our estimates and their standard errors.

Missing Variables			Rows	
Well-Being	Job Satisfaction	Performance	Missing 1 variable	Missing 2 variables
160	160	64	314	35

Table 1: Summary of where the data is missing

One method that will allow us to do this is multiple imputation using the multivariate normal distribution. Instead of treating our explanatory variables as given, we will treat an entire row of our data as a draw from a multivariate normal distribution. This allows us to calculate the conditional distribution of missing data and impute a new value from which we can estimate our coefficients of interest. Creating many different datasets using this technique and aggregating them will give us unbiased estimates of our parameters.

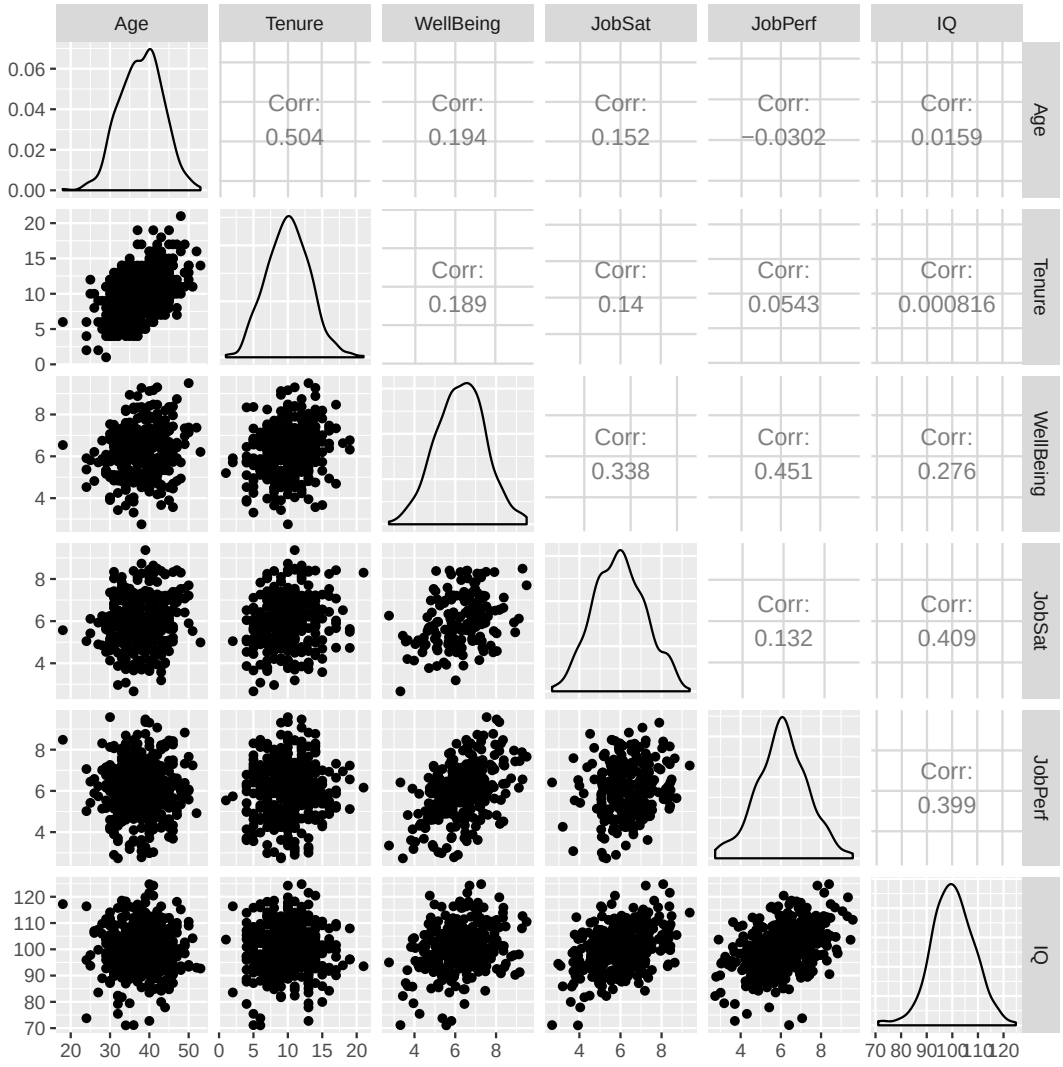

Figure 1: Histograms show reasonable bell curves and scatterplots show linear relations

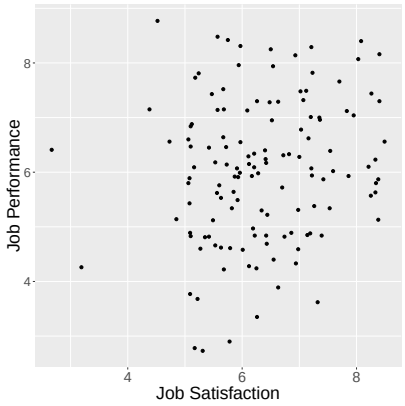

Figure 2: Graphing only complete rows shows few complete responses included low job satsifaction

A strong assumption of this model is that all the variables from a single subject can be modeled with the multivariate normal distribution. If this is not the case, another multivariate distribution would need to be chosen and its conditional distributions found. If the data is multivariate normal, all the marginals are also normal, and all the bivariate marginals should form a linear, oval shape. Figure 1 shows that this assumption appears to be met for all pairwise comparisons of our variables. The density of the job performance variable looks to be a bit too peaked to be normal, but for only just over 400 observations it is not an unreasonable shape for normally distributed data. It appears we can proceed with out multiple imputation algorithm.

Because we have few explanatory variables relative to the amount of data, we will use all six variables in our model without addressing their individual contributions to model fit.

3 Model

Our model for the imputation will be:

$$\begin{pmatrix} y_i \\ \boldsymbol{x_i} \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where the left-hand side is a row of data including the response and explanatory variables, $\boldsymbol{\mu}$ is a vector of means for each variable, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix with each variable's variance on the diagonal and their covariances with other variables on the off-diagonal.

Regardless of whether we're imputing 1, 2, or more variables, our conditional distribution can be calculated from the joint and will follow a normal or multivariate normal distribution. For each row in our dataset we will impute any missing values by randomly sampling from the known conditional distribution, and once we have a complete dataset we will estimate our coefficients the same as any linear regression problem. To deal with the bias introduced by our specific random draws that we take, we will do this many times to average over all possible samples each time estimating our conditionals from the estimate of the mean and covariance structure of our previous dataset. We can then aggregate over each individual estimate to get accurate, unbiased estimates for our coefficients, as well as perform hypothesis tests and create confidence intervals for them.

A sample imputation is shown in Figure 3. This shows that the points we're imputing seem to fall very naturally with the points that were given to us, and give us high confidence that our imputation is working how we would like it.

Running this algorithm, we can see how well the data is fit by the model. Each iteration will produce a slightly different r^2 term, but Figure 4 shows that they are in the 0.22 to 0.34 range. The mean of the r^2's is 0.29, meaning that the six variables we use account for about 30% of the variation in job performance.

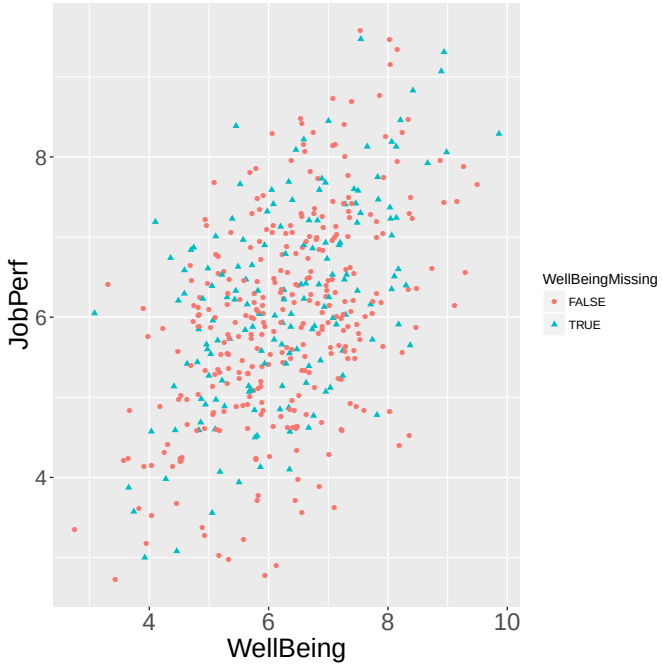

Figure 3: Imputed values graphed against observed values

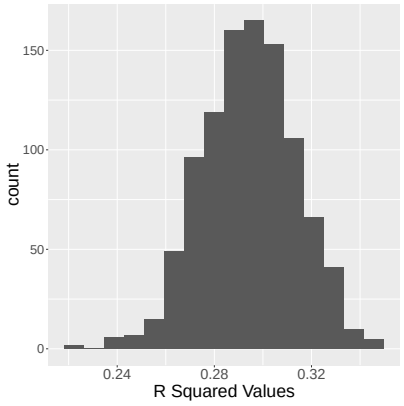

Figure 4: Distribution of r^2

	Estimate	P-value	95% Confidence Interval
Well-Being	0.41	< 0.0001	(0.308, 0.516)
Job Satisfaction	-0.081	0.171	(-0.200, 0.035)

Table 2: Coefficients of interest estimates and confidence intervals

4 Results

Table 2 gives the point estimates, hypothesis tests, and confidence intervals for our parameters of interest. From the p-value for well-being, we can definitively conclude that higher well-being relates to higher job performance. Both measured on a ten point scale, our estimate is that for a one point increase in well-being, job performance will increase by 0.41 points on average (95% confidence interval (0.308, 0.516)).

Our results failed to produce any evidence that higher job satisfaction resulted in higher job performance. Our point estimate resulted in a negative estimate of the effect, but our confidence interval of (-0.2, 0.035) suggested that the effect could be positive, negative, or zero. With a p-value of 0.171 we cannot reject the hypothesis that job satisfaction and performance have no linear relationship.

5 Conclusion

Using this analysis, we were able to establish a firm relationship between well-being and job performance, and we were unable to say anything conclusive about performance and job satisfaction. It is important to remember, however, that this survey is an observational survey, and it is particularly hard to determine if a person's well-being makes them perform better or whether they report higher well-being because they are a high performer. We cannot assign well-being experimentally, so caution should be exercised before making any causal inference.

The model we used was effective at handling the large amount missing data that we dealt with. Although our uncertainty would be less if we knew the exact values of the variables, the method we used did a good job of accounting for that increased uncertainty.

Further analysis may want to incorporate more variables (to account for the significant amount of variance that remains unexplained). If data was gathered that wasn't normally distributed (such as categorical variables), significant revisions would need to made on the imputation method in order to impute valid values for the explanatory variables.