NC STATE UNIVERSITY

ST841 Statistical Practice I

Consulting TAA

Author:	Contact Email:
Drew Hollis	anhollis@ncsu.edu
Xinyu Zhang	xzhang97@ncsu.edu
Qiang Heng	qheng@ncsu.edu

November 16, 2020

Contents

1 Introduction

Businesses must take advantage of the latest technologies to stay relevant and competitive. Many new technologies and tools can automate mundane business administration tasks. Automating administrative tasks leads to greater efficiency and fewer errors and free up employees to focus on tasks that add more value to the company. All of these principles hold for a company's tax and accounting department as well.

The advent of robotic process automation (RPA), the automation of business processes using artificial intelligence and software, has had a particularly significant impact on the accounting profession. [8] The extent to which companies and accounting departments have adopted and integrated automated tax tools into their workplace vary. Our client's study aims to determine what factors seem to have the most significant impact on the initiation, adoption, and routinization of tax analytics and automation (TAA) tools in larger companies.

Initiation, adoption, and routinization are explained in [2]. Initiation refers to an initial evaluation of the suitability of technology by the company, adoption refers to the point at which a company recognizes technology as valuable for its interests and begins to integrate that technology into its business practices, and routinization refers to the full-scale deployment of that technology.

In determining what factors most impact the initiation, adoption, and routinization of a technology by a company, our client employed the technology-organization-environment (TOE) framework that first proposed in [12] and later used by [14] to study the diffusion of e-business technology in companies.

The TOE framework indicates that technology diffusion is influenced by the technological sophistication and preparedness of a company, the organizational and managerial attributes of a company, and the regulatory and competitive context or environment of a company.

Our client is interested in determining which of these factors are related to the three response variables initiation, adoption, and routinization. Our client has already done some preliminary analysis exploring the relationship of the above explanatory variables to routinization [10].

1.1 Data

To answer these questions, our client sent surveys to a chief tax officer at each of the Fortune 1000 companies asking about the initiation, adoption and routinization of TAA within that company's tax department and the level of technological readiness, technological integration, managerial obstacles, competition intensity, and regulatory support for TAA for the company. Of the 1000 companies that received surveys, 68 responded.

Several of the variables like initiation, routinization, managerial obstacles, technology integration, competition intensity, and regulatory environment were measured using a number of 5-point Likert scale survey items. For instance to measure initiation, respondents were asked to use a Likert scale on 7 survey items to rate the significance that each of 7 potential TAA benefits had in their initial decision to pursue TAA tools.

Other variables like technological readiness where measured by asking a series of questions relating to the kinds of technology used at the company and the number of tax employees at the company who specialize in the use of TAA tools.

1.2 Survey Details and Preliminary Data Transformation

The data were collected using a survey with 28 questions. Each survey question is listed in the table in Figure 1 with a description of its format and the model variable it relates to.

Q6 asks the respondent to check all the levels (emerging, intermediate, and advanced) of TAA tools their company adopted and Q22 asks the respondent to check all specific TAA tools their company employs. Thus the responses to these questions are multiple checks corresponding to the technologies a company has employed. In order to turn these responses into a single numerical measure, we worked with the client to construct numerical weights corresponding to each check box. We then computed a weighted average across the check boxes for each observation to obtain a single numerical value summarizing the strength of each response for Q6 and Q22.

To obtain univariate measures for each of the Likert item questions, we performed nonlinear principal components analysis to a one-dimensional summary of each of the Likert scale questions. This procedure is described in detail in section 2.1.

2 Methodology

In this section, we provide more theoretical backgrounds of the methodologies that we adopt in this consulting project for the convenience of the client's future replication and publication. Aside from theoretical supports for the client's future references in this report, we would also provide a separate runnable RMarkdown file for technical supports. This section has been divide into four parts. Section 2.1 introduces how to combine the Multiple Likert Items into one score for the analysis using the non-linear Principal Component Analysis (PCA). Section 2.2 describes the methodology for imputing the missing values using Amelia. Section 2.3 states the main framework of the multivariate regression analysis, which was also adopted by the client before. Besides, Section 2.4 introduces the background of the regression tree methods, which were adopted to give more insights into the visualized tree.

2.1 Creating Composite Scores from Multiple Likert Items

One challenge presented by the TAA dataset is that several response variables of interest like TAA evaluation and routinization and several explanatory variables of interest like technology integration and managerial obstacles are measured using multiple Likert scale items.

When several survey items are used to measure a single underlying variable, it is often desirable to have some way of combining the information from the several survey items into a single composite score representing the variable of interest. A common approach for deriving such composite scores is to obtain the principal components of the several survey items and use these principal components, especially the first principal component as the composite score [5][9].

Our difficulty is that we are dealing with Likert items for which principal components analysis is not really well-defined. One solution to this problem is nonlinear principal components analysis with optimal scaling as described in [7] and [3]. This approach proceeds by finding an optimal transformation of the Likert data into a continuous representation and then performing standard PCA on this transformed representation.

Similar to standard principal components analysis, the nonlinear principal components analysis problem can be represented as a loss minimization problem. We minimize the loss function using a technique called alternating least squares.

Let $\mathbf{X}_{n\times p}$ be a matrix of n observations of p Likert scale items. We assume that x_{ij}, the i-th observation of the j-th item is 1 for the lowest rank item on the Likert scale, 2 if it is the second lowest rank item on the Likert scale, and so on. This represents the raw Likert scale data we would like to condense into a composite score. Let $\mathbf{Q}_{n\times p}$ represent the transformed version of $\mathbf{X}$ under optimal scaling. Let $\mathbf{H}_{n\times q}$ be the matrix of component scores where $q \leq p$. In our application, since we want a single composite score, we have $q = 1$. Finally, let $\mathbf{A}_{q\times p}$ be the matrix of component loadings. $\mathbf{HA}$ can be interpreted as a rank q approximation to $\mathbf{X}$. We wish to find the optimal approximation, so we minimize the loss function:

$$L(\mathbf{Q},\mathbf{H},\mathbf{A}) = n^{-1}\sum_{j=1}^{p}(\mathbf{q}_j - \mathbf{H}\mathbf{a}_j)^T(\mathbf{q}_j - \mathbf{H}\mathbf{a}_j) \tag{1}$$

where $\mathbf{q}_j$ and $\mathbf{a}_j$ are the j-th columns of $\mathbf{Q}$ and $\mathbf{A}$ respectively. Alternating least squares works by iteratively minimizing the loss function with respect to one of $\mathbf{Q}$, $\mathbf{H}$, or $\mathbf{A}$ while holding the others constant until some convergence criterion is met. Thus, the optimal scaling problem (finding $\mathbf{Q}$) is solved simultaneously with the principal components problem.

All that remains is to speak briefly about the form of $\mathbf{Q}$. In general for each of $j = 1, .., p$, we have

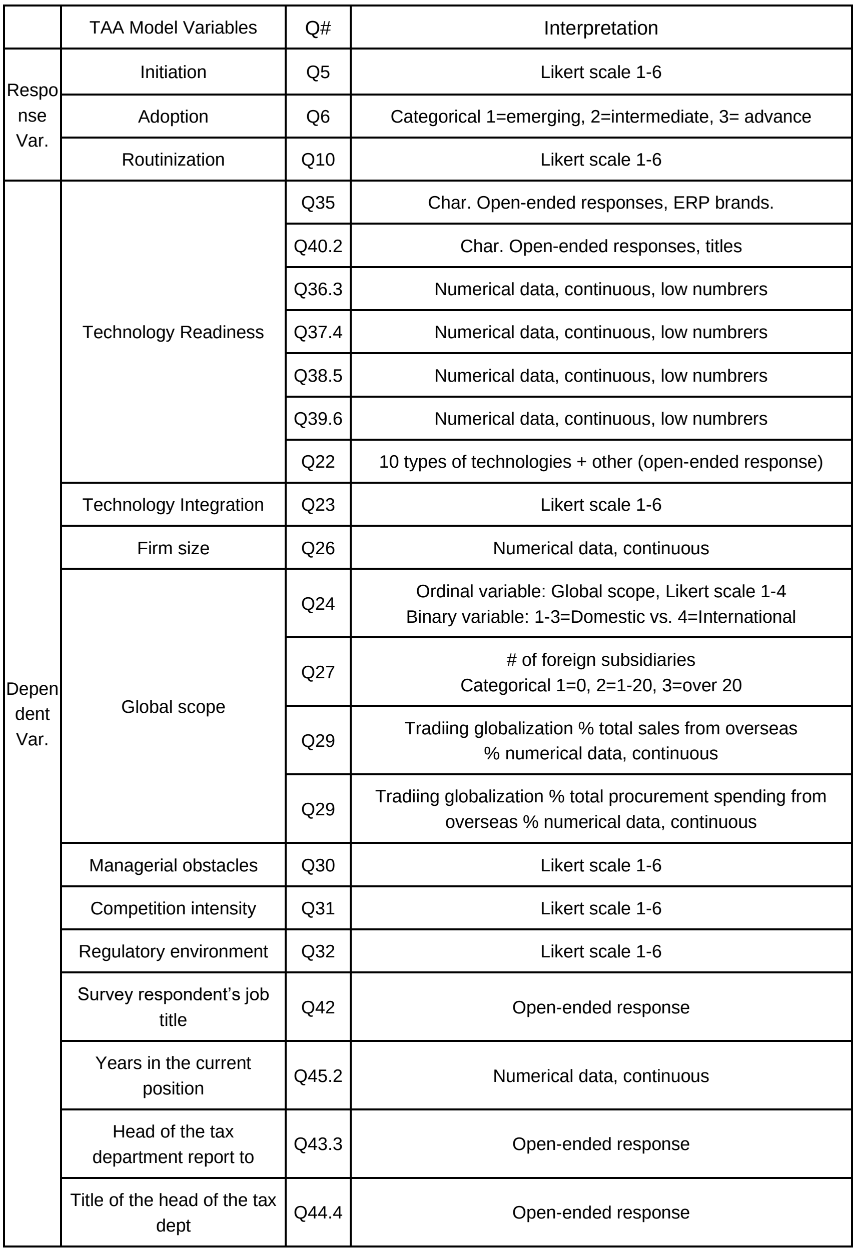
$$q_{ij} = \sum_{k=1}^{b} \alpha_k \phi_k(x_{ij}) \tag{2}$$

where q_{ij} is the i-th entry of $\mathbf{q}_j$ and $\phi_k(\cdot)$ is a basis function representation of x_{ij}. After some preliminary experimentation, a second degree B-spline was determined to do the best job of capturing the structure in the Likert items. If x_{ij} is a missing value, $\phi_k(\cdot)$ is set to $1/b$. This is one of three methods for handling missing values suggested in [3]. It is called the averaging method. After experimentation, it was found that the other two methods produced imputations for the missing values that were too extreme relative to the non-missing values. The optimization of the loss function with respect to $\mathbf{Q}$ involves finding for each $j = 1, ...p$, the set of $\alpha_1, ..., \alpha_k$ that minimizes the loss function.

2.2 Missing Value Imputation

Survey data As introduced before, there are 28 survey questions in total, and 23 of which are selected based on the references and the client's subject knowledge.

Missing value for the original survey data: We conducted an exploratory data analysis on this dataset with 68 observations and 56 original variables. We



	TAA Model Variables	Q#	Interpretation
Response Var.	Initiation	Q5	Likert scale 1-6
	Adoption	Q6	Categorical 1=emerging, 2=intermediate, 3= advance
	Routinization	Q10	Likert scale 1-6
Dependent Var.	Technology Readiness	Q35	Char. Open-ended responses, ERP brands.
		Q40.2	Char. Open-ended responses, titles
		Q36.3	Numerical data, continuous, low numbrers
		Q37.4	Numerical data, continuous, low numbrers
		Q38.5	Numerical data, continuous, low numbrers
		Q39.6	Numerical data, continuous, low numbrers
		Q22	10 types of technologies + other (open-ended response)
	Technology Integration	Q23	Likert scale 1-6
	Firm size	Q26	Numerical data, continuous
	Global scope	Q24	Ordinal variable: Global scope, Likert scale 1-4 Binary variable: 1-3=Domestic vs. 4=International
		Q27	# of foreign subsidiaries Categorical 1=0, 2=1-20, 3=over 20
		Q29	Tradiing globalization % total sales from overseas % numerical data, continuous
		Q29	Tradiing globalization % total procurement spending from overseas % numerical data, continuous
	Managerial obstacles	Q30	Likert scale 1-6
	Competition intensity	Q31	Likert scale 1-6
	Regulatory environment	Q32	Likert scale 1-6
	Survey respondent's job title	Q42	Open-ended response
	Years in the current position	Q45.2	Numerical data, continuous
	Head of the tax department report to	Q43.3	Open-ended response
	Title of the head of the tax dept	Q44.4	Open-ended response

Figure 1: The mapping of useful survey questions to the variables for analysis

checked the proportion of missing values and have observed that there are fifteen variables that have a proportion of missing values larger than fifty percent; ten variables have the $10\% - 50\%$ proportion of missing values; only the five categorical variables corresponding to Q5 have no missing value. For Observations in row 1, 2, and 5, we could also notice they have almost all variables missing.

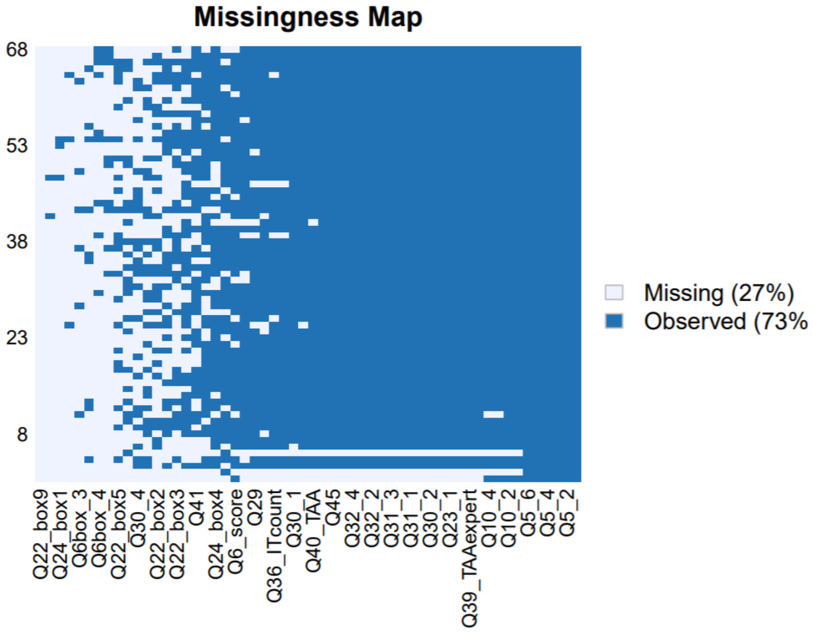
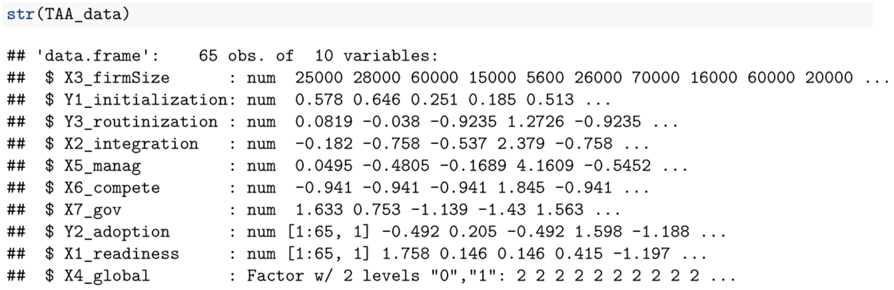


Figure 2: The visualization of missingness in the data. The Y-axis is the id of the observation. The X-axis is the variables ordered with a decreasing proportion of missing values from left to right.

Hence, after the discussion with the clients, we only keep 65 observations, and these 23 questions involve 56 variables mapping the ten variables, three response variables, and seven predictor variables as the table in Figure 1 shows.



```
str(TAA_data)
```

```
## 'data.frame':    65 obs. of  10 variables:
##  $ X3_firmSize      : num  25000 28000 60000 15000 5600 26000 70000 16000 60000 20000 ...
##  $ Y1_initialization: num  0.578 0.646 0.251 0.185 0.513 ...
##  $ Y3_routinization : num  0.0819 -0.038 -0.9235 1.2726 -0.9235 ...
##  $ X2_integration   : num  -0.182 -0.758 -0.537 2.379 -0.758 ...
##  $ X5_manag         : num  0.0495 -0.4805 -0.1689 4.1609 -0.5452 ...
##  $ X6_compete       : num  -0.941 -0.941 -0.941 1.845 -0.941 ...
##  $ X7_gov           : num  1.633 0.753 -1.139 -1.43 1.563 ...
##  $ Y2_adoption      : num [1:65, 1] -0.492 0.205 -0.492 1.598 -1.188 ...
##  $ X1_readiness     : num [1:65, 1] 1.758 0.146 0.146 0.415 -1.197 ...
##  $ X4_global        : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Figure 3: Cleaned data for analysis

AMELIA imputation method: As for the missing value imputation, we adopt the method named AMELIA. [6] Here the method assumes the first assumption that the complete data $\mathcal{D}_{n\times k} = (\mathcal{D}^{obs}, \mathcal{D}^{mis})$ follows the multivariate normal distribution, which is

$$\mathcal{D} \sim N_k(\mu, \Sigma) \tag{3}$$

and a second assumption that missing data are missing at random (MAR), which means the probability of missingness could be fully explained by observed data, defined as

$$p(M \mid \mathcal{D}) = p(M \mid \mathcal{D}^{obs}) \tag{4}$$

where M is the missing matrix with $M_{ij} = I(\mathcal{D}_{ij} \in \mathcal{D}^{mis})$.

Hence, the normal assumption as well as the MAR assumption should be checked after the data are fully prepared and before the conduction of imputation.

Remembering the normality assumption Eq. 3 and the missing at random assumption Eq. 4, we can obain the following likelihood for the observed data $\mathcal{D}^{\text{obs}}$:

$$p\left(\mathcal{D}^{\text{obs}}, M \mid \theta\right) = p\left(M \mid \mathcal{D}^{\text{obs}}\right) p\left(\mathcal{D}^{\text{obs}} \mid \theta\right) \tag{5}$$

where $\theta = (\mu, \Sigma)$ is the parameter of interest. Thus the likelihood of θ based on observed data can be expressed as:

$$L\left(\theta \mid \mathcal{D}^{\text{obs}}\right) \propto p\left(\mathcal{D}^{\text{obs}} \mid \theta\right) = \int p(\mathcal{D} \mid \theta) d\mathcal{D}^{\text{mis}} \tag{6}$$

and with a uniform prior of θ, the posterior could be

$$p\left(\theta \mid \mathcal{D}^{\text{obs}}\right) \propto \int p(\mathcal{D} \mid \theta) d\mathcal{D}^{\text{mis}} \tag{7}$$

Since the $\mathcal{D}^{\text{mis}}$ is missing, Expectation-Maximization (EM) algorithm combined with a bootstrap approach is applied here to obtain the estimated modes of the posterior with its variance estimation. By assuming a start point of θ_0, we define the Q function as

$$Q(\theta \mid \theta^v) = E_{\theta^v}(ln(p\left(\mathcal{D} \mid \theta\right)) \mid \mathcal{D}^{\text{obs}}) \tag{8}$$

where $p(\mathcal{D} \mid \theta)$ is the complete data likelihood. The updates of θ^{v+1} given θ^v as

$$\theta^{v+1} = \arg\max_{\theta} Q(\theta \mid \theta^v) \tag{9}$$

Thus, we have the estimations of θ and can make imputation of the missing data based on the observed data as well as the estimated parameters for the complete data likelihood.

Besides, m simultaneous imputations can be combined together to produce a more stable average imputaion:

$$\bar{T} = \frac{1}{m}\sum_{j=1}^{m} T_j \tag{10}$$

where j represents the j^{th} datasets, and the $\bar{T}$ has the following standard error:

$$SE(T) = \sqrt{\frac{1}{m}\sum_{j=1}^{m} SE\left(T_j\right)^2 + S_T^2(1 + 1/m)} \tag{11}$$

where, $S_T^2 = \Sigma_{j=1}^m \left(T_j - \bar{T}\right)^2 / (m-1)$ is the sample variance across the m estimations, and $SE\left(T_j\right)^2$ is the estimated variance of T_j based on the j^{th} dataset.

Besides, it's also common that to treat the missing value of a categorical variable as a new variable; for the questions related to Likert scales where there's always a level that indicating the customer doesn't know the answer, we can directly change the missing value into that case; some times a mean imputation would also be appropriate if sufficient subject knowledge is grounded. Thus, the missing value imputation of this project can be quite customized to the specific variable with more subject knowledge. However, cleaner data should be get in the first step.

2.3 Multi-response (Multivariate) Linear Regression

A common technique for handling multiple responses is multi-response (multivariate) linear regression. Suppose we have m responses with p features and n observations, multi-response linear regression can be written in matrix form

$$Y = XB + E \tag{12}$$

where $Y \in \mathbb{R}^{n\times m}$ are the dependent variables, $X \in \mathbb{R}^{n\times p}$ are the independent variables, $B \in \mathbb{R}^{p\times m}$ are the coefficients and $E \in \mathbb{R}^{n\times m}$ are the errors. To account for the possible correlation of different columns of Y, we assume that each row $E_{i.}$ of E has a covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. If put in vectorized form,

$$\mathrm{vec}(Y)|X \sim \mathcal{N}([B' \otimes I_n]\,\mathrm{vec}(X), \Sigma \otimes I_n) \tag{13}$$

It can be shown that the maximum likelihood estimate of B is invariant to Σ, which can be computed using ordinary least square

$$\hat{B} = (X'X)^{-1}X'Y \tag{14}$$

Then an unbiased estimate of error covariance matrix Σ is given by

$$\hat{\Sigma} = \frac{Y'(I_n - X(X'X)^{-1}X')Y}{n-p} \tag{15}$$

while the MLE for Σ is given by

$$\tilde{\Sigma} = \frac{Y'(I_n - X(X'X)^{-1}X')Y}{n} \tag{16}$$

The least square estimate $\hat{B}$ has the following distribution

$$\mathrm{vec}(\hat{B}) \sim \mathcal{N}(\mathrm{vec}(B), \Sigma \otimes (X'X)^{-1}) \tag{17}$$

Thus if we want to test $B_{jk} = 0$ against $B_{jk} \neq 0$, it can be seen that

$$\frac{\hat{B}_{jk}}{\hat{\Sigma}_{kk}(X'X)^{-1}_{jj}} \sim t_{n-p} \tag{18}$$

which will provide a p value for each regression coefficient B_{jk}. It is worth pointing out that mathematically, statistical inference for a single entry B_{jk} in the coefficient matrix B is the same for single-response linear regression and multi-response linear regression. However, compared with separate single-response linear regression models, multi-response linear regression takes possible correlation of response columns into consideration, which will enable inference for blocks of B. For example, we might be interested in testing whether several independent variables have a joint significant impact on all the response variables, i.e. we want to test whether a reduced model with $q < p$ covariates is sufficient. The hypotheses can be formulated as

$$H_0 : B_2 = 0 \text{ versus } H_1 : B_2 \neq 0 \tag{19}$$

where $B_2 \in \mathbb{R}^{(p-q)\times m}$ is a part of the partitioned $B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$. Then the likelihood ratio statistic becomes

$$\Lambda = \frac{\sup\limits_{B_1,\Sigma} L(B_1, \Sigma)}{\sup\limits_{B,\Sigma} L(B, \Sigma)} = (\frac{|\tilde{\Sigma}|}{|\tilde{\Sigma}_1|})^{n/2} \tag{20}$$

where $\tilde{\Sigma}_1$ and $\tilde{\Sigma}$ are respectively constrained and unconstrained MLEs for Σ. For large enough n, the modified likelihood ratio test statistic has an approximate chisquare distribution

$$-\nu \log(\Lambda) \sim \chi^2_{m(p-q)} \tag{21}$$

where $\nu = n - p - \frac{1}{2}(m - p + q + 2)$.
If we use E to denote $n\tilde{\Sigma}$ and H to denote $n(\tilde{\Sigma}_1 - \tilde{\Sigma})$, statistics from the Multivariate Analysis of Variance (MANOVA) literature [13] which rely on F-distribution approximations can be adopted for testing (19). Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ denote the eigenvalues of matrix HE^{-1},

- Wilk's lambda $= \sum_{i=1}^{p} \frac{1}{1+\lambda_i} = \frac{|E|}{|E+H|}$
- Pillai's trace $= \sum_{i=1}^{p} \frac{\lambda_i}{1+\lambda_i} = \mathrm{tr}(H(H+E)^{-1})$
- Hotelling-Lawley's trace $= \sum_{i=1}^{p} \lambda_i = \mathrm{tr}(HE^{-1})$
- Roy's largest root $= \max(\lambda_i)$

After the above test statistics are calculated, they are transformed into an F-statistic which will have an approximate or exact F-distribution under the null hypothesis. In some cases, the four tests are identical, which means they will lead to the same F-statistics and probabilities. When they do differ, Pillai's trace is often used because of its power and robustness [1]. We will take advantage of implementations of these tests in `anova` function of `stats` package in R [11] for our statistical computation.

2.4 Regression Trees

Due to the outliers, missing values, and lack of normality in the data, we believed it was worthwhile to also employ regression methodology that was robust to these problems.

Regression trees [4] are a modelling framework that make fewer assumptions than linear regression models, automatically handle missing values without using imputation or ignoring data, are robust to outliers, and do not require the data to be normally distributed.

A regression tree is constructed by assigning each observation to a bin based on the values of the independent variables associated with that observation. The tree model predicts the response of a particular observation to be the mean of the responses in that observation's assigned bin.

The bins, or leaf nodes, are found by recursively splitting the set of observations on values of the independent variables so that at each split the prediction error in the resulting bins is minimized.

We illustrate a regression tree in figure 4. The tree begins by splitting on the independent variable A. Observations with $A > 2$ go to the left, observations with $A \leq 2$ go to the right. On the left the next optimal split is performed on variable B. The boxes at the bottom of the tree represent the leaf nodes. P is the number of observations in that box and mean is the mean of the responses for those observations which is used as the prediction value.

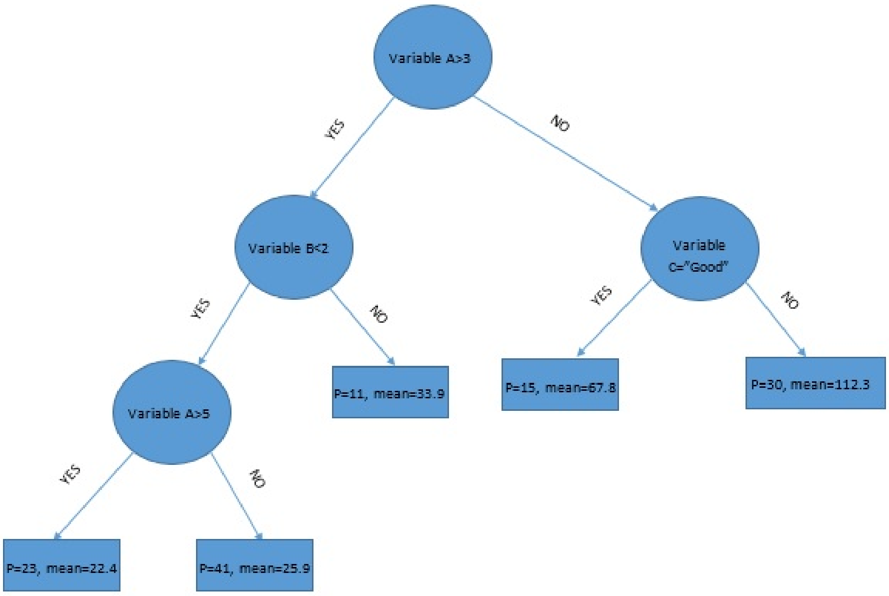


Figure 4: Regression Tree Illustrated

The splitting variables and the values for those variables that are used to define the split are chosen to optimize

$$\sum_{i \in B} \sum_{j=1}^{p_i} (y_{ij} - \bar{y}_i)^2$$

where B is the set of boxes, p_i is the number of observations in box i and $\bar{y}_i$ is the mean of the responses for the observations in box i.

3 Results

In this section we presents results of likert scale data transformation, mulvariate linear regression and regression trees. We first present graphs of likert scale item loadings to visualize how each item contributes to its composite score. Then we look at multivariate regression results to get an idea of how independent variables influence dependent variables. Taking possible correlation between response variables into consideration, MANOVA likelihood ratio tests are also conducted to see which independent variables have a significant impact on at least one of the response variables. Finally we create a regression tree for each response variable as a non-parametric alternative in our methodologies. The results of the regression trees are consistent with our findings in multivariate linear regression.

3.1 Transformation of Likert Scale Data

After using non-linear PCA to construct scores for the Likert data, it is useful to look at the loadings for each of the first principal component scores being used as a one-dimensional approximation. The loadings quantify the strength and direction of the relationship between each of the original Likert scale items and the first principal component. These loadings are depicted for each of the 6 Likert scale sets in Figure 5.

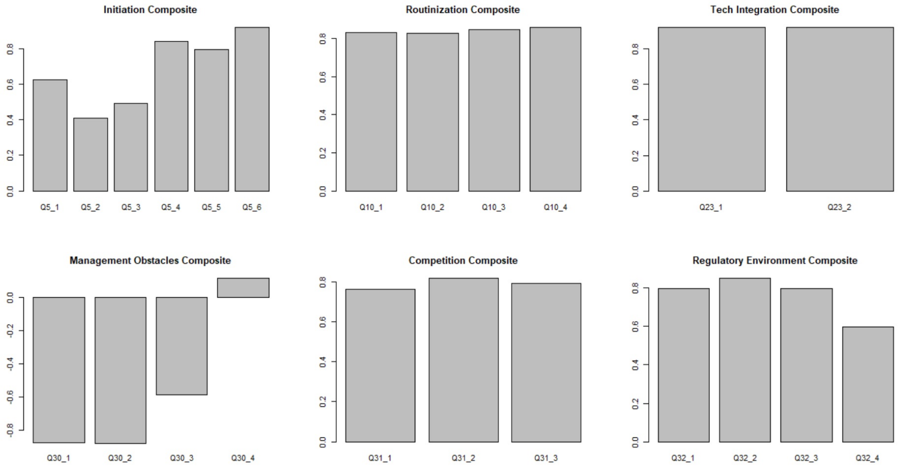


Figure 5: First Principal Component Loadings

We can use the loadings plot in Figure 5 to help us interpret the meaning of the scores that we use to represent each of the 6 variables underlying the Likert scale items. For instance, from the first plot, we can see that the score representing the consideration a firm initially gives to considering different TAA technologies is most influenced by a company's responses to the fourth, fifth, and sixth Likert items. The first, second, and third questions relate to using TAA to reduce costs, expand in-house tax services, and improve the coordination of the tax department with other departments. The last three items relate to improving the reporting, analysis, and productivity of the tax department. We can see that the last three items relate more to the primary output of the tax department and the first three relate to how the operations of the tax department within the company. This provides insight into what the unified score is really measuring. It may be a measure of the role accounting department output played in a company's initial considerations for adopting TAA.

Another interesting variable is the management obstacles variable. For this variable, the firms rarely answered the fourth Likert question. Thus, this question was mostly missing values. This explains its almost negligible impact on the first principal component score. It is also interesting to note that the other questions have a negative relationship with the first principal component. Thus, for the managerial obstacles variable, we would interpret a lower score as indicating more significant managerial obstacles.

3.2 Multi-response (Multivariate) Linear Regression Results

Recall that out response variables are TAA technology initiation, adoption and routinization, our independent variables are technology readiness, technology integration, firm size, global scope, management obstacles, competition intensity and regulatory environment. The explanatory variables are standardized to make the coefficients comparable and exclude intercepts from our model. The following table summarizes the coefficient estimates and their corresponding significance level.

	Initiation	Adoption	Routinization
Technology Readiness	-0.11	0.50***	0.12
Technology Integration	-0.06	-0.05	0.33**
Firm Size	0.06	0.26*	-0.03
Global Scope	0.37**	-0.18	0.07
Management obstacles	-0.11	0.19	0.15
Competition Intensity	0.07	-0.01	0.14
Regulatory Environment	0.06	0.05	0.15

Table 1: Coefficients and their siginifice levels. *** means $p < 0.001$, ** means $p < 0.01$ and $*$ means $p < 0.05$.

Notice that significant coefficients are rare. Now, let's utilize previously mentioned F-tests from MANOVA theory to decide whether a specific explanatory variable has a significant impact on all of the response variables. In other words, we can provide a single p value for each independent variable. The following table summarizes test statistics, the transformed F-statistics and their corresponding p values. Notice that in our case, the four tests, namely Wilk's lambda, Pillai's trace, Hotelling-Lawley's trace and Roy's largest root are equivalent and they produce the same F-statistics and p values, so we only report one F-statistic and p value for each independent variable. Also, Hotelling-Lawley's trace and Roy's largest root turn out to have identical test statistics, so we report them in one column.

	Wilk	Pillai	Hotelling/Roy	F-statistic	p-value
Technology Readiness	0.73	0.28	0.38	7.47	0.0003***
Technology Integration	0.86	0.14	0.16	3.10	0.03*
Firm Size	0.90	0.10	0.11	2.07	0.11
Global Scope	0.86	0.14	0.17	3.29	0.03*
Management Obstacles	0.91	0.09	0.10	1.96	0.13
Competition Intensity	0.98	0.02	0.03	0.50	0.68
Regulatory Environment	0.97	0.03	0.03	0.54	0.65

Table 2: MAONVA test statistics and their p values. *** means $p < 0.001$, ** means $p < 0.01$ and $*$ means $p < 0.05$.

From the table we can see that technology readiness, integration and global scope are statistically significant. Firm size is significant for adoption but not significant overall. We can also see that management obstacles, competition intensity and regulatory environment don't seem to have significant impact on the response variables.

3.3 Regression Trees

We fit a separate regression tree for each of the three response variables. The resulting trees are plotted in the below figures.

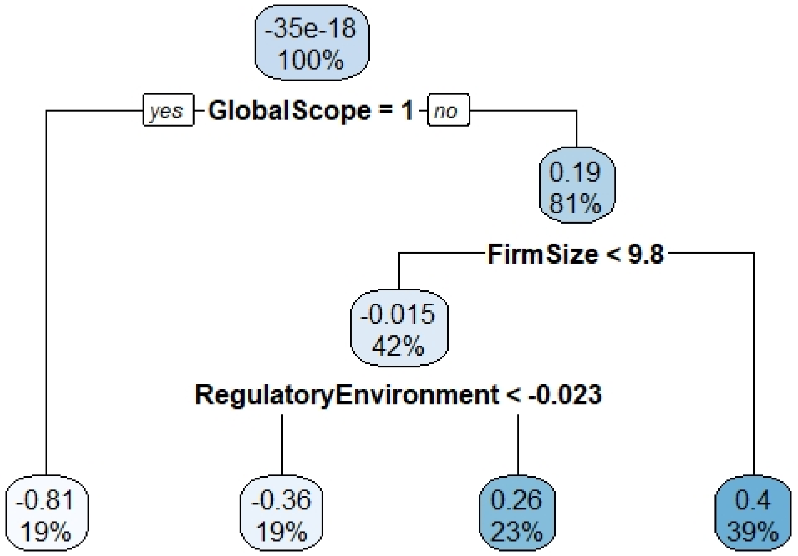


Figure 6: Initiation Tree

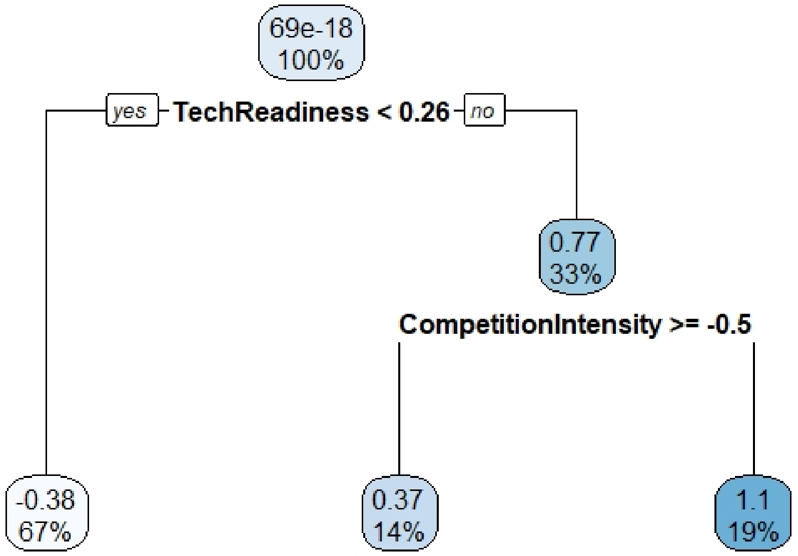


Figure 7: Adoption Tree

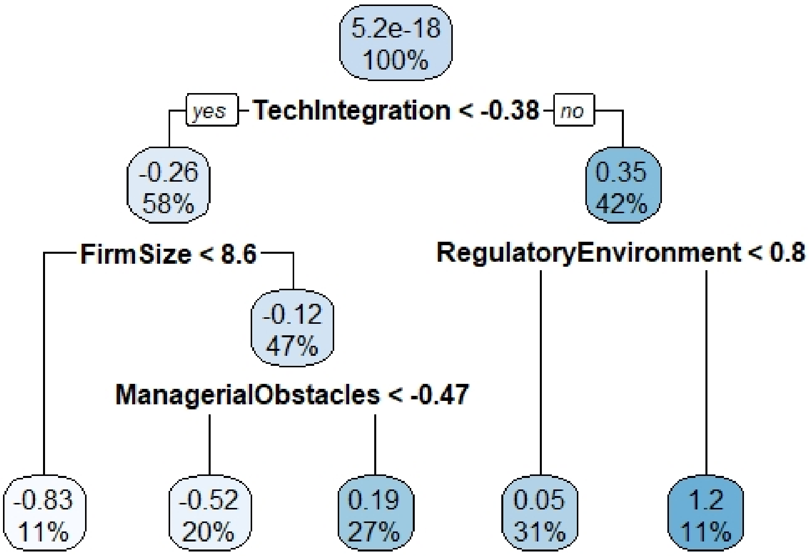


Figure 8: Routinization Tree

The first number in each box is the mean response for that partition of the tree and the percentage is the percentage of observations assigned to that partition of the tree. Thus, for the initiation tree, 19% of the observations were purely domestic and these domestic firms had a mean initiation score of -0.81. 42% had a foreign presence a log firm size less than 9.8 and a corresponding mean initiation score of -0.015, while 39% of firms had a foreign presence and log firm size of more than 9.8 and a corresponding mean initiation score of 0.4.

The first splitting variable is often interpreted as the most important or influential variable in the model. The results agree with the linear regression results in that for initiation the global extent of the firm as measured in global scope seems to be most associated with higher initial consideration of TAA, for adoption, technological readiness seems to be most closely associated with adopting more advanced TAA tools, and for routinization, technological integration seems to be most closely associated with better TAA routinization.

The variables lower down in the tree, however, may not be important, and may show up as being valuable only for partitioning this particular data set. The lower-level variables may not be generally important.

4 Discussion

In this project, we provide multi-aspect statistical supports to help the client gain better understandings of the diffusion of tax automation techniques and develop better methodology in presenting the analysis in the research paper. We transformed the Likert scale using non-linear PCA; imputed the missing values using AMELIA through EM algorithm; analyzed the results using both multivariate linear regression for significant levels and regression tree analysis for interpretable visualizations.

It is important to note that the data set is small, there are outliers, and much of the data is missing. The linear regression results should be interpreted with caution due to the outliers, the small data set size, and the fact that the data don't appear to be normal.

While regression trees are robust many of these weaknesses, single-fit regression trees tend to over-fit the data. This issue may be exacerbated by the small sample size. The splitting variables lower down in the trees may not be generally important, and may show up as being valuable only for partitioning this particular data set.

Both our analyses have consistently identified a number of important variables for each of the stages of technology diffusion, and these findings may serve as a good basis for future studies. Because of some of the issues with the dataset mentioned above, however, we encourage the reader to interpret the results with some caution.

Also, we would provide an RMarkdown file aside from this report to the client as a documentary of replicable results for future research.

References

[1] Gregory Carey. Multivariate analysis of variance (manova): I. theory. Retrieved May, 14:2011, 1998.

[2] Randolph B Cooper and Robert W Zmud. Information technology implementation research: a technological diffusion approach. Management science, 36(2):123–139, 1990.

[3] Jan De Leeuw. Multivariate analysis with optimal scaling. 2011.

[4] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the lasso and generalizations. CRC press, 2015.

[5] William G Henderson, Susan G Fisher, Noel Cohen, Susan Waltzman, Laura Weber, VA Cooperative Study Group on Cochlear, et al. Use of principal components analysis to develop a composite score as a primary outcome variable in a clinical trial. Controlled Clinical Trials, 11(3):199–214, 1990.

[6] James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. Journal of statistical software, 45(7):1–47, 2011.

[7] Mariëlle Linting, Jacqueline J Meulman, Patrick JF Groenen, and Anita J van der Koojj. Nonlinear principal components analysis: introduction and application. Psychological methods, 12(3):336, 2007.

[8] Steven Mezzio. Robotic process automation for tax: Tax is in the vanguard of the fourth industrial revolution. Journal of Accountancy, 228(6):18, 2019.

[9] Reza Motallebzadeh, J Martin Bland, Hugh S Markus, Juan Carlos Kaski, and Marjan Jahangiri. Neurocognitive function and cerebral emboli: randomized study of on-pump versus off-pump coronary artery bypass surgery. The Annals of thoracic surgery, 83(2):475–482, 2007.

[10] Robin Pennington, Al Chen, Jennifer Dirienzo, and Kathy Krawczyk. Technology diffusion in tax: An examination of tax analytics and automation routinization. AMCIS 2020 Proceedings. 5., 2020.

[11] R Core Team et al. R: A language and environment for statistical computing, 2013.

[12] L Tornatzky and Mitchell Fleischer. The process of technology innovation. Lexington, MA: Lexington Books, 165, 1990.

[13] Russell T Warne. A primer on multivariate analysis of variance (manova) for behavioral scientists. Practical Assessment, Research & Evaluation, 19, 2014.

[14] Kevin Zhu, Kenneth L Kraemer, and Sean Xu. The process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business. Management science, 52(10):1557–1576, 2006.