Job Performance Case Study

Spencer Newcomb & Spencer Ebert

Brigham Young University

Abstract. This paper details an effective method for imputing missing values when the data is modeled as coming from a Multivariate Normal (MVN) population distribution. We then pool across those imputations to find parameter estimates in a linear regression context. Then we use the MVN model to analyze data on university professors with the goal of answering various questions how about Job Performance and Tenure by conditioning for them. We conclude that the most important predictors of Job Performance are Age, Well-Being and IQ, whereas Tenure is significantly affected by Age and Job Satisfaction.

1 Introduction

1.1 Background

One of the primary goals of any business is to maintain the productivity of its employees. Many go to great lengths to assure employee satisfaction, believing that in turn they will be more effective workers. If a business can pinpoint which factors influence the satisfaction and well-being of their employees, they can then make informed decisions to maximize employee productivity.

1.2 Goals of Analysis

The data that we examined consisted of measurements on 480 university professors, including Age, Tenure, Job Performance, Job Satisfaction, overall Well-Being, and IQ Score. In this analysis we aimed to make inference on what affects the productivity and tenure of these professors. It is important for the university that the professors perform well both as instructors and researchers. In particular, we wanted to address the following questions:

1. Is job performance significantly affected by a professor's well-being or job satisfaction?
2. Do common student notions hold, such as older professors performing poorly or smarter professors relating poorly to students?
3. Do professors that are satisfied with their job tend to stay longer at the university?

In addition to these questions of interest, we had a primary goal of using all the data available to us. If there were any missing values, we wanted to find a reasonable way to not lose the information that actually was still there.

1.3 Exploring the Data

As we explored the features of the dataset, we quickly found that there was a severe issue with missing values. Of the 480 observations, 72.7% had at least one measurement missing. This meant that we only had 131 complete observations to work with. The following is a breakdown of missing values for each variable.

	ID	Age	Tenure	WellBeing	JobSat	JobPerf	IQ
% Missing	0.00	0.00	0.00	33.33	33.33	13.33	0.00

Table 1: The percentage of missing values for each variable

Though 131 observations may seem like a decent sample size to work with, it would be dangerous to ignore the sheer quantity of missing values in the data. If the measurements were omitted for some systematic reason, it would induce bias into our analysis. Our tests would also lose power if we just threw out those observations.

In order to get an idea for an appropriate statistical model, we looked at a scatter-plot matrix and found that the relationship between each pair of variables was approximately linear.

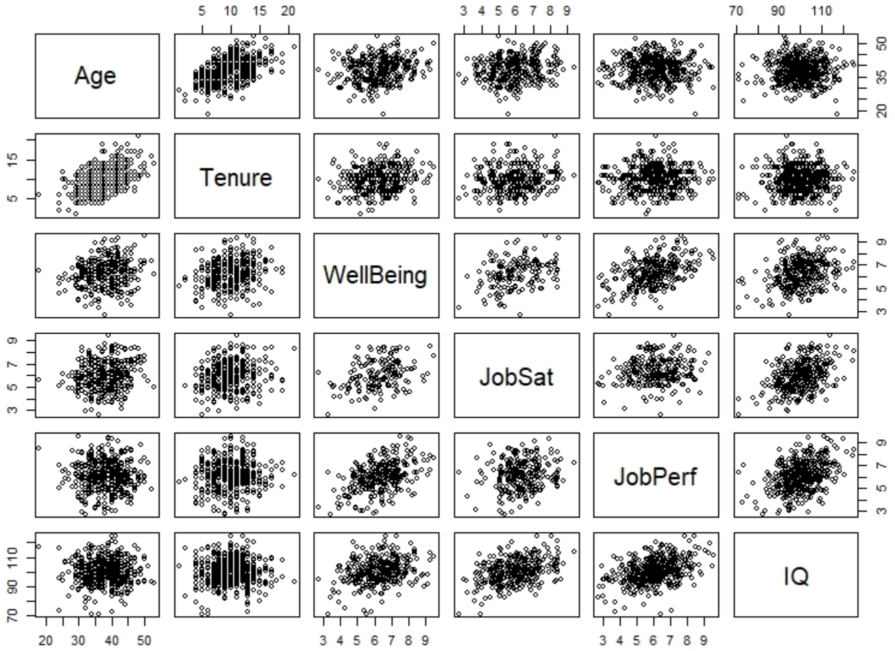

Fig. 1: Scatter-plots of each bivariate relationship.

2 Methodology

2.1 Multivariate Normal (MVN) Model

For this analysis, we fit a multivariate normal model to the data and filled in the missing values by using multiple imputation.

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{Where} \quad \boldsymbol{Y} = \begin{bmatrix} \text{Age} \\ \text{Tenure} \\ \text{Well Being} \\ \text{Job Satisfaction} \\ \text{Job Performance} \\ \text{IQ} \end{bmatrix} \tag{1}$$

In equation 1, $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{Y}$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{Y}$. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ aren't given in the context of this problem but we estimate them from the data given by using unbiased estimators $\hat{\mu} = \sum_{i=1}^{n} \boldsymbol{Y}_i$ and $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \hat{\mu})(\boldsymbol{Y}_i - \hat{\mu})'$. One of the really nice things about using a MVN model is that all of the conditional and marginal distributions are normal as well. By using this fact, we can fill in missing values from our dataset using conditional distributions. A discussion on multiple imputation is found in the next section. The MVN distribution is also very useful in the context of this problem because we can look at effects on any variable. We aren't just stuck with one response variable. The equations for conditional distributions from the MVN distribution are as follows.

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \end{bmatrix} \quad \text{Where } \boldsymbol{Y_1} \text{ and } \boldsymbol{Y_2} \text{ are partitions of } \boldsymbol{Y}. \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_1} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_2} \end{bmatrix} \tag{2}$$

$$\boldsymbol{Y_1} | \boldsymbol{Y_2} \sim N(\boldsymbol{\mu_{1|2}}, \boldsymbol{\Sigma_{1|2}}) \tag{3}$$

$$\boldsymbol{\mu_{1|2}} = \boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_2}^{-1} (\boldsymbol{Y}_2 - \boldsymbol{\mu_2}) \tag{4}$$

$$\boldsymbol{\Sigma_{1|2}} = \boldsymbol{\Sigma_1} + \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_2}^{-1} \boldsymbol{\Sigma_{21}} \tag{5}$$

In our analysis, we were interested in looking at Job Performance and Tenure as response variables. To do this, we looked at two different conditional normal distributions with $\boldsymbol{Y_1}$ = JobPerf, and $\boldsymbol{Y_1}$ = Tenure. Looking at both these variables as conditioned on the others, allows us to set up a multiple linear regression model for both responses.

$$\text{JobPerf}_i = \beta_0 + \beta_1 * Age_i + \beta_2 * Tenure_i + \beta_3 * WellBeing_i + \beta_4 * JobSat_i + \beta_5 * IQ_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, {\sigma_{jp}}^2) \tag{6}$$

$$\text{Tenure}_i = \theta_0 + \theta_1 * Age_i + \theta_2 * WellBeing_i + \theta_3 * JobSat_i + \theta_4 * JobPerf_i + \theta_5 * IQ_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, {\sigma_t}^2) \tag{7}$$

The coefficients β are the linear effects that each variable has on Job Performance. For example, if the Well-Being score goes up by 1 we expect the Job Performance score to increase by β_3 holding all else constant. The same applies for the θ coefficients, but instead with tenure as the response.

By using conditional models for both Job Performance and Tenure, we can investigate direct relationships that other variables have on the response from the estimated βs and θs. The main goals of our analysis are answered by estimating our regression coefficients and we estimate them in each step of our multiple imputation.

2.2 Handling Missing Data

The employee satisfaction dataset has myriad missing values. We didn't want to eliminate those lines because that gets rid of most of the dataset. We filled in the values by drawing from a conditional MVN distribution, according to the algorithm below.

1. Estimate $\hat{\mu} = \sum_{i=1}^{n} \boldsymbol{Y}_i$ and $\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \hat{\mu})(\boldsymbol{Y}_i - \hat{\mu})'$ from the complete data.
2. Identify an observation with missing values.
3. Find which variables are missing in that observation and look at a conditional distribution with $\boldsymbol{Y}_1$ as the missing values.
4. Draw from the conditional MVN distribution using equations 3-5 to fill in the missing values.
5. Repeat steps 2-4 for all of the observations with missing values.
6. Fit linear models conditioned for Job Performance and Tenure, and store away estimated coefficients and standard errors.
7. Estimate $\hat{\mu}$ and $\hat{\Sigma}$, but this time using the new filled-in dataset.
8. With these new estimates, repeat steps 2-7 replacing the previous MVN draws with the new ones a large number of times. (When identifying observations with missing values, use the observations from the original dataset not the new filled in data.)

We implemented MVN multiple imputation in this case because it used all of the data that we had, and drawing from a conditional MVN distribution is straightforward. Other, perhaps simpler, methods of imputation come with many shortcomings, such as biased variance estimates, artificially increased correlations, and underestimated standard errors, etc. We decided to use the multiple imputation algorithm to avoid these problems, and we account for uncertainty in the parameter estimates by drawing randomly from the conditional distribution.

To test to make sure multiple imputation worked we looked at trace plots for our estimated $\hat{\beta}$s shown in figures 2 and 3.

These trace plots show that the parameter estimates mix well, thus we felt comfortable with the multiple imputation step.

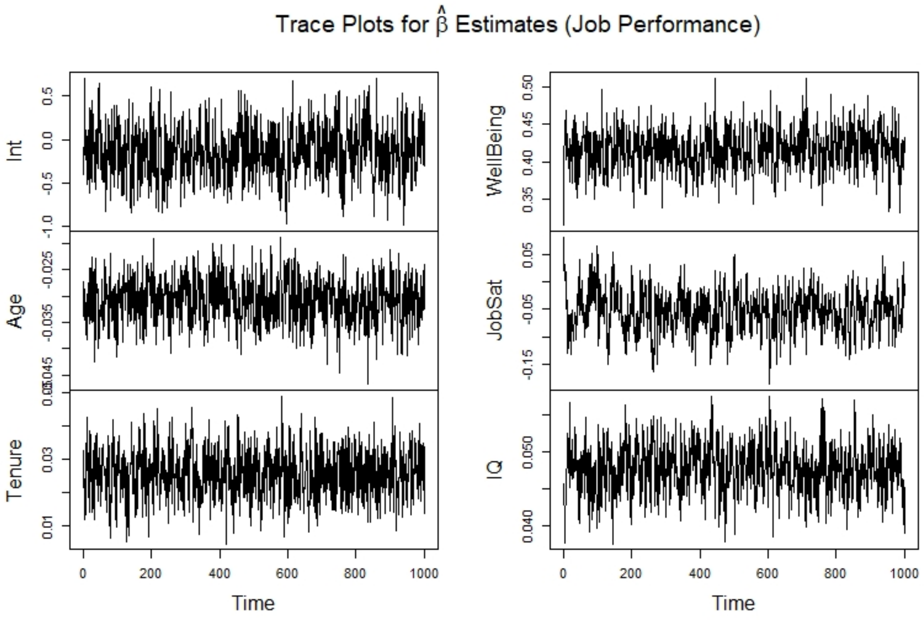
Trace Plots for $\hat{\beta}$ Estimates (Job Performance)
Int
-1.0
-0.5
0.0
0.5
Age
-0.045
-0.035
-0.025
Tenure
0.01
0.03
0.05
WellBeing
0.35
0.40
0.45
0.50
JobSat
-0.15
-0.05
0.05
IQ
0.040
0.050
0
200
400
600
800
1000
Time
Time

Fig. 2

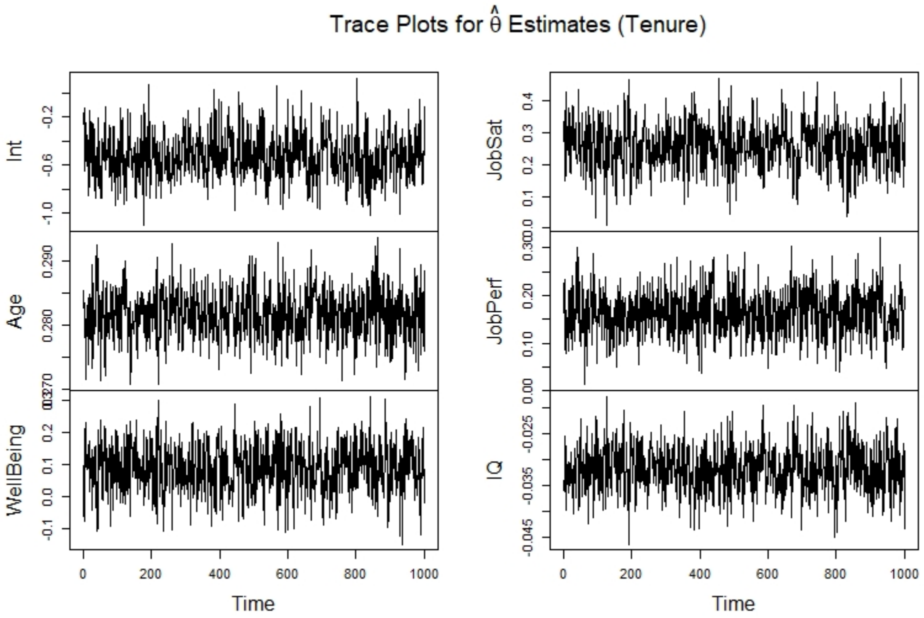
Trace Plots for $\hat{\theta}$ Estimates (Tenure)
Int
-1.0
-0.6
-0.2
Age
0.270
0.280
0.290
WellBeing
-0.1
0.0
0.1
0.2
0.3
JobSat
0.0
0.1
0.2
0.3
0.4
JobPerf
0.00
0.10
0.20
0.30
IQ
-0.045
-0.035
-0.025
0
200
400
600
800
1000
Time
Time

Fig. 3

3 Model Justification & Performance

3.1 Multivariate Normal Assumptions

When we jointly treat the variables as following a MVN distribution for Linear Regression, the model comes with the following assumptions:

1. The marginal distribution for each variable is normal
2. The relationship between any two variables is linear
3. Each observation is independent of the others

To address the issue of independent observations, it is most important to consider the nature of the data points and how they were gathered. There was no outstanding reason as to why the observations would depend on each other. Additionally, we assumed that the data were gathered through an appropriate random sampling mechanism.

When exploring the normality of the marginal distributions we simply looked at the sample histogram for each variable separately. As seen below in 4, each distribution is approximately normal in its shape.

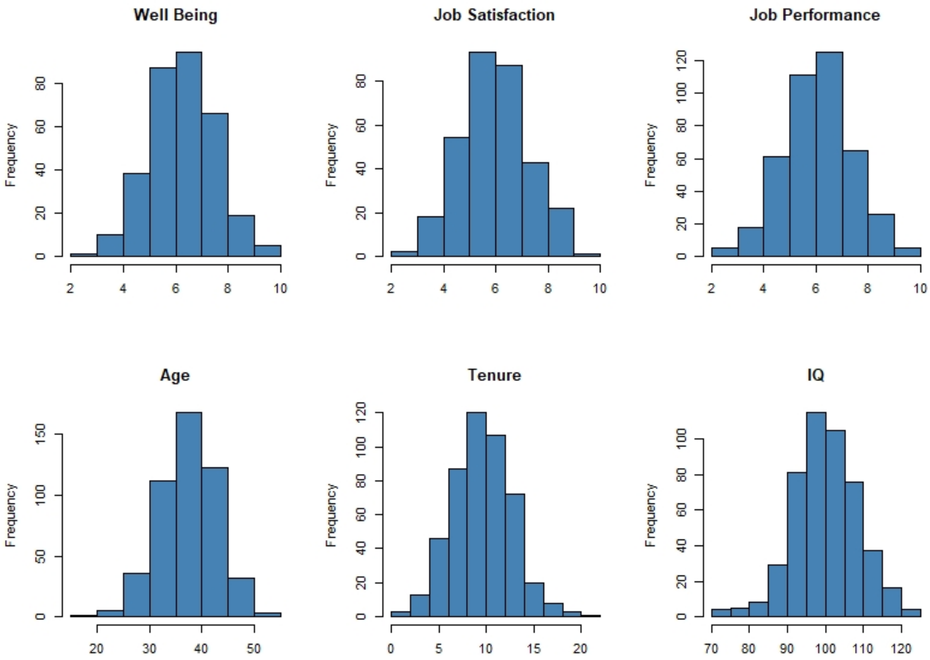

Fig. 4: Histograms for each marginal distribution.

To assess the linearity of the bivariate relationships in the data, we examined a scatter-plot matrix (see (1) and found the assumption to be met because each plot showed a linear relationship, resembling the contour of a bivariate normal distribution.

3.2 Model Fit

To evaluate the fit of the model, we computed R^2 for the conditional models of Job Performance and Tenure. We calculated our overall R^2 by pooling the R^2 for each iteration in our multiple imputation step. For the conditional for Job Performance, we got an R^2 of 0.30. This means that about 30% of the overall variance in Job Performance is explained by the other variables in the model. Based on the random nature of the imputations, the calculated R^2's differ so we looked at quantiles of 0.025 and 0.975 to assess our uncertainty of R^2 and got (0.26, 0.34). For the conditional model for Tenure we got an R^2 of 0.27 with 0.025 and 0.975 quantiles of (0.26, 0.28). That means about 27% of the overall variance in Tenure is explained by the other variables with uncertainty (0.26, 0.28). Given that this is a summary of several variables, our R^2 for both conditional cases is acceptable. The distribution of our R^2 in both cases is found in figure 5.

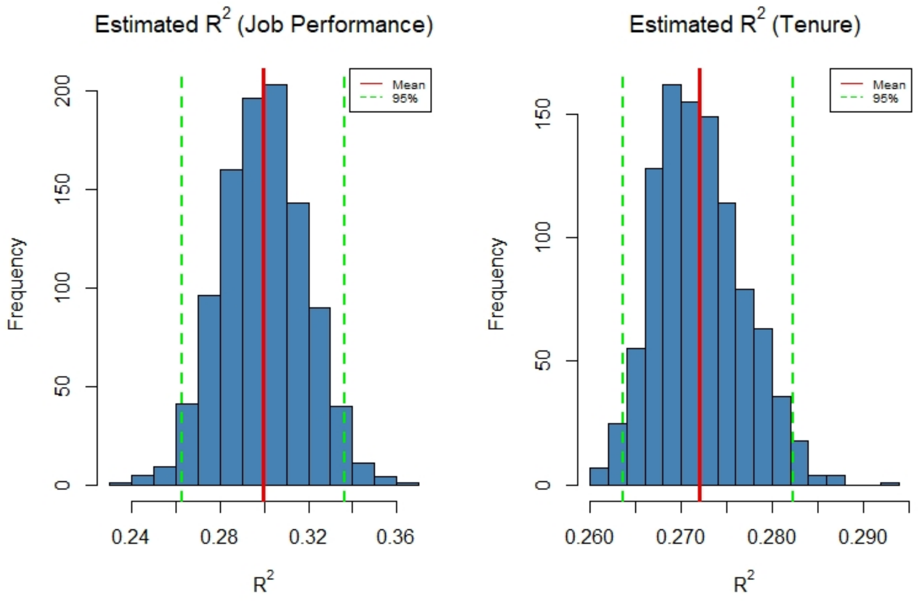

Fig. 5: Histograms for distribution of R^2

3.3 Missing Data Effect

We also wanted to see the effect our missing data had on the model. To do this, we calculated the fraction of missing information (FMI). M is the number of iterations in the multiple imputation step (we ran it 1000 times).

$$\text{FMI} = \frac{V_b + V_b/M}{V_t} \tag{8}$$

$$V_w = \frac{1}{M}\sum_{m=1}^{M} SE^2(\hat{\beta}_m) \qquad V_b = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_m - \text{Average of Betas})^2 \tag{9}$$

$$V_T = V_w + V_b + V_b/M \tag{10}$$

In basic terms, the FMI is the proportion of parameter variances ($\hat{\beta}_i$) that comes from the missing data. So if there is an FMI close to 1 then most of the parameter variance is a result of the missing data. On the other hand, if FMI is close to 0 then the missing data results in a small proportion of the variance. Our results for both conditionals (Job Performance and Tenure) are below.

	Int	Age	Tenure	WellBeing	JobSat	IQ
JobPerf	0.165	0.147	0.140	0.250	0.421	0.212

Table 2: FMI values when JobPerf is the response.

	Int	Age	WellBeing	JobSat	JobPerf	IQ
Tenure	0.012	0.024	0.291	0.301	0.140	0.065

Table 3: FMI values when Tenure is the response.

Most of the FMI's are low, especially for Tenure, Age, and Job Performance. We found higher FMI values for Well-Being and Job Satisfaction in both models, which makes sense because those variables had the most missing values compared to the others. Overall, our parameter FMI values show that most of the parameter variance was explained by the model rather than the missing data.

With these FMI values we calculated degrees of freedom, which affect our T statistics and confidence intervals in the results (see Section 4). The formula for degrees of freedom is below, where ν is the degrees of freedom.

$$\nu = (M-1)\left(\frac{1}{\text{FMI}^2}\right) \tag{11}$$

As FMI gets smaller our degrees of freedom gets larger, so our confidence intervals get smaller and we have greater power in the resulting tests.

4 Results

Our results can be divided into inference on what affects Job Performance and inference about Tenure. For both, we provide a table of estimates for the linear effect of each of the other variables, as well as the uncertainty associated with them (lower and upper bounds of a 95% confidence interval). Included in each table is also the p-value associated with a t-test of significance for each variable. We can use these results to answer and interpret our questions of interest. As discussed previously, these estimates are ”pooled” estimates based on 1000 iterations of filling in missing values.

4.1 Job Performance

	Estimate	Lower	Upper	T-stat	P-value
(Intercept)	-0.1347	-1.5838	1.3144	-0.1823	0.8554
Age	-0.0307	-0.0532	-0.0082	-2.6784	0.0074
Tenure	0.0260	-0.0127	0.0647	1.3151	0.1885
WellBeing	0.4164	0.3136	0.5192	7.9384	<0.0001
JobSat	-0.0533	-0.1702	0.0635	-0.8952	0.3707
IQ	0.0476	0.0332	0.0620	6.4767	<0.0001

Table 4: Parameter estimates associated with Job Performance

From Table 4, we see that three variables have a statistically-significant effect on Job Performance: Age, Well-Being, and IQ. Note that Job Satisfaction does not have a significant effect on Job Performance (p-value = 0.1885), but Well-Being does (p-value <0.0001). The estimated effect that Well-Being has on Job Performance can be interpreted as follows: for a one point increase in the Well-Being score, we expect to observe an average increase in Job Performance of between 0.3136 and 0.5192, holding all else constant. This is particularly interesting since Job Performance and Well-Being are on the same scale, ranging between 1-10.

Using a similar interpretation for Age, we found that as professors get older, they tend toward lower Job Performance. This supports one of the hypotheses that most students claim. The other student claim that professors perform poorly when they are ”smarter” is, however, not backed by the data. We estimate that for every one point increase in IQ Score, the average Job Performance for a given professor will actually increase by 0.0332 to 0.0620, with 95% confidence.

4.2 Tenure

Intuitively it makes sense that Age would be a significant predictor for Tenure. The more interesting result from this regression is that Job Performance has no significant effect on Tenure (p-value = 0.1882). Other than Age, the only other variable that significantly affects Tenure is actually Job Satisfaction (p-value = 0.0482). Based on our results, we estimate with that for a 1 point increase in the Job Satisfaction score, Tenure will increase by between 0.0020 and 0.5169 years on average (holding all else constant).

	Estimate	Lower	Upper	T-stat	P-value
(Intercept)	-0.5257	-3.8698	2.8183	-0.3081	0.7580
Age	0.2812	0.2344	0.3280	11.7823	<0.0001
WellBeing	0.0937	-0.1891	0.3765	0.6495	0.5160
JobSat	0.2594	0.0020	0.5169	1.9754	0.0482
JobPerf	0.1631	-0.0798	0.4060	1.3158	0.1882
IQ	-0.0320	-0.0668	0.0028	-1.8014	0.0716

Table 5: Parameter estimates associated with Tenure

5 Conclusions

In conclusion, the MVN approach to regression allowed us to effectively accomplish each goal that we had going into the analysis. Firstly, we were able to fill in missing values using the correct conditional distribution, given the complete data. After iteratively filling in those missing values with the algorithm specified in Section 2, the model lent itself to easily conditioning for a variable interest for linear regression to get estimates for the effects of the other ones. These regression models helped us answer the questions of interest as discussed in Section 4.

These results can help university administrators focus on what leads to a productive professor that stays with the university as long as possible. Age is not something that can be entirely controlled for, but we do see a marginal benefit to universities in searching out younger professors. The age issue can be countered by emphasizing overall well-being among the professors (perhaps instituting programs that increase well-being), since that has the largest positive effect on job performance. Interestingly, helping professors to find satisfaction in their job is not expected to increase productivity, but it does have a significant effect on how long the professor will stay. So we see a combination of well-being and job satisfaction being important for good professors to stay put.

During the analysis, many new and interesting questions arose that we think would be suitable for follow-up analysis. For example, with so many missing values, it would be helpful to fit a statistical model that could evaluate the risk of a measurement being not available. It would be interesting to explore through simulation other possible side-effects of multiple imputation. This would help identify specific shortcoming that our approach has that we may not be able to pinpoint right now. Additionally, since we have identified what affects Job Performance and Tenure at a high-level, a follow-up study with more in-depth assessments on employee satisfaction and job performance could lead to new insights.

6 Teamwork

Spencer’s Responsibilities:

1. Code up multiple imputation algorithm
2. Wrote majority of Sections 2 3
3. Master of formulas and all things math
4. Reviewed final product
5. Convince us to take a break to eat ice cream

Spencer’s Responsibilities:

1. Fit linear models and create tables of parameter estimates
2. Wrote majority of Sections 1, 4 and 5
3. Produced histograms and graphs
4. Reviewed final product
5. Convince us to take a break to eat dinner