Analysis of Customer Satisfaction Survey Data

Pete Rotella*, Sunita Chulani†,
*Cisco Systems, Inc. – 7200 Kit Creek Road, Research Triangle Park, NC, USA 27709
†Cisco Systems, Inc. – 821 Alder Drive, Milpitas, CA, USA 95035
Emails: protella@cisco.com, schulani@cisco.com

***Abstract*—Cisco Systems, Inc., conducts a customer satisfaction survey (CSAT) each year to gauge customer sentiment regarding Cisco products, technical support, partner- and Cisco-provided technical services, order fulfillment, and a number of other aspects of the companys business. The results of the analysis of this data are used for several purposes, including ascertaining the viability of new products, determining if customer support objectives are being met, setting engineering in-process and customer experience yearly metrics goals, and assessing, indirectly, the success of engineering initiatives. Analyzing this data, which includes 110,000 yearly sets of survey responses that address over 100 product and services categories, is in many respects complicated. For example, skip logic is an integral part of the survey mechanics, and forming aggregate views of customer sentiment is statistically challenging in this data environment. In this paper, we describe several of the various analysis approaches currently used, pointing out some situations where a high level of precision is not easily achieved, and some situations in which it is possible to easily end up with erroneous results. The analysis and statistical territory covered in this paper is in parts well-known and straightforward, but other parts, which we address, are susceptible to large inaccuracies and errors. We address several of these difficulties and develop reasonable solutions for two known issues, high missing value levels and high colinearity of independent variables.**

***Keywords*-CSAT survey (customer satisfaction survey), customer satisfaction, data imputation, missing values, mean substitution, listwise deletion, dominance analysis, colinearity.**

I. Introduction

The Cisco yearly customer satisfaction survey (CSAT) results contribute key data to the customer perception layer of what we call the "quality pyramid." Our quality pyramid currently consists of three levels of quality-centric metrics: Internal (in-process) engineering, customer experience, and customer perception. Figure 1 depicts this software quality paradigm (and similar representations can be drawn for hardware quality, services, etc.).

Our intent with this quality pyramid is to identify "linkages" (also sometimes referred to as "drivers" or "predictors" in this paper) that connect the engineering layer to the customer experience layer, and then the customer experience and customer perception layers. Customer experience alone is not sufficient information—customer expect to encounter bugs and other issues, but how many, and of what severity, should be encountered over a given time period? We have no way of knowing if n number of bugs or issues are acceptable or not for a given customer. Guessing about this is not good enough, so we go directly to a large sample of customers yearly to receive direct feedback.

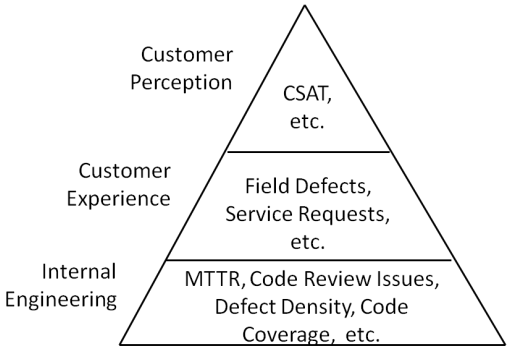

Figure 1. Quality pyramid

"Linkages" are found using multiple linear regression models, with the successful independent variables comprising the drivers that presumably influence the metric (dependent variable) at the next higher level in the pyramid. There are also several linkages needed within the customer perception layer itself, to connect specific attributes of CSAT to the CSAT experience layer—this additional step is described in Section II-C. (In addition, we are in the process of adding two additional layers above customer sentiment: Customer loyalty and, above that, revenue/profit.)

The objective of establishing accurate linkages is to identify specific engineering practices and processes (such as structured code reviews) that influence the customer experience with the software (experiencing field bugs, for example). We then assess the extent of the influence of customer experience events on customer sentiment, measured with the yearly CSAT survey (as well as with other mechanisms, such as customer feedback sessions and CIO surveys). Once we identify a sequence of linkages that suggests a way to improve customer satisfaction, we promulgate the use of the practice/process across the targeted business unit, technology group, or even across all of engineering. This method has resulted in the implementation of numerous practices and processes across all of engineering.

978-1-4673-1761-0/12/$31.00 © 2012 IEEE

MSR 2012, Zurich, Switzerland

To successfully accomplish this process/practice improvement approach, we need to be sure that the CSAT drivers, linking the CSAT attributes to the CSAT experience level, are accurate the entire chain of linkages depends on the reliability of this last step. In late 2010, we noticed that several of these linkages did not make much sense. We saw, for example, that product quality turned out to be the main driver of the ease of doing business with Cisco experience-level responses. Several of these "sanity check" issues motivated us to study the analysis method in more depth to ascertain if the method needed improvement. The CSAT survey has been in place for over 10 years, and the analysis method has evolved over that time, but perhaps not fast or thorough enough to ensure the level of accuracy we need at this time.

This paper describes in detail several areas in which we have made CSAT analysis improvements. There are other areas that are still problematic, but more study is needed before we have viable solutions. Some of these as-yet unresolved issues are briefly described in Section IV, Summary & Future Work. (The CSAT responses are ordinal values on a Likert scale of 1 to 5, with 5 representing the highest level of satisfaction. Typical mean values are in the 4.2 to 4.4 range, so there invariably are heavy negative skews, and we are continuing to examine several important non-parametric issues related to this.)

We do not find many references in the research literature to in-depth analyses of large-scale surveys. Most of the research literature refers to small case studies, or to analyses done on dummy databases. We have found that our results, using a large and complex survey, are not always consistent with results reported on in the published research literature for similar analyses these instances will be pointed out in this paper.

II. Data Imputation Issue

A. Background

Data imputation (dealing with missing values) is an important issue to be faced in analyzing data from most surveys, particularly those in which skip logic is used. Skip logic is the branching of questions/responses that occurs when specific questions are answered in certain ways. Considerable research has been done in the research community to try to come up with reasonable ways to deal with missing values over the many types of survey data landscapes this is only a partial list of these methods:

- mean substitution [1], [2], [3]
- listwise deletion [4], [5], [6]
- multiple imputation [7], [8], [9]
- hot deck substitution [10], [11], [12]
- selection rate weighting [5], [13], [14].

Cisco had been using the mean substitution method extensively for several years, starting at a time when the missing value levels were low ($< 10\%$), but as the CSAT survey evolved over time, the missing value levels in parts of the survey gradually but steadily increased, for some responses to as much as 85% missing. Mean substitution uses the mean value of all non-missing cells, for a given variable, to populate all the missing cells for that variable. This is intuitively appealing, since the mean represents, in a sense, a typical response. However, the research literature [15], [16], [17] over the past 10 years or so has been leaning away from the use of mean substitution for situations in which the missing value levels are greater than $\sim 10\%$, Since we were seeing unexpected and strange results coming from some of the regression models used for driver analysis, and we typically saw that these models used at least one variable that had elevated ($> 10\%$) missing value levels, we studied this area in detail. More will be said below on why we think mean substitution, although intuitively appealing, introduces large errors in many cases, even those with $< 5\%$ missing. Our initial analysis examined the use of listwise deletion, and we then evaluated the multiple imputation approach.

B. Analysis Difficulties with Missing Values

Skip logic is an important part of the CSAT survey. We have found that customers will generally only devote 15 min. to answering survey questions the responses after 15 min. are not very reliable and not as complete as the responses prior to that time. Therefore, we need to use the time well, and skip logic branching enables us to get more responses in under 15 min. For example, if a (hypothetical) question, "do you have direct experience using Cisco products?," is answered affirmatively, the respondent is branched into a section of the survey that asks numerous questions about the specific products used. Similarly, for customers' purchasing staff, we try to gather as much specific information about order fulfillment, ease of doing business, etc. There are similar branches for executives and managers, for those who work primarily with Cisco's partners, etc.

Another source of missing values occurs when a survey has non-mandatory questions. Approximately 90% of the current CSAT survey questions are non-mandatory, which is not atypical for large surveys [18], [19], [20].

A major downside of all this branching (as well as having a large number of non-mandatory questions) is that we have, in some areas of the survey, high missing value rates. Table I shows missing value levels for responses to specific questions included in our fiscal year 2011 CSAT survey (conducted from Aug., 2010, through July, 2011).

Since many missing value levels are large, we find ourselves having two choices: 1. Calculate and report results for each major skip logic branch, thereby generating five or more sets of results that need to be interpreted separately or aggregated using a weighting scheme, or 2. employ a data imputation approach that accurately represents the total response population. Choice number one is not a good one

Table I
DATA POPULATIONS FOR CSAT VARIABLES

CSAT questions/ responses	Satisfaction with:	Non-missing values level, % of Q08A	n
Q08A	Ease of doing business	100%	84741
Q112	Product quality	75%	63617
Q15	Partner, pre-sales	62%	52524
Q18	Partner, post-sales	51%	43254
Q34	Technical support team	39%	33371
Q22	Account team	33%	27935
QCS2	Order fulfillment	16%	13814

reporting five or more sets of results to each business unit would cause too much confusion. Having a single result for each survey topic enables the business unit executives to focus on key problem areas (or leverage key improvement areas, where the CSAT results show improvements occurring). Therefore we need to come up with a method that deals with the missing value levels so we can be sufficiently certain that the responses accurately reflect the full population and can be used for linkages-type modeling.

A key point is: If we were sure of the main drivers, we could use only those data rows in which the dependent and important independent variables are fully populated. But we can only be sufficiently sure of the main drivers if the "slice" of data is large enough to accurately reflect the full number of dependent variable rows. Therefore we need to conduct sensitivity analyses to gain confidence that we are accurately representing the full dependent variable population. This is done by examining various levels of missing values, and with various imputation approaches. This way we can gauge when the missing value levels (that are then later imputed or deleted) are low enough to not substantially adversely influence the results, and also determine which imputation method can enable a larger slice of the data to more accurately mimic the full population.

C. Imputation Tests: Listwise Deletion

1) Testing at High Missing Value Levels: "Listwise deletion" simply means using only data rows that are fully populated for the dependent and independent variables. We exclude from the modeling and other analyses all rows that are not fully populated. The major downsides of this method are that we:

- exclude real and valuable data, and often substantial amounts of data
- on occasion, deplete the row count to a level where the modeling results are not statistically significant, or the confidence bands are too wide to enable the results to be useful in practice
- run the risk of slicing the data too thinly, thereby examining unrepresentative sub-populations that do not accurately represent the full population.

Keeping these limitations in mind, we ran a number of experiments with fiscal year 2010 CSAT data to determine if missing value level has a substantial effect on the driver ranks, impact levels, and error bars. An initial set of tests evaluated linear regression models in which the dependent variable is "ease of doing business (EODB)" (Q08A question 08A in the survey, an internally-used designation) responses, and a set of responses to 21 questions are used as independent variables. We want to know which "attributes" (independent variables) affect the general feeling regarding "ease of doing business with Cisco" most strongly. These initial multiple regression models quickly identified three variables that are much more important than the others: a. "Overall satisfaction with Cisco product quality" (Q112); b. "satisfaction with the Account Team" (Q22); and c. "satisfaction with the Technical Support Team" (Q34). (Another variable that is very predictive of EODB in one small branch sub-population is "order fulfillment satisfaction" (QCS2), but we excluded it from the modeling since the missing value level is so high (84%), and this thin slice was later found to be unrepresentative of the full population.) Over the full data population, these three independent variables are the most predictive of EODB of the 21 variables evaluated.

Our modeling approach was this: We started out regression modeling with full mean substitution using these three variables (Q112, Q34, and Q22), then stepwise reduced the mean substitution levels by reducing the imputed data rows using listwise deletion. This was done using the R statistics package we first randomly selected imputed rows for the lowest missing value level variable (Q112), then deleted one percent of those rows at a time and reran the full sequence of models until the Q112 variable had no imputed rows remaining. Then we took the next most remaining imputed variable (Q34) and repeated these steps. We say "remaining" here, since listwise deletion of one of the variables invariably reduces the others to some extent. The final model in this series, then, is the full listwise model in which there are no imputed rows remaining for any of the variables. (We start out with no imputation of the dependent variable—we "listwise deleted" from the full population to only the populated rows for the dependent variable, ease of doing business (Q08A). Also, we have assumed MCAR (missing completely at random) behavior for all these analyses.)

This sensitivity approach works well in assessing whether or not the independent variables display substantial changes in the variable coefficients (or t value, standard error, or model F value), or if there is so much change in coefficient values that coefficient "crossover" occurs. We define "crossover" as the situation in which the coefficient of variable A, say, is found to be less than that of variable B at a certain level of imputation, then becomes greater than that of variable B at a different level of imputation.

We have found that, depending on mean substitution rate, the "ease of doing business" model coefficients, predictor t

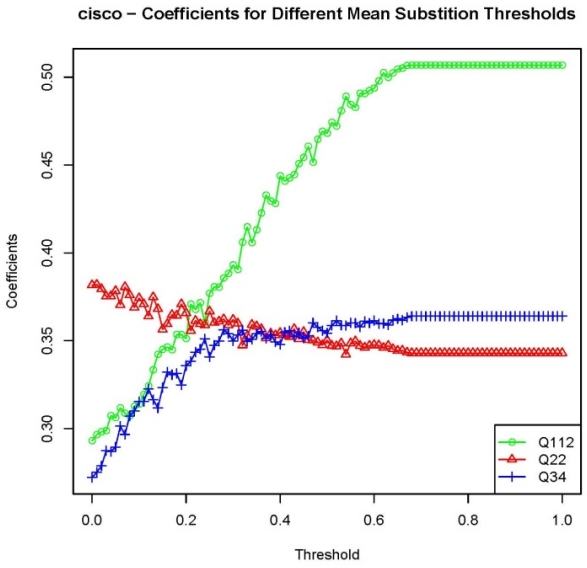

Figure 2. Mean substitution in EODB models

values, and r-squared values, using these three independent variables, change substantially in this stepwise imputation evaluation, as depicted in Figure 2 below (where "threshold" on the x axis is the fraction of rows that have been mean-substituted, up to the missing value level of each independent variable (as given in Table I), and the y axis gives the resulting regression model coefficient).

We see similar coefficient "crossover" patterns using other dependent variables: Q08D ("is Cisco customer-focused?"), Q26C ("satisfaction with Cisco hardware"), Q28 ("satisfaction with Cisco software"), and a number of other key customer perception (experience level) topics, all of which are important in influencing yearly product goals, support team goals, etc. It is clear from these analyses that the coefficient values can change dramatically, depending on the level of mean substitution. Also, the model fit, measured by adjusted r-squared value, also can change substantially, depending on mean substitution level, from R^2 of ~ 0.25 for full mean substitution to ~ 0.40 for full listwise deletion.

Dominance analysis was used (see Sections III.B and III.C for details) to estimate the relative importance of the predictive variables (and these results are fairly consistent with results using the direct coefficient value method of estimating impact). Using listwise deletion, there is a 46% drop in impact for Q112, a 64% increase for Q22, and a 22% increase for Q34, as seen in Table II.

We have also found that care needs to be exercised in including highly colinear variables. When we include 'pre-sales partner satisfaction' (Q15) and "post-sales partner satisfaction" (Q18) responses with the other three independent variables, the resulting models (see Figure 3 below)

Table II
CROSSOVER EFFECT IN Q08A,Q08D MODELS

	Q112	Q34	Q22
Q08A: Ease of doing business			
Mean substitution	48%	27%	25%
Listwise deletion	26%	33%	41%
Q08D: Customer focused			
Mean substitution	44%	36%	20%
Listwise deletion	27%	33%	40%

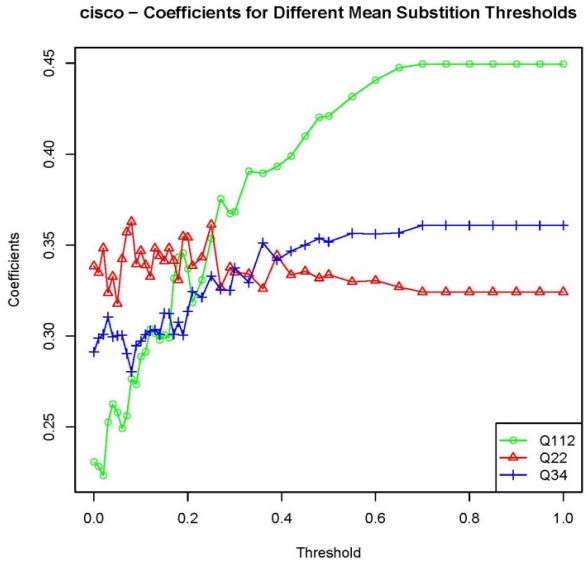

Figure 3. Including Q15+Q18 in EODB models

yield a variable impact profile quite different than that shown in Figure 2, with Q34 ranking high above Q112, in this configuration, at low mean substitution levels. The colinearity between these two variables is 0.69, and we observe strong Q15/Q18 interference, even though the VIF value here of 3.2 is substantially lower than the accepted rule of thumb value of 5 (or in some references, 10), which is commonly used as a threshold for colinearity concern. This topic will be covered more fully in Section III.A. See Figure 3.

2) Testing at Low Missing Value Levels: The research literature often states that low mean substitution levels are usually tolerable, and resulting models and analyses are likely to not be materially affected by mean substitution in the vicinity of 5% or lower [1], [2], [3]. Some researchers state that less than 10% mean substitution is usually tolerable. In mid-2011, our team switched to a largely listwise deletion approach to deal with missing value levels greater than 10%. Confidence intervals did expand substantially for dimensions (primarily the product satisfaction dimension, since these data slices tend to be the thinnest) with low-volume response levels. Even with this change, the region

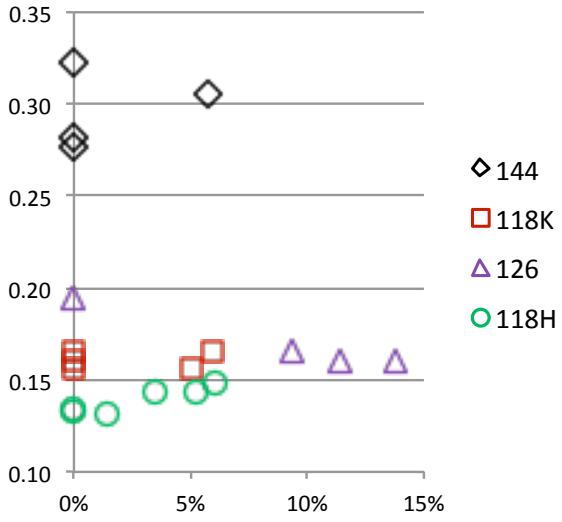

Figure 4. 'Satisfaction with software' (Q28) models

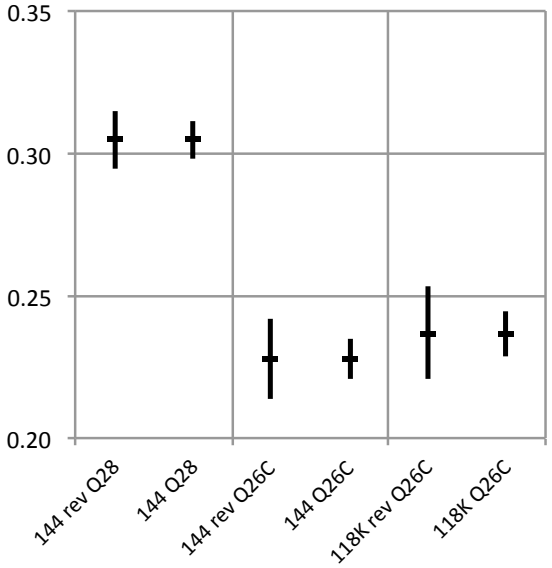

Figure 5. Error bars inflation with mean substitutions

lower than 10% did appear to be problematic in some situations, so we continued our study of that region. Since virtually all of the recent driver analyses include key independent variables that employ mean substitution in that range, it is important to quantify this impact, particularly the impact on the confidence intervals we place around predicted driver strength.

Figure 4 below shows the variation of coefficient level (y axis) for various mean substitution levels (x axis) for the Q28 ("satisfaction with Cisco software") dependent variable models, which use the same 21 independent variables that were used in the 'ease of doing business' models described above. We chose to examine the Q28 models since, in addition to the importance of the Q28 results, the missing value levels of the key variables are all in the $< 10\%$ range, when we use full listwise deletion against the dependent variable. See Figure 4.

There is some "crossover" effect seen here, where variables that have weaker impact at higher mean substitution levels (or vice versa) gain in apparent impact as the mean substitution level is reduced by listwise reduction. Q126 ("satisfaction with software installation, upgrade, and migration") has the highest mean substitution level in the partial mean substitution equation, at 13.08%, and does show a coefficient increase from 0.161 at this level of imputation, to 0.195 in the full listwise deletion model, at $\sim 7\%$ overall mean substitution. We also see fairly large variation in coefficient level for Q118K ("satisfaction with product reliability"), and particularly for Q126, at the zero mean substitution level, although no direct crossover is seen. But since a wide variation is observed, even with fairly low depletion rates and high row counts, we consider the region between 0% and 5% mean substitution to introduce a coefficient (beta) error that turns out to be much larger than the result of the simple standard error (SE) calculation for the coefficient: Beta $\pm 1.645 * SE$, where 1.645 is $t_{.05}$, and the resulting range establishes the 90% confidence interval.

For the error introduced by mean substitution, we use the standard deviation of the coefficient values represented in Figure 4, times 1.645, for a direct comparison with the error bars generated for the coefficients in the original equation. These representations are shown below, in Figure 5, for the Q28 ("satisfaction with Cisco software") and the Q26C ("satisfaction with Cisco hardware") models, with the two models using different independent variables.

There are many data points near the 0% missing value level because, as we depleted each independent variable's rows, the rows of the other variables, both independent and dependent, are depleted somewhat, even in situations in which there were no missing value rows left for these non-targeted variables. Most of the coefficient variance for these variables comes from the variation near their 0% mean substitution level, when the mean substitution levels of the other important independent variables are changed. Approximately three-fourths of the variance here is a real effect of altering mean substitution levels listwise deletion variance has been tested in a similar way, where random rows are deleted after full listwise deletion is achieved, and the associated error is much smaller, roughly one fourth of that seen in the mean substitution test (see Figures 4 and 5). The variance of the coefficient at the zero missing value level should be included in the total variance calculation,

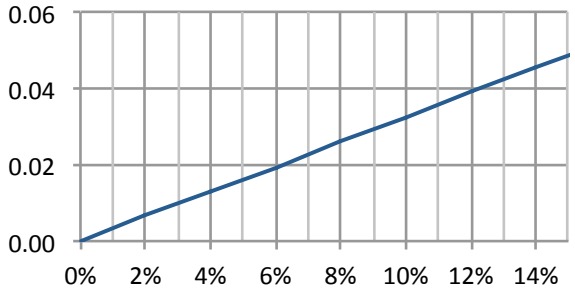

Figure 6. General inflation factor for error bars

we believe, since it is true jitter that results from imputation level alone, and not partially the result of the type of listwise deletion variance described.

The revised (i.e., "rev") error bars shown if Figures 4 and 5 are $\sim$2.5 times the coefficient error bars for the Q28 models, and $\sim$ 3.5 times the coefficient error bars for the Q26C models. From this, it is clear that mean substitution, in the $\sim$7% region for these two separate dependent variables, enlarges the error bars of the coefficients by about a factor of three. Extending this analysis to the other models developed, we have calculated a standard error correction factor that is a function of missing value level. For every one percent in missing value level, in the region between 2% and 13% missing, the correction factor increases by $\sim$ 0.42 units. Figure 6, below, shows the approximate factor (on the y axis) that should be added to the mean standard error (times 1.645) of the model coefficients to construct 90% confidence intervals, as a function of missing value percentage level (on the x axis).

3) Testing Using Training Datasets: We took the listwise deletion set of rows for the Q08A/Q112/Q22/Q34 model (n=16653) and assumed this is a complete dataset. Then we randomly blanked out cells for Q112, Q22, and Q34 to emulate the missing cell rates for the full dataset, that is 24.9% missing for Q112, 67.0% missing for Q22, and 60.7% missing for Q34. We then mean-substituted for the randomly assigned missing values and ran the correlation matrix and models, calling these the "re-mean-substituted" results. Then we listwise-deleted that dataset after returning the mean substitution cells to blanks (with the resulting dataset n=1613), calling this the "re-listwise" set. Since the n=16653 dataset is the "correct" answer, a priori, this approach tests whether or not the mean substitution method gives an answer that is close to being correct, as defined by the listwise deletion results (and confirmed as so in Section II.C.4). If the "re-listwise" dataset gives an answer that is close to the original n=16653 dataset, we know that we have not sliced the data too thinly.

The results, using both coefficient comparisons and dominance analysis (see Sections III.B and III.C) clearly show

Table III
RE-MEAN SUBST. & RE-LISTWISE DELETION

case	data view	n	Q112	Q34	Q22	R^2
1	mean subst.	84471	0.506	0.365	0.334	0.24
2	listwise	16653	0.293	0.272	0.382	0.39
3	re-mean subst.	16653	0.484	0.345	0.334	0.26
4	re-listwise	1613	0.301	0.320	0.387	0.41

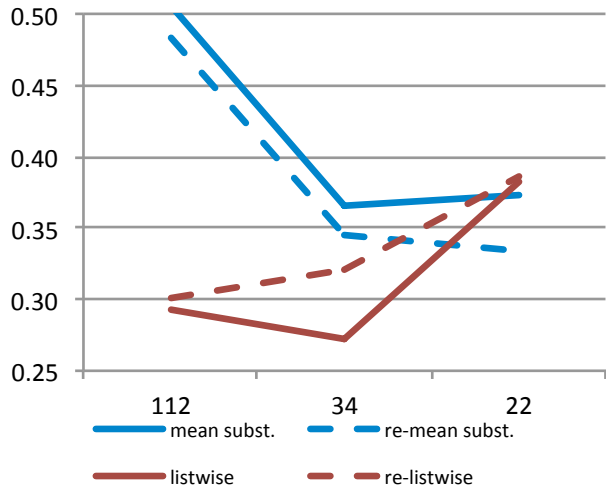

Figure 7. Re-mean substitution and re-listwise deletion plots

that mean substitution is inappropriate, even at low substitution levels, giving results far different from the full n=16653 test dataset, and in fact giving results similar to the mean substitution results for the full n=84471 dataset. This alone does not mean that listwise deletion is the best approach, but in conjunction with the sample size experiment done in Section II.C.4, builds a case that listwise deletion probably yields accurate results when the data volume is greater than $\sim$ 1000 for this highly skewed data (the mean skew is -1.4, for these variables, with a range of -1.0 to -1.8). This analysis also confirms that mean substitution, even at only a few percent, does not yield accurate results with this type of data. The slight variations between the original data and the "re-" variants indicate that there is some inaccuracy at low data volume, although not much. Table III and Figure 7 display the coefficients comparison.

4) Sample Size Testing: One way to circumvent the confidence interval expansion described in Section II.C.2 is to eliminate mean substitution in situations where too-thin data slicing is found not to be a problem. To test whether or not we are slicing the dataset too thinly, and thereby inadvertently selecting unrepresentative portions of the full dataset by using listwise deletion as an alternative approach, we constructed univariate EODB models using the lowest populated variables (Q22 at 33%, 27935 rows, of Q08A's level; Q34 at 39%, 33371 rows) to see if the "crossover"

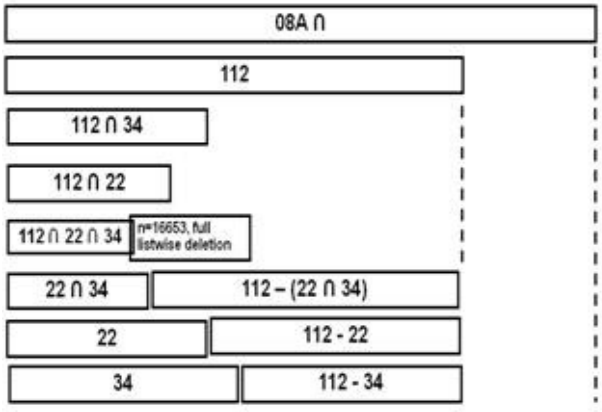

Figure 8. Thin slice testing of EODB model data

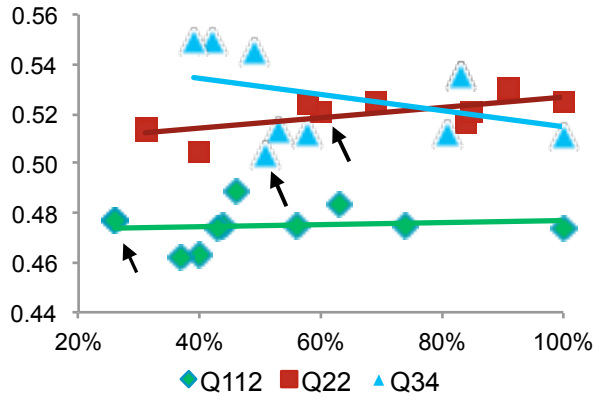

Figure 9. Correlation Matrix for Predictive Variables

of coefficients and t values, as seen in the above examples, occurs. For these models, we sliced the independent variable data in several different ways, over various intersection regions of these independent variables, as is shown in Figure 8, and compared the resulting regression coefficients.

In this test, we increased the maximum "coverage" of Q22 from 59% of the complete Q08A data to 81% and continue to see the crossover, where Q22's coefficient is 26% higher than Q112's, and the crossover yielded a similar 31% higher with the smaller subset of Q22 rows. Likewise, we increased the coverage of Q34 from 49% of the complete Q08A data to 83% and continue to see the crossover, where Q34's coefficient is 15% higher than Q112's, and the crossover yielded a similar 12% higher with fewer Q34 rows.

In Figure 9, below, the correlation matrix R values (on the y axis) of the three most predictive independent variables against Q08A (EODB) show that from full listwise deletion (see black arrows; with data completion level on the x axis) to zero deletion (100%), the correlations with the dependent variable are fairly constant, indicating that overly thin slicing of any of these key variables is not occurring. See Figure 9.

5) Testing 'Nonsense' Variables: When we model only with variables that seem to have nothing to do with Q08A ("ease of doing business"), we find that the apparently 'nonsense' independent variables, at high levels of mean substitution, achieve regression coefficient values, F values, t values, and R2 values that are comparable to those of models using more "sensible" independent variables that also incorporate high levels of mean substitution. In Figure 10, "technical innovation" (Q315), "high availability features" (Q118J), and "interoperability" (Q118F), at high levels of mean substitution, yield model parameters that are comparable to those of variables that make more sense as influential variables, such as "satisfaction with the Account Team" (Q22), and "satisfaction with Technical Support" (Q34), when the sensible variables experience high levels of mean substitution. See Figure 10.

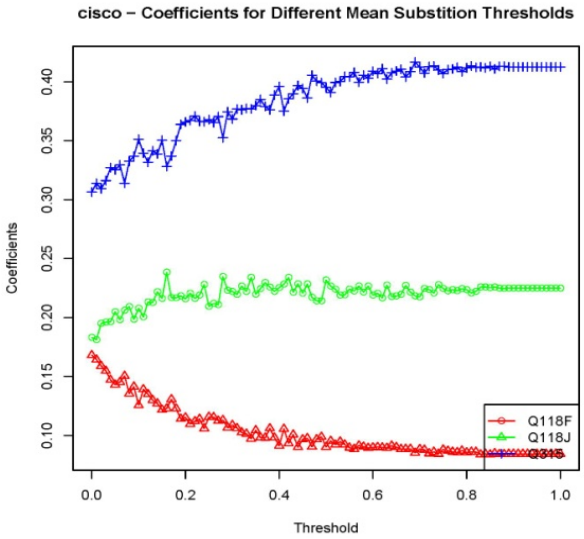

Figure 10. "Nonsense" variables for EODB

Including these apparent "nonsense" variables, together with the more sensible variables analyzed above ("overall satisfaction with Cisco product quality" (Q112), "satisfaction with the Account Team" (Q22), and "satisfaction with Technical Support" (Q34)), fortunately retains the "proper" order of coefficients, but it does reveal an interesting situation: If a 'nonsense' variable has a low missing value level, we can get a nonsense result that mimics other models where the main driver appears to be another low missing value variable.

Since we intuitively consider "ease of doing business" to be influenced by a different set of drivers than Q08D, "customer focused," it is surprising and disconcerting to see the similarities of the models. It appears that missing value level has a very strong influence on the model results with this type of data.

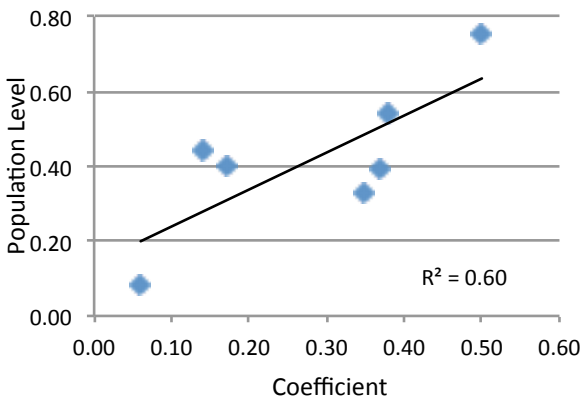

Figure 11. Missing value levels v. coefficients

In this vein, if we plot the inverse of missing value level (calling this "population level"), for both the presumably impactful and the presumably "nonsense" variables, against regression coefficient value, we see an interesting correlation that suggest that the higher the population level, the higher the coefficient value, apparently largely independent of the actual survey values in the populated cells. See Figure 11:

6) Imputation Tests: Multiple Imputation: We are currently using R/Amelia to generate models, using the multiple imputation method, to compare to the mean substitution and listwise deletion models reported on here. These multiple imputation models are found to be very similar to the mean substitution models described. We are not reporting these detailed results at this time - further work is needed and is planned.

III. Colinearity Issue

A. Background

The traditional way to compensate for high colinearity among independent variables is to rely on the increase in the coefficient standard errors that results from high colinearity (and therefore high VIF, variable inflation factor):

$$s_{b_k} = \sqrt{\frac{1 - R^2_{YH}}{(1 - R^2_{X_k G_k}) * (N - K - 1)}} * \frac{s_y}{s_{X_k}}$$

$$= \sqrt{Vif_k} * \sqrt{\frac{1 - R^2_{YH}}{(N - K - 1)}} * \frac{s_y}{s_{X_k}}$$

where S_{bk} is the total of the standard errors, H is the set of all X independent variables, G_k is the set of all X variables except X_k, and R represents the variance for each set. The larger $R^2_{X_k G_k}$ is (i.e., the more highly correlated X_k is with the other independent variables in the model), the larger the standard error will be. One major problem is that, as the X_s become more highly correlated, it becomes more difficult to determine which X is actually producing the effect on Y. Also, the reciprocal of $1 - R^2_{X_k G_k}$, the variable inflation factor (VIF), attempts to show how much the variance of the coefficient is being inflated by colinearity. The resulting standard error is proportional to the square root of VIF, so a VIF of 5 (corresponding to a colinearity of 0.80) results in a standard error that is 23% larger than in a situation in which the VIF is 3.3 (corresponding to a colinearity of 0.70).

Adding more highly colinear variables increases the size of the standard errors, especially if the extra variables do not produce substantial increases in R2, as is the case for our models. Adding more variables decreases the (N K -1) part of the denominator also. More variables also decrease the tolerance (i.e., 1 VIF) of the variable, and thereby increases the standard error, therefore adding marginal or extraneous variables also reduces the precision of all the estimates. See [21] for a full explanation of this topic.

In order to adequately compensate in regression models for situations in which there are many highly colinear interactions, in 1987 Kruskal [22] developed the "relative importance" method, and in 1993 Budescu [23] developed "dominance analysis," a more robust alternative to Kruskal's approach. Dominance analysis considers pairs of attributes and their contributions to the squared multiple correlation (R^2), and therefore is effective in focusing on the marginal impact of each variable on the regression model's predictive capability. This notion of "pairwise dominance" involves the pairing of all the variable inter-correlations, and the consideration of how each pair affects the dependent variable. By considering all permutations of pairs, this approach enables us to test the impact of adding any number of highly colinear variables to an existing model, thereby resulting in an accurate estimate of the relative strengths of the variables in the model [24], [25].

B. Colinearity Tests

Dominance analysis was used by us to estimate the relative importance of the predictive variables for both the best mean substitution EODB model and the best listwise deletion EODB model, with both models incorporating Q112, Q34, and Q34 responses. In addition to the aforementioned crossover difficulties encountered, we see another type of crossover effect here when we compare the coefficient levels approach with the dominance analysis approach—clearly different impacts for the three main variables result, between the dominance analysis method and the strict comparison of regression coefficients.

For the full mean substitution model, the relative importance levels, as gauged by dominance analysis, are 48% for Q112, 27% for Q34, and 25% for Q22, as shown in Table IV's "percent" row. For the listwise deletion model (no mean substitution), the relative importance levels are 41% for Q22, 33% for Q34, and 26% for Q112. If we use listwise deletion, therefore, we see a 46% drop in importance for Q112, a 64%

Table IV
MARGINAL R^2 AND k CHARTS: MEAN SUBST.

Mean substitution - Q08A	R^2	Marginal r-squared contribution 112	34	22
112	0.17	–	0.00	0.03
34	0.12	0.05	–	0.03
22	0.08	0.12	0.07	–
112 x 34	0.17	–	–	0.07
112 x 22	0.20	–	0.04	–
34 x 22 0.15	0.09	–	–	
112 x 34 x 22	0.24			

k	112	34	22
0	0.17	0.12	0.08
1	0.09	0.04	0.03
2	0.09	0.04	0.07
M(C)	0.12	0.07	0.06
Percent	48%	27%	25%

Table V
INDEP. VARIABLE HIGH COLINEARITY.

	Q28_1	Q118K	Q126	Q118G	Q315	Q144	Q118M
Q28_1	1.00						
Q118K	0.71	1.00					
Q126	0.79	0.75	1.00				
Q118G	0.65	0.76	0.68	1.00			
Q315	0.64	0.67	0.62	0.65	1.00		
Q144	0.79	0.68	0.80	0.62	0.59	1.00	
Q118M	0.76	0.72	0.67	0.65	0.68	0.70	1.00

Table VI
COEFFICIENT & DOMINANCE APPROACHES

Q08A	Coefficient (Mean Subst.)	Dominance (Mean Subst.)	Coefficient (Listwise Deletion)	Dominance (Listwise Deletion)
Q112	42%	48%	31%	26%
Q22	28%	25%	40%	41%
Q34	30%	27%	29%	33%

increase for Q22, and a 22% increase for Q34. See Table IV for an example.

For these dominance analysis computation matrices, we display only the three most predictive independent variables, to demonstrate the method more clearly, since adding the additional significant variables changes the major impact levels only slightly. The column, in the top chart, labeled 'R^2' lists the squared multiple correlation for each model. The next three columns show the change in the model R^2 that occurs when each variable is added to the models listed in the first column. In Table IV, when Q112 is added to the model $y = b_1 * Q34 + b_2 * Q22$, this results in a marginal increase of 0.09 to the R2 level. The value of 0.00 shows that Q112's impact is equivalent to that of Q34, since the resulting regression equation's correlation coefficient is unaltered.

The bottom chart (with "k" in the upper left-hand cell) displays how to compare the relative importance of each variable. The first row shows the squared multiple correlation for each independent variable against the dependent variable. The next rows show the average marginal increase in R^2 when each variable is added to an existing model composed of one or two variables. For example, the third "k chart" row in Table IV shows that when Q22 is added to the Q112+Q34 model, the increase in R^2 is 0.07. Similarly, when Q34 is added to the Q112+Q22 equation, the result is an increase of 0.04. The fourth row is the average marginal contribution (M(C)), and the last row ("percent") converts each variable's average marginal contribution, to the squared multiple correlation, to a percentage of the total.

This overall approach is consistent with the Budescu method [23], [25], [26], and is quite close to Kruskal's "relative importance" approach [22]. Both the Kruskal and Budescu methods attempt to factor in the effects of high colinearity among independent variables, since the VIF approach does not adequately include contributions from all interactions beyond the direct interaction of the independent variable with the dependent variable. The effect of high multicolinearity is particularly strong in models, such as the ones we see with the CSAT data, in which there are many independent variables with colinearity values greater than 0.60, and some cases where the values are greater than 0.75. Table V, below, shows an example of this for the Q28 ("satisfaction with Cisco software") correlation matrix.

1) Impact: Coefficients and Dominance Analysis: To compare dominance analysis with the coefficient value method, we include the results (including derived percentages) from Table IV and the equivalent matrices for listwise deletion with percentages-derived results from Table II. We see large differences, in both mean substitution and listwise deletion mode, as shown here in Table VI and Figure 12.

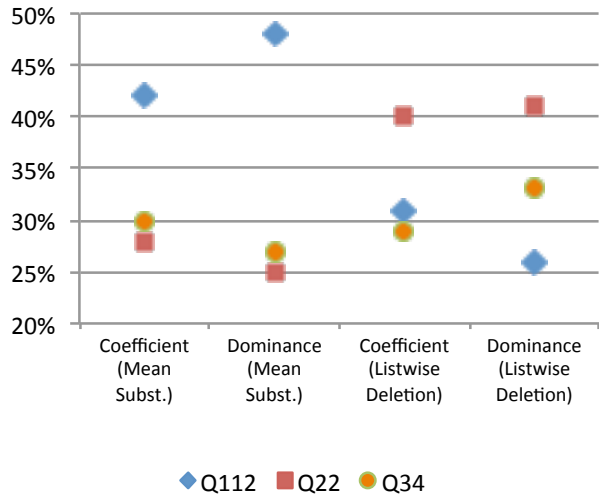

Figure 12. Coefficient & Dominance Approaches

IV. Summary & Future Work

It is essential that we identify drivers that "link" internal engineering metrics and processes all the way to customer perception (as gauged by CSAT survey results). This will enable us to provide guidance (and goaling) for engineering to improve software reliability. The work described in this paper has strengthened our confidence in the order and impact of specific drivers within the customer perception layer of the quality pyramid described in Section I. Key findings of this research are:

- Mean substitution, even at low levels, can cause serious driver impact interpretation problems, whereas listwise deletion appears to give accurate results in a wide range of situations, if the responses volume is adequate. Other imputation approaches are being studied.
- Confidence intervals for the drivers can be constructed using the standard errors of the regression coefficients when listwise deletion is used exclusively. For low response volume situations where mean substitution (at low levels) must be used, a simple method has been developed to modify the standard coefficient confidence intervals to address the observed uncertainty.
- Dominance analysis appears to be a promising approach for gauging the impact of regression variables in models in which many of the independent variables are highly colinear. Further work is needed to construct accurate confidence intervals around the proportion percentages resulting from dominance analysis.

Acknowledgments

The authors are very indebted to John Intintolo of Cisco and Foyzur Rahman of the University of California at Davis for their creative ideas, detailed analyses, and thoughtful critiques extensive and much-appreciated help with all facets of this work.

References

[1] P. Roth, F. Switzer, and D. Switzer, "Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques," Organizational Research Methods, Vol. 2(3), pp. 211-232, July, 1999.

[2] J. Cohen, P. Cohen, S. West, and L. Aiken, L. S. (2003). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Edition. Mahwah, N.J.: Lawrence Erlbaum, 2003.

[3] K. Fung and B. Wrobel, "The treatment of missing values in logistic regression," Biometrical Journal, Vol. 31(1), pp. 3547, 1989.

[4] R. Little and D. Rubin, Statistical Analysis with Missing Data, New York: Wiley, 1987.

[5] T. Raghunathan, "What do we do with missing data? Some options for analysis of incomplete data," Annual Review of Public Health, Vol. 25, pp. 99117, 2004.

[6] W. Vach, Logistic Regression with Missing Values in the Covariates, New York: Springer, 1994.

[7] D. Rubin, Multiple Imputation for Nonresponse in Surveys, New York: John Wiley and Sons, 1987.

[8] L. Collins, J. Schafer, and C. Kam, "A comparison of inclusive and restrictive strategies in modern missing data procedures," Psychological Methods, Vol. 6, pp. 330-351, 2001.

[9] G. Schlomer, S. Bauman, N. Card, "Best practices for missing data management in counseling psychology," Journal of Counseling Psychology, Vol. 57(1), pp. 1-10, 2010.

[10] F. Scheuren, "Multiple imputation: How it began and continues," The American Statistician, Vol. 59, pp. 315-319, 2005.

[11] J. Kim and W. Fuller, "Fractional hot deck imputation," Biometrika, Vol. 91, pp. 559-578, 2004.

[12] C. Musil, C. Warner, P. Yobas, and S. Jones, "A comparison of imputation techniques for handling missing data," Western Journal of Nursing Research, Vol. 24(7), pp. 815-29, 2002.

[13] J. Hendrickx, H. Ganzeboom, "Occupational status attainment in the Netherlands," European Sociological Review, Vol. 14, pp. 387-403, 1998.

[14] T. Bhekisipho, "An empirical comparison of techniques for handling incomplete data using decision trees," Applied Artificial Intelligence, Vol. 23, pp. 373-408, 2009.

[15] J. Schafer and J. Graham, "Missing data: Our view of the state of the art," Psychological Methods, Vol. 7, pp. 147-177, 2002.

[16] J. Graham, B. Taylor, A. Olchowski, and P. Cumsille, "Planned missing data designs in psychological research," Psychological Methods, Vol. 11, pp. 323-343, 2006.

[17] J. Graham, "Missing data analysis: Making it work in the real world," Annual Review of Psychology, Vol. 60, pp. 549-576, 2009.

[18] E. DeLeeuw, "Reducing missing data in surveys: An overview of methods," Quality & Quantity, Vol. 35, pp. 147-160, 2001.

[19] R. Groves, "Nonresponse rates and nonresponse error in household surveys," Public Opinion Quarterly, Vol. 70(5), pp. 646-675, 2006.

[20] J. Krosnick, "The causes off no-opinion responses to attitude measures in surveys," in Survey Nonresponse, edited by R. Groves, et. al., New York: John Wiley & Sons, Inc., pp. 303-314, 2002.

[21] D. Belsley, E. Kuh, and R. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Wiley-IEEE, 2004.

[22] W. Kruskal, "Relative importance by averaging over orderings," American Statistician Vol. 41, pp. 6-10, 1987.

[23] D. Budesku, "Dominance analysis: A new approach to the problem of relative importance of predictors in regression," Psychological Bulletin Vol. 141, pp. 542-551, 1993.

[24] D. Allen and T. Rao, Analysis of Customer Satisfaction Data, Milwaukee: ASQ Quality Press, 2000.

[25] R. Azen and D. Budescu, "Comparing predictors in multivariate regression models: An extension of dominance analysis," Journal of Behavioral and Educational Statistics, Vol. 31, pp. 157-180, 2006.

[26] R. Azen and D. Budescu, "Dominance analysis: A method for comparing predictors in multiple regression," Psychological Methods, Vol. 8, pp. 129-148, 2003.