The Effect of Uncertainty Tone on Earning Forecast Accuracy in Conference Call

Dan Huang

643 Regression Project

2018 Fall

Contents

1 Introduction

Conference calls are implemented as one of the channels to communicate with outside stakeholders and provide more information for them increasingly every quarter or every year. Annual reports or earnings release are written in the paper for stakeholders to read instead of asking or answering questions to providers. However, one particular feature of conference calls is to arrange a section for analysts to ask and answer questions. During the management presentation period, managers will present the performance of the last periods and tell the audiences for expectations of future periods before questions and answers. Analysts can prepare specific questions before asking managers or blow brains in thinking about question in one second. Hence, it is difficult for managers sometimes to answer the questions easily or properly. In addition, managers might refuse to answer the questions unexpected or provide ambiguous answers to them.

Therefore, managers' communication with outside stakeholders may hide some significant information due to the inefficient answers. Textual analysis research is a classical technique method by using specific tone to disclose or hide information spoken by managers. In the past studies, tone of conference calls gives informative process to stakeholders. Various tones that are spoken by managers are used in different ways such as positive tone or abnormal tone. The primary purpose of this project is to examine whether the uncertain tones during the presentation period can influence the analysts' earnings predictions. The performance of uncertain tone during conference calls affects analysts' interpretation for uncertainty, which could cause the less precision of earnings forecasts. The effect of uncertain tone used in conference calls on earnings forecast accuracy will be investigated in this project.

2 Description Data

The data set is provided by Institutional Brokers' Estimate System (IBES) in 2015. There are 731 observations in the data that contains 7 independent variables and one response variable.Also there are 37 observations of values in the response variable (Y), which means there is no difference between the resorted earnings and the analysts' forecast earnings in these observations. In addition, there are no missing values in this data file. The response variable is error median difference (Y) that is measured by the difference between the reported earnings and the median of analysts' forecast earnings within 3 days after the conference call. My consultant Miss Karry told me that error forecasts will be more precise when error-MED is decreasing. FACTIVA database provided all of the transcripts of conference calls. The uncertainty (X_1) represents the count of uncertainty words listed in Loughran and McDonald (2011). X_2 and X_3 represented as positive tone and negative tone, respectively. LnLength (X_4) is measured as the total word count of the conference call transcripts. LnAssets (X_5) is calculated by the logarithm of assets. X_6 represents the number of analysts from IBES. STD-RET (X_7) is called the standard deviation of returns.

3 Estimation and Analysis for Models

3.1 Full Model Checking

In this project, ordinary squares method is implemented to estimate the full linear regression model. Given the data file, there are 731 observations, 7 independent variables ($X_1 - X_7$) and one response variable (Y). These measured variables are indicated below:

Y is the difference between the reported earnings and the median of analysts' forecast earnings within 3 days after the conference call;

X_1 is the count of uncertainty words listed in Loughran and McDonald (2011);

X_2 is the count of positive words listed in Loughran and McDonald (2011).

X_3 is the count of negative words listed in Loughran and McDonald (2011);

X_4 is the total words count of the conference call transcript;

X_5 is the logarithm of assets;

X_6 is the number of analyst from IBES;

X_7 is the standard deviation of returns.

By running SAS, we obtain the results of the estimated full model and all tests of assumptions are done at significance level α=0.5. SAS outputs are shown in **Appendix I** Table 1.1-1.6. The predicted model is:

$$\hat{Y}=0.02513+0.061965X_1-0.10395X_2+0.1465X_3-0.00507X_4-$$
$$0.00001173X_5+0.00022514X_6+0.21839X_7$$

From the Table 1, we know that the global F-test is statistically significant with F-value=7.55 and p-value (F>7.55) is less than 0.0001, implying that there is at least one predictor that is useful for prediction of the difference between the reported earnings and the median earnings, representing the response variable Y. However, the R^2 of full model is 6.81% and Adjusted R^2 is 5.91% from Table 2, suggesting only 6.81% of the total variability in error median difference is explained by the regression model and hence the full estimated model is not good.

From the analysis of normality test and homogeneous variance test, Table 1.4 and Table 1.5 tell us that all p-values of normality test Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling are less than 0.05, which indicates that errors are not normally distributed. The residual analysis shows that F-value is 3.79 and p-value (F>3.79) is 0.0005, suggesting that variances of errors are not equal.

For further analysis, however, we want to figure out how well each independent variable explains the response variable in the presence of all other variables. In the Table 1.3, we notice that only p-value of partial tests of regressor X_7 is less than 0.0001, which suggests X_7 is significant in the full model containing all other variables.

According to the tests above, two assumptions that are normally distributed errors and homogeneous variances are both violated and R^2 and adjusted R^2 are pretty low, which suggests this full model is not fitted well. In addition, only X_7 is significant in the full model in the presence of all other variables. Therefore, we need to take transformations for stabilizing variance and making errors normal. Also improving fit and prediction capability is another goal to entertain.

3.2 Transformation on Response Variable and Regressor Variables

Transformations on response variable or one or more of the independent variables are undertaken to stabilize variance and make errors normally distributed. Another goal is to improve fit by increasing R^2, adjusted R^2 and lowering mean square error and prediction capabilities of the model.

In the situation that homogeneous variance and normally distributed errors are violated, taking the transformation on response variable is not only to stabilize variance but also make errors normally distributed. In this case, the response variable error median difference (Y) is the range of 0 and 1, suggesting the reciprocal transformation on response variable might be appropriate. Here, we take reciprocal transformation on response variable Y to stabilize variance and make errors normal. There are 37 observations that have values of zero in the original response variable data. Since we take reciprocal transformation that makes no sense on zeros, I delete all these zeros of response variable for further analysis after consulting to Miss Karry. Observing the Table 2.1-2.3 in **Appendix II**, we know that the global F test in which F=1.49 and p-value=0.1661 tells us there is no significant difference explained by predictors on the response and errors don't follow normal distribution after reciprocal transformation on response variable to fit the full regression model. However, the assumption of equal variances is satisfied now.

We decide to take Box-Cox transformation on response variable and Box-Tidwell on predictor variables to improve the fit and prediction capability of the model. The

Box-Cox method that gives us the appropriate transformation on response variable is YT=$Y^{0.01}$. Through the Box-Cox transformation, the fit of model and prediction capability has been improved a lot from the analysis of Table 2.6-2.8 in **Appendix II**. Fitting the full model with transformed response variable only, we notice that F-value in the global F test is 16.93 and p-value (F>16.93) is less than 0.0001. The R^2 and adjusted R^2 has been increased to be 14.73% and 13.86%, respectively, suggesting 14.73% of the total variation in the response is explained by this regression model. From the parameter estimates table, X_3, X_4, X_5 and X_7 are significant in the model containing other variables.

Based on the formula of Box-Tidwell transformation result, we know that the appropriate transformation on predictor variables are:

$$\text{XT1}=X1^{-165},\ \text{XT2}=X2^{3.2},\text{XT3}=X3^{2.87},\ \text{XT4}=X4^{15.17},\ \text{XT5}=X5^{-3.23},$$
$$\text{XT6}=X6^{2.56},\ \text{XT7}=X7^{1.12}.$$

Since values of X1 are too small, the transformed values are very large after transformation to explain. We decide to take natural log of X1. The final transformation results of fitted model by implementing transformed response variable and predictor variables are displayed in the Table 2.10-2.12. The estimated transformed full model is:

$$\widehat{Y}^{0.01}=0.93365 +0.00045\text{XT1} -38.23\text{XT2} +112.09\text{XT3} -2.87\text{E-17XT4} +1.18\text{XT5}$$
$$-0.0000016\text{XT6} +0.1\text{XT7}$$

Overall, the global F-test value has been improved to be 17.18. R^2 and adjusted R^2 has increased to 14.92% and 14.05%, respectively, indicating 14.92% of total variability in the response YT is explained by the regression model. At the same time, the partial t-tests indicate that the p-values of independent variable XT3, XT4 and XT7 are less than 0.05, suggesting each of them is significant in the model containing other independent variables. Perhaps there exists multicollinearity among the

predictor variables. Therefore, we will discuss the diagnostics of multicollinearity in the next section.

3.3 Multicollinearity Diagnostics

In this section, we discuss the untransformed full model, transformed full model and reduced model with significant independent variables through partial t-tests in both centered and scaled data and also scaled data only. The results of multicollinearity diagnostics are shown in **Appendix III** Tables 3.1-3.6.

From the centered and scaled data of estimation of untransformed full model in Tables 3.2-3.3, all of variance inflation factors (VIF) are not larger than 4, indicating then there almost certainly is no multicollinearity in the data and there is no need to continue with the diagnostics. From the scaled data only of estimation of untransformed full model in Table 3.1, as the eigenvalues sum to p=8 and there is a column in the Collinearity Diagnostics box with title "Intercept". The largest condition index is 110.99296 that far exceeds 30, which suggests that there is one serious dependency. The dependency is between the intercept and X_4 by looking at the proportion variation, implying X_4 has small range of values.

We then analyze multicollinearity diagnostics for transformed full model in the scaled data only. Looking at Table 3.4, the largest condition index is 52.61388, which is greater than 30, which suggests that there is one serious dependency. The dependency is between the intercept and XT1 by looking at the proportion variation, indicating XT1 has small range of values.

Table 3.5 is the result of centered and scaled data both for transformed full model. All of variance inflation factors (VIF) are less than 4, suggesting there almost is no multicollinearity in the data for transformed predictor variables. Table 3.6 is the result of centered and scaled data for the chosen model. All VIF's are less than 2, which indicates there is no multicollinearity issues among regressor variables.

Overall, there is no multicollinearity issues existing among the regressor variables

for untransformed and transformed models.

3.4 Model Selection and Cross-Validation

In this section, we split the data into two parts, using n_1 part to select the "best" model and n_2 part to validate the model.

3.4.1 Splitting Data

After deleting the zero values of response variable, we have 694 observations left to be divided into twp parts: n_1=602 and n_2=92. I decide to use R language to generate two parts data randomly to separate.

3.4.2 Model Selection Criteria

To fit and choose the "best" model, R^2, adjusted R^2, MSE, forward procedure, backward procedure, stepwise procedure, Cp and PRESS are used for model selection criteria. We implement 602 observations that are transformed to fit the model by SAS and the summary results of model selection criteria are presented in **Appendix IV** Table 4.1-4.11.

Given the comparisons from all the results presented in Table 4.1-4.8, we decide to choose the model with regressor variables XT3, XT4, XT5, XT6, XT7 as the "best" model. All three selection approaches such as forward, backward and stepwise choose this model. R^2 and adjusted R^2 also rank this model as the top 4. PRESS, MSE and Cp choose different models given fitting the untransformed model.

Next we will use the chosen model to test the significance and fit in the n_1 data set. All the predictor variables are significant at level $\alpha = 0.5$ except XT6 variable shown in Table 4.9-4.11. In the chosen model, the global F-test value is 21.08 and p-value (F>21.08) is less than 0.0001, which means at least one regressor is significant for the prediction of response variable. R^2 and adjusted R^2 has been improved to be 15.03% and 14.31%, respectively, meaning that 15.03% of the total variability in the

response variable is explained by the regression model.

3.4.3 Cross-Validation

In this section, we will validate the "best" model chosen previously in the n_2 data set containing 92 observations. The correlation between the observed response values and predicted response values is presented in Table 4.12. The correlation is 0.37668 and the p-value is 0.0002, which suggests there is positive correlation between the observed values and predicted values on response and the model is validated successfully. The "best" model chosen is good for the prediction of the response.

3.5 Residual Analysis and Influence Diagnostics

In this section, we will do residual analysis and influence diagnostics to detect if there exists high leverage points, outliers and high influence points. After identifying potential high leverage points, outliers and high influence points, next step is to determine the extent of the influence, in terms of what components of a regression analysis is being affected by the impact positive or negative influence.

HAT diagonal, Studentized residuals and R-student statistics are used to detect where exists high leverage points, outliers and/or high influence points. For the best model we chose, h_{ii} >2p/n=(2*6)/694=0.017. We present the summary of residual analysis and influence diagnostics in **Appendix V** Table 5.1. The outputs indicate that no Cook's D values are larger and equal than 1, suggesting there is no point affecting the regression coefficients. Hence, there is not necessary to check the DFBETAS. Also, the absolute values of DFFIT are no larger than 2, which means there is no point having impact on the fit of the regression model. Therefore, COVRATIO standard is also satisfied.

Finally, we can summarize the data set is satisfying for predicting the linear regression of the best model even though there are few high leverage points and outliers that don't affect the fit and regression coefficients. Therefore, we can keep these

point in the data set.

3.6 The Best Model Fitting

In this section, we finally fit the chosen model with predictor variables XT3, XT4, XT5, XT6, XT7 in the 694 observations data set. We provide the results in **Appendix VI** Table 6.1-6.6. From the ANOVA table, the global F-value has increased to 23.73. R^2 and adjusted R^2 are fitted to be 14.71% and 14.09%, respectively, which indicates this best model explains 14.71% of the total variability for the response variable. The parameter estimates table indicates that variables XT3, XT4, XT7 are significant in the presence of other variables in the partial t-tests. The assumption of normal distributed errors and equal variances are both still violated.

4 Conclusion

Combining all results of multicollinearity diagnostics, model selection and cross-validation, we obtain the best transformed model to fit the data with transformed independent variables XT3, XT4, XT5, XT6 and XT7 to predict the transformed response variable.

The estimated model is:

$$\widehat{Y}^{0.01}=0.9307 + 119.29XT3 \text{ -29E-18}XT4 +1.2227XT5 +154E\text{-}8XT6 +0.103XT7$$

This model explains 14.71% of the total variation in the response variable when R^2=14.71%. Comparing the untransformed full model and transformed full model, there are five regressor variables in the best model. To be honest, this chosen model does not have very good fit for the prediction of the response variable. However, Miss Karry said this R^2 makes sense based on her research on the data file.

5 Executive Summary

Given this project data file with 731 observations and 7 independent variables to process linear regression model for predicting error median difference and investigating the effect of uncertainty tone in conference call on earnings forecast accuracy, the full regression model doesn't satisfy the assumptions of normal distributed errors and equal variance at very beginning analysis. There is only 6.81% of the total variability in the error median difference explained by the linear regression model. The mean square error is 0.00173. In addition, only X_7 variable out of 7 variables is significant in the presence of all other variables in the model. Our primary purpose in this project is to find out how many independent variables are truly needed for prediction of the difference between the reported earnings and the median forecast earnings and test the fit of model.

Firstly, we estimate the full model with all observations and all independent variables. The predicted full model is:

$$\hat{Y}=0.02513+0.061965X_1-0.10395X_2+0.1465X_3-0.00507X_4-$$
$$0.00001173X_5+0.00022514X_6+0.21839X_7$$

This regression model explains 6.81% of total variation in the error median difference, indicating this regression full model is not good even though there is at least one predictor that is significant for the predicting of the error median difference because the global F-test value is 7.55 and p-value (F>7.55) is less than 0.0001. In this model, only one regressor is important for the prediction of error median difference in the presence of all other variables. There may exist multicollinearity in the data set. Moreover, both assumptions of errors that are normal distribution and are equal variance are violated. Therefore, transformation on response variable or predictor variables are necessary to stabilize the variance or make error normal or improve the fit and prediction capability.

Observing the data carefully, it is easy to find out reciprocal transformation on error

median difference may be appropriate and then we delete 37 observation of zeros in the error median difference (Y). However, we obtain the result that there is no significant difference in the global F-test although the equal variance is satisfied now. Next, we take transformations on response variable error median difference and predictors to improve the fit and prediction of capability. Here, we choose to transform the response variable Y with power 0.01. Taking the full regression model with YT= $Y^{0.01}$, we improve the R^2 to 14.73% that means 14.73% of the total variability in response variable is explained by the regression model and there are four significant independent variables X_3, X_4, X_5, X_7 in the model containing other variables. We transform all independent variables and combine the transformation on response variable, which are presented like

$$\text{YT}= Y^{0.01},\ \text{XT1}=\log(X_1),\ \text{XT2}=X2^{3.2},\text{XT3}=X3^{2.87},\ \text{XT4}=X4^{15.17},$$
$$\text{XT5}=X5^{-3.23},\ \text{XT6}=X6^{2.56},\ \text{XT7}=X7^{1.12}.$$

We get the fitted model:

$$\hat{Y}^{0.01}=0.93365 +0.00045\text{XT1} -38.23\text{XT2} +112.092\text{XT3} -2.87\text{E-17XT4} +1.179\text{XT5}$$
$$+0.00000156\text{XT6} +0.101\text{XT7}$$

In this transformed full model, the F-value has been improved to 17.18 and p-value (F>17.18) is less than 0.0001, which suggests at least one independent variable is significant for the prediction of response variable YT. R^2 and adjusted R^2 are 14.92% and 14.05%, respectively, which means 14.92% of the total variability in the response variable is explained by the regression model. In addition, there are three significant independent variables XT3, XT4 and XT7 in the transformed full model containing other variables.

For checking the multicollinearity among the independent variables, there is no multicollinearity in the untransformed full model with scaled data only. Also There is no multicollinearity in the transformed full model and chosen model with three significant independent variables in the scaled and centered data.

During model selection procedure, we split the data into two parts n_1=602 and n_2=92. Through model selection criteria, the chosen model consists of variables XT3, XT4, XT5, XT6 and XT7 at the significant level $\alpha = 0.1$. We fit this chose model with regression in the n_1 data set and find out the global F-test value is 21.08 and p-value (F>21.08) is less than 0.0001. In the partial t-tests, all independent variables XT3, XT4, XT5 and XT7 are significant except XT6. R^2 has been improved to be 15.03% as well, meaning 15.03% of the total variability in the response is explained by the regression model. We then use n_2 data set to validate the chosen model. The validation is successful because there is positive correlation between observed values and predicted values, which means the chosen model is appropriate for the prediction. Then we want to know whether there exists high leverage points, outliers or high influence points. During residual analysis and influence diagnostics, we find out there are few high leverage points and outliers but no observations affecting the model prediction.

Finally, we fit the best model in 694 observations:

$$\hat{Y}^{0.01}=0.93065 +119.29\text{XT3} -2.878\text{E-17XT4} +1.22\text{XT5} +0.00000154\text{XT6} +0.103\text{XT7}$$

This mode explains 14.71% of the total variation in the response variable YT=$Y^{0.01}$. The transformed independent variables are difficult to interpret for the prediction of response variable. After consulting with Miss Karry who provides this data file, she considers this final regression model is reasonable for the prediction of the error median difference since she is working with her professor and they get the result that R^2 is about 10%. In addition, the uncertainty word (X_1) does not work according to Miss Karry's research, which makes sense that the final best model also excludes XT1 variable.

6 Appendix