Employee Case Study

Aubrey Odom & Jeremy Meyer

Brigham Young University

1 Background

1.1 Goals of Analysis

Traditionally, one of the metrics used to identify company productivity has been some assessment of worker contentment with their job duties. In order to gauge the happiness and performance of employees at a particular university, a survey of employees was administered that assigned quantitative scores to various measures. These included job satisfaction, individual well being at the university, tenure length, and objectively evaluated performance in their position. In addition, descriptive factors such as employee IQ and age were also recorded. By evaluating these features and their relationship to efficiency and achievement on the job, we hope to identify the largest contributing factors to the university's performance. This information can help universities be more effective with their employees. We will also address the validity of certain student hypotheses such as if older professors seem to care less about their jobs or if smarter professors have a harder time relating to students. In sum, the goals of the analysis are to address the following research questions:

1. Does employee well-being and/or job satisfaction impact job performance?
2. Does job performance tend to decrease with age (tenure)?
3. Will a higher IQ result in a lower job performance?
4. Does satisfaction and well-being lead professors to stay longer at the university?

1.2 The Data

The data consists of 480 employees, with data on the 6 different measures mentioned earlier. These can be seen from the histograms in Figure 1. Unfortunately, one issue with the dataset is the abundance of missing data. Details for this can be found in Table 1 below. In total, 72.7% of employees appeared to be missing at least one score of the three categories, and of these 10% of employees were missing two scores. This is problematic because creating linear models without addressing these *NA* entries would result in removal of the professors that have missing entries, and thus permitting usage of only 27.3% of our original dataset to estimate our model parameters. We would not only be losing information, but the analysis could be biased if there were important reasons the data is missing. We would like to fill in of our missing observations by using the complete data.

Table 1: Total Missing Observations

IQ	Age	Tenure	Well-Being	Job Sat	Job Perf	Employees missing $\geq$1 var
0	0	0	160	160	64	349
0.0%	0.0%	0.0%	33.3%	33.3%	13.3%	72.7%

2 Description of Models

2.1 The Multivariate Normal Distribution

In order to utilize all of our data in a linear model, we decided to fill in our missing observations through multiple imputation. We chose this method because it allows us to use all of the data without drastically

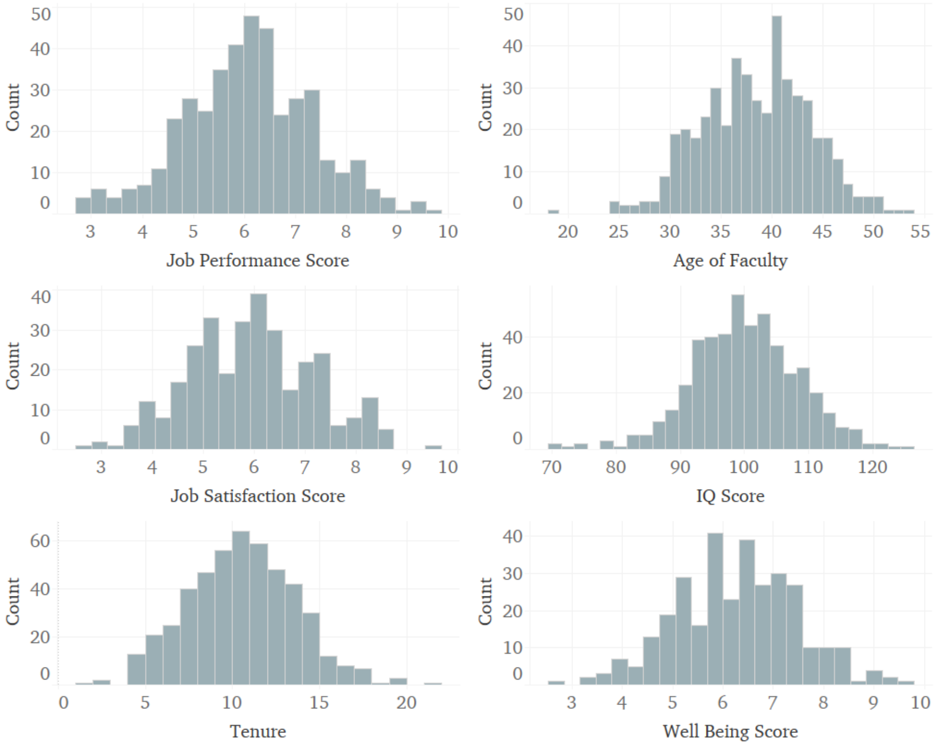

Fig. 1: These histograms represent the approximate distributions of each of the variables in our data set, excluding the missing observations.

altering estimates of the variances. To proceed with the imputation, we considered each employee as a draw from a 6 dimensional Multivariate Normal (MVN) distribution, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Each dimension in the MVN distribution corresponds to each numeric variable in the dataset. We can describe this relationship using the relationship presented in (1).

$$\text{Each Employee} \sim \text{MVN}_6(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

Initial estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were computed by using the data that was not missing. We then iterated through each employee with at least one missing value. Since every conditional and marginal distribution of the MVN by nature is also normally distributed, the missing values can be filled in by sampling from the correct conditional distribution. For example, if well-being is missing for a particular employee, the missing value can be sampled from the appropriate conditional distribution of all the other known variables. Samples can be drawn of one or multiple values, so this was able to successfully complete the data. However, we also want to make sure the parameters and missing data converge, so the estimates will be updated after completing the data, and are then used to complete the data again. We repeated this process for 10,000 iterations.

In order to fulfill the goals of the analysis, two different multiple regression models with job performance and tenure as the response will be fit during each iteration. The MVN distribution will allow us to switch variables of interest with ease. The assumptions for a multiple regression model will hold as long as the data is MVN since multiple regression comes from a conditional MVN distribution. We will save the parameter estimates and pool them so we can make variable inference. This will allow us to determine if a particular variable such as age or IQ impacts job performance.

2.2 Assumptions

One assumption that needs to be verified is if the data actually follows a Multivariate Normal distribution. Recall that one of the properties of a MVN is that every marginal, conditional, and joint distribution is normal. We will check assumptions on the complete data before using the MVN to fill in the missing values. The marginal distributions from Figure 1 look fairly normal in nature.

From the pairs plots in Figure 2, we can examine the bivariate relationships. Note that the data points in each plot resemble a cloud shape, with a higher concentration of points in the center. This suggests that all the 2-dimensional joint relationships have one peak and resemble a bivariate normal distribution. The points follow a fairly linear relationship, so the conditional distributions are linear, which is a property of the MVN. Since the data points do not funnel as we move along the horizontal axes, we can see that the conditional distributions have a constant variance. Due to the difficulty of checking higher dimensional conditionals/joint distributions, we will assume these checks are sufficient. As such, we will use the MVN distribution to impute the data.

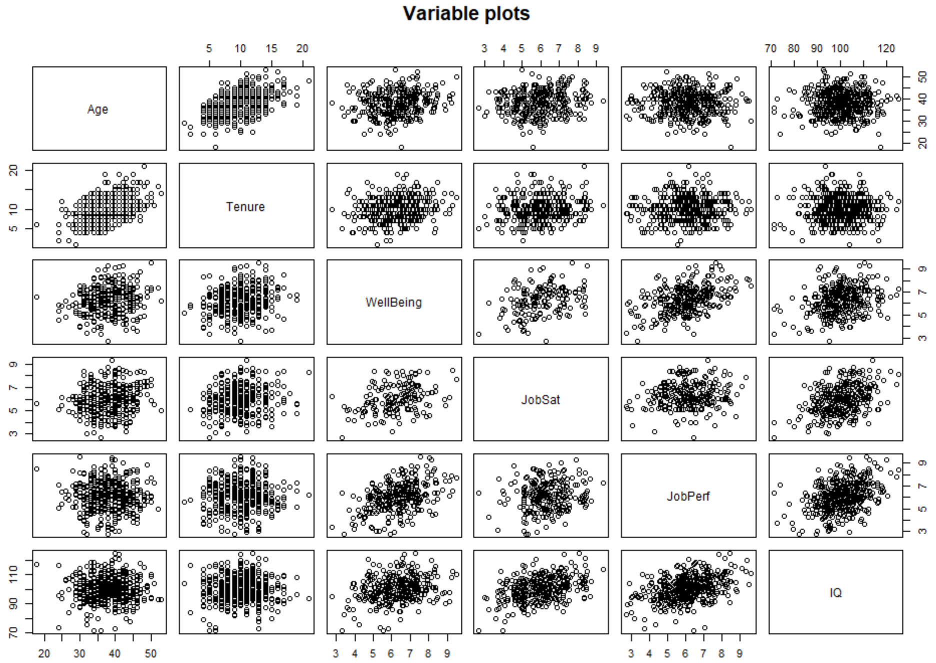

Fig. 2: These pairs-plots show all bivariate relationships. These are be analogous to a 2D kernel density plot.

After imputing the data for 10,000 iterations, we needed to make sure the sequence of parameters and missing data converge. The plots showing this convergence are seen in Figures 3 and 4 later in the report. These will be explained more in Section 3.1 following our introduction of the multiple regression models used for the research questions.

2.3 The Models

Our first model, described in equation (2), is intended to help us answer our first three research questions, using work performance as the variable that we would like to measure, and all other variables as predictors of performance. Since we are interested in all 6 variables for the research questions, we include them all. The βs in (2) are linear regression coefficients that act as weights on our predictor variables. β_0 is the estimate when all other variables are zero, and each of the subsequent β_is become the slope coefficients for the x_i inputs.

Although we do not know the true population values of these parameters, we can still make inferences on our limited data set. We were able to derive the $\hat{\beta}$s, or estimates of the βs, from the linear regression model. In order to obtain the pooled parameter estimates, we ran a linear regression on each of our 10,000 data sets and saved the $\hat{\beta}_i$s. Upon looking at a histogram of the estimates, we found their distribution to be Gaussian, allowing us to use the mean of the $\hat{\beta}_i$s as the pooled model parameter estimate.

$$\text{JobPerf}_i = \beta_0 + \text{Age}_i \cdot \beta_1 + \text{Tenure}_i \cdot \beta_2 + \text{Well-being}_i \cdot \beta_3 + \text{JobSat}_i \cdot \beta_4 + \text{IQ}_i \cdot \beta_5 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad (2)$$

With regards to answering our research questions, we are able to use the estimates in comparison to each other and the individual parameter confidence intervals to determine whether a certain predictor is significant to the model. For example, in the case that we would like to use job satisfaction and well-being to predict professor tenure length as in model (3), we can look at whether zero is in the confidence interval that describes those specific $\hat{\beta}$s. From this, we are able to determine whether the variables have a significant effect in predicting tenure, as long as it is unlikely that those $\hat{\beta}$s could equal zero. These ideas are further discussed in section 4.

$$\text{Tenure}_i = \beta_0 + \text{Age}_i \cdot \beta_1 + \text{JobPerf}_i \cdot \beta_2 + \text{WellBeing}_i \cdot \beta_3 + \text{JobSat}_i \cdot \beta_4 + \text{IQ}_i \cdot \beta_5 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad (3)$$

The example provided in the previous paragraph allows for the introduction of our second model, described in (3). The βs in this model do not refer to the βs in model (2), but their interpretation is the same. This model seeks to understand the relationship between our surveyed variables and the tenure length of professors. In this case, using multiple imputation to correct for missing values of the data was very appropriate because it allowed us to continue with using the data sets we already created, but now with a completely different model and response variable.

3 Justification & Performance

3.1 Convergence

To check the convergence of the data imputation, we have included trace plots of a few β and standard error estimates in Figures 3 and 4. These plots show the values of the coefficients and standard errors over each iteration and while they vary over time, their center and spread is the same. The same results can be seen for the missing data values themselves or the $\boldsymbol{\mu}$ estimates. This means that the values for the beta estimates do not depend on the data set iteration and so we can pool them together for the analysis.

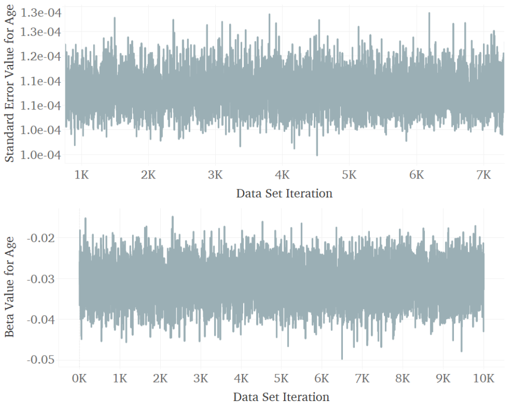

Fig. 3: This plot reveals the convergence of the $\hat{\beta}$ estimates and their standard errors over 10,000 iterations of the multiple imputation algorithm for the age covariate in (2). The age covariate is necessary for answering the question of whether age and tenure lead to decreased job performance.

After pooling the beta estimates together, we can measure how much the missing data affects the beta estimates by using a metric called the fraction of missing information (FMI). The two sources of variation in our estimates are variation within each iteration and variation between each iteration. FMI is defined as the percentage of variation in the estimates between each iteration relative to the total variation of the estimate. If the FMI is close to 1, then the imputed data has a high influence on the estimates. If the FMI is close to 0, that indicates that the way the data is filled has low influence on the analysis. The FMIs for each predictor in each of the models are listed in tables 2 and 3:

Table 2: FMI values for Job Performance Model

Intercept	Age	Tenure	WellBeing	JobSat	IQ
0.1596	0.1496	0.1482	0.3077	0.4495	0.2155

Table 3: FMI values for Tenure Model

(Intercept)	Age	JobPerf	Well-Being	JobSatis	IQ
0.0126	0.0247	0.1474	0.2929	0.2937	0.0673

Even though only well-being, job performance and job satisfaction have missing values, the $\hat{\beta}$ for the other variables still change depending on how the data is filled in. This is because the betas represent an expected change while holding all other variables constant. Thus the complete variables have some between iteration variability, though it tends to be very low, as can be seen in the FMI tables. It is worth noting that job satisfaction in the job performance model has a moderately high FMI: about 44.95% of the variation in $\hat{\beta}_4$ from model 2 is determined by how we impute our missing values. In all cases, less than half of the variability is due to data imputation, so we can safely make inferences on the betas.

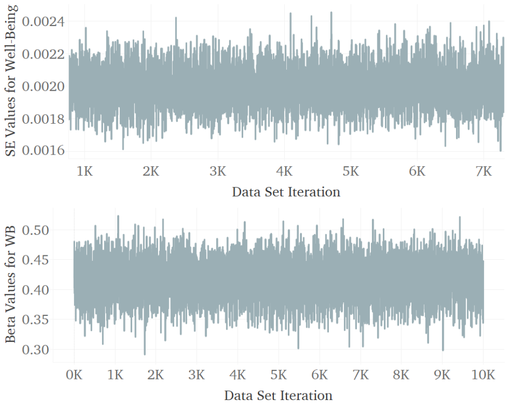

Fig. 4: This plot reveals the convergence of the $\hat{\beta}$ estimates and their standard errors over 10,000 iterations of the multiple imputation algorithm for the well-being covariate in (3). The well-being covariate is necessary for answering the question of whether well-being and satisfaction lead to increased professor tenure length.

3.2 Model Fit

Another question we wish to address is how well the two models fit the data. By fitting the models for each of the 10000 datasets, we can obtain R^2 values, which represent the percent of variation in the variable of interest that can be explained by the model. To pool the estimates together, we simply took the average across all iterations and found the middle 95% percentile values. The results can be shown in Table 4.

Table 4: R^2 values for each model

Model	R^2 estimate	2.5%	97.5%
Job Performance	0.3004	0.2628	0.3371
Tenure	0.2723	0.2637	0.2836

This means that over 10000 iterations, the Tenure model explained on average about 27.2% of the variation and about 95% of the time, the model explained between 26.37% and 28.46% of the variation.

It is worth noting that the Job Performance model R^2 has a higher spread (see Figure 5). This is a consequence of the Job Performance model having a higher FMI values. Although these percentages are low, since the sample of our data is large (480), these R^2 values indicate reasonable fit.

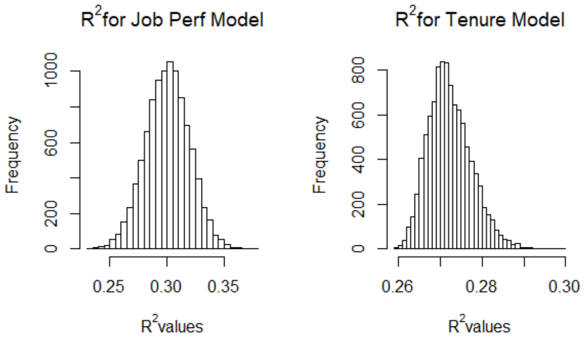

Fig. 5: R^2 values for all iterations. Note that the Job Performance model has a higher spread than the Tenure model.

4 Results

To answer the initial questions posed, we will be using the results presented in Tables 5 and 6. Our initial and overarching question is whether employee well-being and job satisfaction affect faculty performance at the university. While we cannot conclude that they are the leading cause due to the limitations of observational studies, we can address whether these variables do in fact have a real and recognizable prediction effect. In Table 5, we see that the estimate for well-being is 0.4169, with a confidence interval of (0.3117, 0.5222). This indicates that for every one point increase in well-being, we expect the job performance score to go up by 0.4169 points holding all other covariates constant. Furthermore, we are 95% confident that the true parameter effect estimate for well-being using this technique is between 0.3117 and 0.5222 units. This is a significant effect, as zero is not contained within the interval.

As for the job satisfaction score, note that the parameter estimate is -0.0552, with a estimated confidence interval of -0.1745 and 0.0641 units. Appropriate interpretations of these values are similar to those presented for faculty well-being scores. However, attention should be focused in this case to the span of the confidence interval for job performance, which does contain zero. This indicates that if we were to conduct a hypothesis t-test on whether the corresponding $\hat{\beta}$ is equal to zero, we would fail to reject the initial assumption, and conclude that job satisfaction is not an important predictor of faculty performance. Then to answer the question at hand, we resolve that faculty well-being does positively influence their work performance, while job satisfaction does not have a remarkable effect.

To answer our second research question of whether job performance decreases with age and tenure, we again look at Table 5. We find that our estimate of $\hat{\beta}_{tenure}$ is 0.0266, with a 95% confidence interval of (-0.0123, 0.0654). Given that zero is also within the interval span, there is not enough evidence to say tenure has a significant effect on job performance. With this information, we set our sights on faculty age as a viable predictor of job performance. We find that we have a negative parameter estimate (-0.0311), in line with our original hypothesis, and a 95% confidence interval of (-0.0536, -0.0085). In this interval, we have strictly negative parameter estimates which reveals that the age is has a significant effect on job performance, and that as professors gain age their job performance appears to simultaneously decrease at a rate of -0.0311 points per year. Thus, although age seems to have a negative impact on job performance, the total length of time at a university does not impact performance.

Our third research question is whether higher IQ's result in lower job performance. Again, due to the observational nature of this study we cannot really conclude causation, but we can again look at whether IQ is a significant predictor in the linear model. From Table 5, we obtain a $\hat{\beta}$ estimate of 0.0475, and a 95% confidence interval of (0.0330, 0.0619), which excludes zero and allows us to conclude that IQ has a positive, significant impact on job performance in contrast with our original inquiry-hypothesis.

Finally, our fourth research question uses the model in (3) to determine whether satisfaction and well-being influence professors to stay longer at the university. In Table 6, we find that the estimate

	Est	2.5%	97.5%	St. Error
(Intercept)	-0.1076	-1.5539	1.3386	0.7379
Age	-0.0311	-0.0536	-0.0085	0.0115
Tenure	0.0266	-0.0123	0.0654	0.0198
WellBeing	0.4169	0.3117	0.5222	0.0537
JobSat	-0.0552	-0.1745	0.0641	0.0609
IQ	0.0475	0.0330	0.0619	0.0074

Table 5: This is a table of $\hat{\beta}$ estimates, their confidence intervals, and pooled standard errors for the model presented in (2). It important to note that the standard errors and degrees of freedom for this analysis were pooled across all of the data sets gathered through multiple imputation.

for $\hat{\beta}_{JobSat}$ is 0.2622, contained within a 95% confidence interval of (0.2622, -0.0008) that does barely include zero. It also appears that well-being does impact professor tenure; the 95% confidence interval of (-0.1972, 0.3709) contains zero and so we must conclude that neither well-being or job satisfaction heavily influence professor tenure length.

	Est	2.5%	97.5%	St. Error
(Intercept)	-0.5196	-3.8641	2.8249	1.7064
Age	0.2815	0.2347	0.3282	0.0239
JobPerf	0.1670	-0.0769	0.4109	0.1245
WellBeing	0.0869	-0.1972	0.3709	0.1449
JobSat	0.2622	-0.0008	0.5251	0.1342
IQ	-0.0321	-0.0670	0.0027	0.0178

Table 6: This is a table of $\hat{\beta}$ estimates, their confidence intervals, and pooled standard errors for the model presented in 3. Parameter estimates were obtained using multiple imputation.

5 Conclusions

Reflecting on our analysis, we were able to tremendously ameliorate the effects of missing data through multiple imputation, and gained greater insight into the impact of several factors on our measured response variables of tenure and job performance. While the assumptions of multiple imputation were reasonably met for this particular application, it should be noted that multiple imputation is not always possible if the dataset at hand has categorical variables, non-linear bivariate relationships, or non-Gaussian univariate distributions.

From the whole of our analysis, we discovered that well-being affects performance while satisfaction does not, and that professor job performance tends to decrease with age while tenure has no effect. We also found that contrary to the original student hypothesis, a higher IQ actually has a positive impact on job performance. Finally that satisfaction and well-being have no real effect on job tenure length.

Looking forward, it seems reasonable to try and expand this study to many other universities so that we can conduct inference on whether these same effects hold true across academia. Also, we would like to point out that the observational nature of this study was somewhat limited in the scope of inferring causation, and that a future study designed as an experiment would be more useful in pulling out more concrete answers for our research questions.

6 Teamwork Statement

Aubrey worked on some figures, sections 1.1, 1.2, 2.3, 4, and 5. She also worked on the bulk of the computation output (multiple imputation algorithm w/ 10,000 iterations, pooling of linear model parameter estimates, pooling of standard errors, R^2 extraction.)

Jeremy worked making sure the model assumptions were met (2.2), outlined the methodology of filling in the missing data in the paper (2.1), evaluated model fit and convergence (3.1, 3.2) and edited several parts of section 1. We both had a really fun time!!