

AN AUTOMATED BUSINESS INTELLIGENT STOCK MARKET MATCHING ENGINE FOR THE OIL AND GAS INDUSTRY: MATERIALITY OF VIOLATING THE NORMALITY ASSUMPTION OF LINEAR REGRESSION CONTROL ANALYSIS

Avi Rushinek, Ph.D.

*Sara F. Rushinek, Ph.D.****

This study develops an automated matching engine (AME) that correlates an ideal target to available options for use on an Internet web. This investigation looks at a case of mutually exclusive options of oil company investments. With the aid of business intelligence the AME helps an investor select an oil company that most closely matches the investor perception of an Ideal Target oil company. The AME downloads 4 sample oil company options, with 61 decision criteria. The AME converts the categorical textual data into binomial numerical data, so that the AME can regress the data. After that conversion, the AME calculates central tendencies of the data, producing a default Ideal Target for the investor. The user can over-write the default Ideal Target that the AME produced. If the user accepts the AME choice of the default, the user does not have to enter any data.

After the data entry has been completed, the AME correlates the Ideal Target oil company with each of the available options. Then the AME rank orders the options in descending order of magnitude

* The authors are professors at the University of Miami.

** The authors will provide additional appendices, tables, charts, and program code listings, upon written request.

of their correlation coefficients with the Ideal Target. The AME picks the highest correlating option company as the Top Pick best choice, discarding the other options. This study focuses on the case of mutually exclusive options, and the materiality of violating the normality of distribution assumption of linear regression analysis. The AME repeats the same analysis before and after normalizing the raw data, then the AME compares the top ranking company. If the top ranking is not materially affected by the normalization of the data, the AME concludes that the top choice of the raw data analysis is valid, disregarding violations of the normality assumptions. Otherwise, if the normalization materially alters the top ranked choice, the AME concludes that the violations of the normality assumptions render the linear regression matches insufficiently reliable. Thus, the business intelligent system recommends other methods of selection.

To make the analysis more robust, the AME uses additional statistical decision criteria such as the Intercept, Slope, R-Square & Correlation, as well as the concept of theoretically perfect correlation. The AME picks the choice that minimizes the difference between the theoretically perfect correlation parameters and the parameters of each optional choice company. To control for the violations of the normality assumptions, the AME calculate the pre and post normalization values and compares them by the stability of the top pick. The AME conducts several linear regression analyses and an ANOVA (Analysis Of Variance) to finalize its selection of the best choice. Likewise, the AME tests the null hypothesis and rejects or accepts them depending on the statistical results.

We designed such a decision process for the Internet and continuous selection process of matching available options supply with the demands of users. Although we have selected only 4 options and 61 criteria, the implications are applicable to a much larger number of options and criteria. Furthermore, neither the number of options nor the criteria have to be fixed at a set number. They can be variable and change from time to time.

Fuzzy Logic Definition & Explanations

Fuzzy logic is an approach to computing based on “degrees of truth” rather than the usual “true or false” (1 or 0) Boolean logic. Dr. Lotfi Zadeh of the University of California at Berkeley first advanced the ideas of fuzzy logic. Dr. Zadeh was working on the problem of computer understanding of natural language. Translating natural language (like most other activities in life and indeed the

437 An Automated Business Intelligent Stock Market Matching Engine

universe) into the absolute terms of 0 and 1 can be very tedious and lengthy, even for a machine. Whether everything is ultimately describable in binary terms is a philosophical question that may be worth pursuing. In practice, much data we might want to feed a computer is in some state of work in progress. Frequently, these are the results of computing. Fuzzy logic includes 0 and 1 as extreme cases of truth. It also includes the various states of truth in between so that, for example, the result of a comparison between two things could be not “tall” or “short” but “.38 of tallness.” Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths that we aggregate further into higher truths. In turn, when the data exceed certain thresholds, it may cause certain further results such as motor reaction. A similar kind of process is used in artificial computer neural networks and expert systems. It may help to see fuzzy logic as the way reasoning really works and binary or Boolean logic is simply a special case of it.¹

Statistical Regression Versus Fuzzy Regression

Statistical linear regression and fuzzy linear regression have been developed from different perspectives, and thus there exists several conceptual and methodological differences between the 2 approaches. The characteristics of both methods, in terms of basic assumptions, parameter estimation, and application are described and contrasted. Their descriptive and predictive capabilities are also compared via a simulation experiment to identify the conditions under which one outperforms the other. It turns out that statistical linear regression is superior to fuzzy linear regression in terms of predictive capability. Their comparative descriptive performance depends on various factors associated with the data set and proper specificity of the model.² This is the focus of our study. The utility of the present model will be optimized under these conditions. We direct our methodology to cases that promote linear regression analysis over fuzzy linear regression.

In contrast, fuzzy linear regression performance becomes relatively better, vis-à-vis statistical linear regression, as the size of the data set diminishes and the aptness of the regression model deteriorates. Such cases are beyond the scope of the present study.

¹ <http://whatis.com/fuzzylog.htm>, 2675 bytes, 03Mar97.

² Kim, Kwang Jae, Moskowitz, Herbert; Koksalan, Murat “Fuzzy versus statistical linear regression” *European Journal of Operational Research*, Vol: 92 Iss: 2 Date: Jul 19, 1996 p: 417–434.

However, in future studies we may focus on linear fuzzy regression, instead of linear regression.

Most researchers of matching engines have focused on computer science and educational applications.³ In contrast, this study focuses on matching an investment based on financial analysis consensus estimate of EPS (Earning Per Share) & P/E (Price Earning Ratios). Other studies focus upon engineering and science. An example of this includes “Motion detection is limited by element density not spatial frequency”.⁴ Another example of scientifically oriented studies with a medical emphasis includes: “A fast dynamic link matching algorithm for invariant pattern recognition”.⁵ There are also programming handwriting analysis and speech pattern recognition studies that represent a computer science orientation.^{6,7} Another area of focus that is closer to our area is the study of intelligent planning machine since it deals with planning much like our own study.⁸ However, the focus of his study is more environmental engineering and architecture rather than investment equity planning, the focus of the present study. Our study that matches investments in equities of an Ideal Target based on financial analysts’ forecasts is based on our previous developmental research work in the Oil and Gas Industry.⁹**10111213**

³ Hawkes, Lois Wright and Derry, Sharon J. “Advances in local student modeling using informal fuzzy reasoning,” *International Journal of Human Computer Studies*, 1996 Dec Vol 45(6) 697–722.

⁴ Eagle, Richard A. and Rogers, Brian J. “Motion detection is limited by element density not spatial frequency.” *Vision-Research*, 1996 Feb Vol 36(4) 545–558.

⁵ Konen, Wolfgang K, Maurer, Thomas, and von der Malsburg, Christoph, “A fast dynamic link matching algorithm for invariant pattern recognition.” *Special Issue: Models of neurodynamics and behavior. Neural-Networks*; 1994 Vol 7(6–7) 1019–1030.

⁶ Matsushima, Takaji; Morikiyo, Yoshiyuki, and Fujishima, Yutaka, “Analysis of handwriting by the dynamic programming matching” *Japanese Journal of Psychology*; 1993 Dec Vol 64(5) 351–359.

⁷ Ainsworth, W. A. and Pratt, S. R. “Feedback strategies for error correction in speech recognition systems.” *International-Journal-of-Man-Machine-Studies*; 1992 Jun Vol 36(6) 833–842.

⁸ Wyatt, Ray “An intelligent planning machine,” *Computers,-Environment-and-Urban-Systems*; 1991 Vol 15(3) 203–214.

⁹ Rushinek, A. and Rushinek, S. “Forecasting Sales, Expenses & Stock Market Values by Quarterly Financial Statement Ratio Analysis: A Microcomputer Software Development Model for Applied Magnetic Corp. and the Electronic and

439 An Automated Business Intelligent Stock Market Matching Engine

Normal & Log Normal Transformation & Linear Regression Assumption Violation

Another problem that pertains to the present study is the transformation of raw data.¹⁴ Ratings of suppression were log transformed to improve the approximation to a normal distribution. Second, linear regression models for the mean values of seizure regularity and suppression were determined using age, gender, as predictor variables. This study is also using regression, except, that we use simple normal transformation to improve the approximation to normal distribution. In the future, we may also use the log normal transformation and compare the results to simple normal transformation.

Business researchers have also used the log normal transformation. An expression is found for the mean and the variance of the Incurred But Not Reported (IBNR) claims in a lognormal linear regression model. The chain ladder model is considered as a special case. The unique uniformly minimum variance unbiased estimator (UMVUE) and the maximum likelihood estimator (MLE) of those quantities are derived. The variance of the UMVUE of the mean of the IBNR claims is calculated. An estimator not involving an infinite series is found that provides an excellent approximation to the UMVUE of the mean of the IBNR claims. Finally, the claims

Electrical Equipment Industry,” *MANAGERIAL AUDITING JOURNAL*, Vol. 10, No. 2, 1995, 7–33.

¹⁰ Rushinek, A. and Rushinek, S. “Net Income-to-Sales Ratio Grouping For the Crude Petroleum and Natural Gas Industry Sector: A Company to Peer Comparative Variance Analysis”, *OIL AND GAS TAX QUARTERLY*, June, 1995, Volume 43, No.4, 713–728.

¹¹ Rushinek, A. and Rushinek, S. “Oil & Gas Market, Sector & Company Relative Strength Congruity Trading Theory (MSCRSSCTT): An Internet Automation Approach Applied to Chevron Corporation’s Stocks” *Oil, Gas & Energy Quarterly*, Vol. 48, No.4, June, 2000, 671–682.

¹² Rushinek, A. and Rushinek, S. “Automated Internet Trading Forecast Variance Analysis: An Oil and Gas Industry And BP Corporation Case Study,” *Oil, Gas, & Energy Quarterly*, Volume 52 Number 1 September 2003, 35–54.

¹³ Rushinek, A. and Rushinek S. “Internet Web Server Comparative Forecasting Sales Models: Reversed Accounting of Texaco Corp and The Oil & Gas Industry linear, Exponential, Power, Polynomial & Logarithmic Univariate Regressions,” *Oil, Gas, & Energy Quarterly*, Volume 54, No. 2, November 2005, 87–95.

¹⁴ McCall, W. Vaughn, Robinette, G.-Dave, and Hardesty, David, “Relationship of seizure morphology to the convulsive threshold.” *Convulsive-Therapy*; 1996 Sep Vol 12(3) 147–151.

experience of an insurance company is used to compare the various estimators of the IBNR reserve developed. Several tests and graphs are used to verify model assumptions.¹⁵ This is similar to our model, which also develops several tests to verify model assumptions.

The robustness to nonnormality of the null distribution of the standard F-tests for regression coefficients in linear regression models is investigated. Assuming the errors to be nonnormal with finite moments, the null distribution of the F-statistics is derived. This differs from the normal theory F-distribution. Besides the sample size and the degrees of freedom of error sum of squares, the major determinant of the sensitivity to nonnormality is the extent of the nonnormality of the regressors or the extent of presence of leveraged (influential) observations. The small effect in one direction when all observations are equally influential and the much larger effect in the opposite direction when the observations are extremely heterogeneous in their influences provide extremes of sensitivity within which the sensitivity of the tests will be found. Several specific functions of regressors that can be used to judge the extent of nonnormality of the regressors or the extent of presence of leveraged observations are identified.

The difference between this approach and our approach is that we do not only look for criteria to “judge” the extend on “nonnormality of the regressors” by humans. We want the AME to judge automatically. This task is much more restrictive compared to human judgment.

Initial & Subsequent Data Inputs, Transformations, & Analysis

The Initial Raw Data Input consists of a matrix of 4 optional companies and 61 categorical & ratio decision criteria. The investor will pick the best choice among these optional companies based on 61 decision criteria. We define the scope of the decision narrowly, to focus on one problem at a time. These investment opportunities, the companies, are mutually exclusive. Thus, the investor will pick only one company that is the Top Pick. This is analogous to dating, or mating, where a mate, or date deal with only one partner at a time. Thus, for every dating experience, the options may change and the

¹⁵ Doray, Louis G, “UMVUE of the IBNR reserve in a lognormal linear regression model” *Insurance: Mathematics & Economics*, Vol: 18 Iss: 1, Date: May 1996, p: 43–57.

441 An Automated Business Intelligent Stock Market Matching Engine

investor will have to make another separate and independent decision.

This system downloads these data from the data base. Thus, these data items represent a small subset that an investor can use. The important issue is the inclusion of both categorical and ratio data types. Similar procedures can be used with other sources of data such as Compustat, or the Internet¹⁶¹⁷

The 4 optional oil and gas investment opportunities in our case are: EIDU, EXXON, TEXAC, & CHEV. The decision criteria include categorical (items 2–3) and ratio data (items 4–61). The categorical data include the Exchange and the Primary SIC (Standard Industry Classification) code. In the present case the Primary SIC includes the following industries: 2911 — Petroleum refining, and 1311 — Crude petroleum. The ratio data include company data such as number of employees, shareholders, as well as financial analysis's consensus estimates & forecasts. The financial analysts' data includes 3 years information focusing future forecasts of Earning Per Share (EPS) and ranging from the Number Of Estimates to Price Earning (P/E) Ratios.

Subsequent Raw Data Entry For One Default Ideal Company & 61 Decision Criteria

After the initial data download (from the database or Internet) has been completed, the decision criteria have been established. The AME analyzes the data and established the parameters of valid data entries. For example, the AME can determine that the number of employees is invalid (negative). Since one of the criteria is the number of employees, a valid user entry has to exceed the value of 1, for one employee. Further, the AME can default to the smallest, largest or average number of employees in this way the AME will not force the user to enter any data. Thus, the AME calculates the parameters of the data such as: (1) Min-minimum, (2) Max-maximum, & (3) Avg-average. This gives the user a choice of 3

¹⁶ Rushinek, A. and Rushinek, S. "The Role of the Forensic Accountant in Calculating the Damages Using the 'But-if Analysis in a Case of Internet Day Trader & Online Broker Misconduct Litigation," JOURNAL OF FORENSIC ACCOUNTING, Vol. 1, No. 2, 241–250, 2000.

¹⁷ Rushinek, A. and Rushinek, S. "Forensic Audits of Stock Market Forecasts For Texaco & The International Oil and Gas Industry & Financial Models for Online Broker Litigation," OIL, GAS, & ENERGY QUARTERLY, Vol. 51, No. 4, June, 2003, 659–678.

default values for the Ideal Target choice. This way the AME produces the Default Ideal. In the event that the user does not wish to over-write it, the Ideal Target the AME will equalize the Ideal Target to the Default Ideal.

Converting Categorical To Binomial Data To Regress The Ideal On The Options

In order to regress the Ideal on each of the Options, the AME converts the categorical data into binomial data. Accordingly, the AME will convert the Categorical Variable, Primary SIC, into multiple Binomial Variables. The AME produces one Binomial Variable for each unique value of the Categorical Variable. Thus, if the Primary SIC values were: 2911 — Petroleum refining, & 1311 — Crude petroleum, then the AME will produce 2 Binomial Variables. One Binomial Variable will be the first value of the Categorical Value 2911 — Petroleum refining and the other one will be the second Categorical Value 1311 — Crude petroleum. For companies that belong to this SIC code, 1311 — Crude petroleum, this Binomial Variable, will have a value of 1 = True. Otherwise, for companies that belong to another SIC code, this Binomial Variable will have a value of 0 = False. Such conversion of categorical to nominal data will enable the AME to conduct a regression analysis without losing any information, on categorical verbal variable.

The conversion of categorical to binomial data challenges the AME. The number of variables varies with the number of unique value in the categorical variable. Therefore, the AME does not know in advance how many observations the regression analysis will have to process. Accordingly, the AME has to calculate these values, before it can proceed with these decisions.

Subsequent Raw Data Entry For Ideal Target Company

Whenever the user disagrees with the predetermined default values, the user will have to overwrite them, and enter different values. So, if the user does not like to have the maximal Number of Estimates, or any calculated value, like minimum or average, the user can simply enter another value. The AME will over-write the default with the user entry. At this point this AME solicits the user, investor, for the Ideal Target option, and the user may enter them over-writing the defaults, or may choose to accept the defaults, not entering new values. In either case, the Ideal Target values will be established, and the next step is to decide the top choice.

443 An Automated Business Intelligent Stock Market Matching Engine

Ranking The Options In Descending Order Of Magnitude Of The Deviations From A Theoretical Perfect

The theoretically perfect value of the Intercept equals to zero, 0. Therefore, the Top Pick in this decision statistic is the TEXACO Company option. This company has the smallest absolute value of the Intercept, 7,089. Accordingly, this AME ranks these companies Intercept as number 1. This Intercept is the closest value to zero, the theoretically perfect Intercept value, and should be the Top Pick.

We could use only one decision statistic, such as the Intercept, to choose the Top Pick. However, to make the analysis more robust, we have added 3 additional statistics. The AME will rank all 4 oil company options for each of these 4 decision statistics. The company option that has the highest rank will be the Top Pick. The AME ranks each one of these statistics in ascending order of its deviation from the theoretically perfect values. These values form a base line for this analysis. The next section demonstrates how we calculated these values.

When the Ideal Target (the Ideal) perfectly correlates to an Existing Option (the Option) the Correlation Coefficient should equal a value of one. Likewise, the univariate regression of the Ideal, Y-variable, on the Option, X-variable, will produce a Multiple R, and an R Square, as well as the slope (Beta), all with a value of a perfect 1. In contrast, the Standard Error and the Intercept should both have a value of zero, 0. The residuals should all equal to zero, 0, and the statistical significance level would not be computable.

The plot of the residuals in such a case will be a flat line that merges with the horizontal axis, Variable X. This describes the values of the Ideal Target and an Existing Option. For the vertical axis, Variable Y, the plot will have only values of zeros.

This condition of a perfect correlation between the Ideal Target and an Existing Option is an artificial base line that can help us evaluate other cases. Such cases will most likely have less than perfect correlation. However, the closer they are to the perfect condition the better is the match between the Ideal Target and an Existing Option. Accordingly, we define the best Existing Option, the Top Match, as the option that comes closest to the perfectly correlated existing option.

Subsequent Normalized Data Entry

After the raw data analysis, the AME transforms the data. It

normalizes the data. To normalize the data, the AME subtracts the mean and divides each value by its standard deviation, creating a normal distribution. Such transformation recast the unit of measurement of the data in terms of the numbers of stand deviations away from the variables' (criteria) means. The resulting distribution of the data is more likely to approach a perfectly normal (bell shaped) distribution. This higher proximity to a normal distribution will reduce the degree of potential violation of the basic linear regression assumptions, especially the assumption regarding the normal distribution.

After this normal transformation of the data, the AME regresses and correlates the Ideal with each of the options again, repeating the previous analysis verbatim. The only difference between the initial and the secondary analysis is standardizing the data. If the raw data were perfectly normally distributed, than the ranking before and after the normal transformation should be identical. However, if the raw data were not normally distributed, the ranking of the transformed data may differ from the ranking of the raw data. To quantify the impact of normalizing the data, the AME compiles a % Percent Differenced ($(\text{Raw-Standardized Data})/\text{Raw Data}$) summary table.

Even though the deviations differ between the pre and post normalization, the differences may not be statistically significant. Further, even if they were statistically significant they may be immaterial for the purpose of the Top Pick. This automatic machine base determination, whether the violation of regression analysis are material for this decision making is the thrust of this study. This study develops a methodology to determine if the differences are material for the purpose of picking the best match using a regression analysis.

The AME uses the Smallest Differenced (Raw-Standard or Normalized) Ranking Average for all 4 decision statistics, to evaluate the materiality. If the Top Pick remains the same, then the assumption violations are immaterial. Otherwise, if the Top Pick does not remain the same, then the violations are material. In our case, the Average of the Differences for TEXACO is the smallest absolute differenced deviate, 025. Thus, the AME ranks it as follows: 1 for the Raw Values, 1 for the Standardized Values, 1 for the Differences Values. Therefore, it remains the Top Pick, and the AME concludes that the violations are immaterial. This would have been true even if this company were ranked first in only 2 out of these 3 types of data transformations. The single most important

445 An Automated Business Intelligent Stock Market Matching Engine

measure is the stability of the top ranked choice, across the (1) Raw, (2) Standardized, & (3) Differenced data. When the ranking is perfectly stable, the linearity violation assumption is immaterial for picking the top ranked option. The same may not be true for other ranks. Therefore, we have to be careful and restrict ourselves to mutually exclusive investment opportunities.

This concludes the data processing part of this study, and brings us to the hypothesis testing, analysis and discussion. Our main hypothesis focuses on the possible violation of some linear regression assumptions and the needs for additional controls to decide whether such violations have material effect on our decision. If the regression produces insignificant statistical results and we are going to discard it, and not base our decision on it, we do not have a problem. However, if the regression is statistically significant, and we intend to base our decision on it, then this AME provides some additional steps to mitigate the adverse affect of not meeting the normality of distribution assumption of a linear regression analysis. Therefore, our first step is testing the hypothesis concerning the statistical significance of the regression parameters.

Regression ANOVA Null & Alternate Hypothesis Testing & Interpretation

The basic regression & ANOVA (Analysis Of Variance) null hypothesis states that no significant statistical parameters exist. In the case of our regression analysis the parameters we test include the R-Square or the Correlation, The Slope, & The Intercept. In the case of the ANOVA, this AME focuses on the differences among the above regression parameters and the theoretically perfect values of these parameters.

A “typical user” that conducts a similar analysis of correlating an Ideal Target to several Existing Options will first test the statistical significance of the regression model. If the results are insignificant, the Ideal Target is not correlated to the Options at all, the user will be inclined to abandon that option. However, if the option highly correlated, the user will retain that option for additional review. Further, if that option is also the most correlated the user will tend to pick it as the Top Choice.

Following the foot steps of such a user we will test the null hypothesis of the regression analysis of the Ideal Target on an Existing Option company. Our typical null hypothesis will state that the regression parameters including R-Square, & the Slope are all

equal to zero. In contrast, the alternative hypothesis states that these parameters are different from zero.

The AME regresses the raw data of the ideal target company on the EIDU Oil, and on each of the other oil companies. Since testing the statistical significance equally applies to all of them we will use this company to illustrate the procedure. Due to the low Significance F value, 8.8409E-134, being less than .05, the AME rejects the null hypothesis stating the R-Square is equal to zero. The “typical” user will repeat the same analysis for each of the options, and then simply pick the option with the highest R-Square as the Top-pick. The problem of this approach is that if indeed the linearity assumptions of the distribution have been greatly compromised, than the Top Pick may be due to that violation rather than simply to the similarity of the Top Pick to the Ideal Target. To ensure that this is not the case, our AME takes an extra ANOVA step. This step confirms that if the normality assumptions have not been violated (as we know that they usually do get violated) the top pick would have still remained the same choice company.

For this purpose the AME standardizes the raw data transforming it into normally distributed variables. Then, after the data have been normalized, the AME repeats the entire process all over again. The AME calculates the difference of the Pre-normalization, & the Post-normalization deviations of the calculated parameters from the theoretically perfect parameters, producing the Differenced deviations. Then, the AME applies an additional (Two Factor Without Replication) ANOVA to these Differenced deviations. This ANOVA reveals that the null hypothesis can not be rejected, for both the columns at the .05 level of statistical significance. The F value of the Rows, 3.0111 is less than its respective F critical, 3.8625, and the F value of the Columns, 1.3032, is less than its respective F critical, 3.8625. This confirms that the AME can not reject the null hypothesis stating that the differences between pre & post normalization are equal to zero. Thus, the violations do not alter the impact on the top choice.

The analysis is still valid since the top ranked company remains on top, whether we use the raw data or the normalized data. Note that some of the 2nd, 3rd, and 4th ranks may not be as stable as the top rate and may have been altered due to the normalization of the data. However, that does not matter since we have limited our analysis to mutually exclusive options.

447 An Automated Business Intelligent Stock Market Matching Engine

Summary

In summary, this study demonstrated ways to automate the evaluation of the materiality of violating the linearity assumptions of the regression analysis for mutually exclusive option decisions. It shows how an Automated Matching Engine (AME) for the Oil and Gas Industry can make such decisions using predetermined rules of standard statistical procedures. Using univariate linear regression analysis and Analysis Of Variance (ANOVA), the AME decides which option is the best choice, and whether violating the linearity assumption will materially impact that decision.

To carry out this task the AME downloaded financial data of 4 oil companies from a computer data base. Then, the AME transformed categorical data into nominal data and calculated their central tendencies: Minimum, Maximum, & averages. The AME applied one of these statistics to a default value for an Ideal Target, letting the user over-write or accept these defaults. The AME regressed this Ideal on each of the Option companies, ranking them in diminishing order of magnitude of their Correlations. Finally, the AME selected the single most correlated company (to the ideal) as the top pick, discarding the other choices. To ensure immateriality of the normality assumption violations, the AME normally transformed the data and then repeated this process standardized data. After confirming that the top picks remain the same choice, and further proving it with ANOVA (2 factor without replication), the AME concluded its analysis.

Conclusions

We can assess the impact of violating the normality of distribution assumptions of a univariate regression analysis, using standard statistical tools. Further, an AME can carry out such analysis automatically using predetermined rules. While we have demonstrated this for mutually exclusive choices, it may not be applicable in the same way for mutually inclusive choices. This is because the top choice may be more stable than the ranks of all other choice in cases of normality assumption violations. Therefore, dealing with inconsistent 2nd, 3rd and 4th rankings may be more challenging than just dealing with the top rank. However, for many problems, that are mutually exclusive in nature, such as some investment decisions, mating and dating decisions, this approach can be useful. In such cases, after the initial date has been selected, the selection process has to be repeated all over again. However, due to the advent of business intelligence, the additional receptions are fast enough to

remain practical. Furthermore, if a mate of an investment choice has already been selected, kept or discarded; it will have to be excluded from the future pool of options. It is no longer qualifying as any option that has not been picked yet. In such case, a mutually exclusive process is mandatory.

Since financial data of the oil and gas industry is usually skewed and tends not to be normally distributed, the procedure that we have used should be adopted as a standard quality control measure. This is especially applicable in cases of automation, where additional calculations do not pose significant difficulties. This is more relevant in cases in which linear regression & ANOVA are readily available, such as in standard spreadsheets. Other non linear regression procedures are not as prevalent. Thus, wrong decisions can be made just because of lack of an additional analysis.

Implications

Due to the volatility in the Oil and Gas Industry we try to find ways to reduce errors. The implication of this study is that investors make incorrect decisions due to the violations of the normality assumption and the lack of controls to measure its materiality. Future studies may extend this approach to mutually inclusive options, and the violations of other assumptions. Oil investors many times have to violate a variety of assumptions due to time and other constraints. Therefore, it is important not only to focus on how to do it properly in the first place, but also on how to control for when things are not done under optimal laboratory conditions. It would be interesting to test these models over time and with various other Oil and Gas companies.