CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Analysis of Audit Data

AUTHOR: Yang Zhou

SUPERVISOR: Dr. Sanjoy K. Sinha

DATE: May 04, 2019

Abstract:

This Honors project is a study of logistic regression modeling based on the external corporate audits from tax year 2009. As more tax administrations go digital, they seek to crack down on evasion and fraud, to collect it more efficiently. As Canada and other countries receive digital tax compliance-related data from an increasing number of sources, it is important to identify observable factors that help to predict tax compliance risk in the population.

By generating useful statistical models, we can improve the way we derive business intelligence to better detect high risk audit files, to reduce assessing time and improve the quality and efficiency of tax assessment results. The study is based on 766 completed records from nine industry sectors. The logistic regression model uses three predictors: Para A value, Para B value, Money value, and Industry Sectors. The correlations between these three predictors show that three predictors are weakly correlated; they are considered as statistically independent.

The results of maximum likelihood estimation (MLE) of the parameters in the regression modeling are significant, while the result of Lack-Of-Fit test is not. As the amount and type of electronic data to evaluate becomes more complex, statistical methodologies are becoming more and more important for decision making. Use of logistic regression modeling to understand the impact of different variables will improve the quality and accuracy of selected high-risk audit files.

Contents:

List of Tables and Figures

1 Introduction:

This Honors Project focuses on building logistic regression models for risk assessment and introduces some practice for detecting potential tax fraud. Canada's tax system is built on the principle of self-assessment, where individuals and businesses calculate and submit taxes according to the *Income Tax Act* and the *Excise Act*. Fraud, or the deliberate misrepresentation of fact, is one of the most common means taken by those who wish to hide taxable income from tax authorities. Underground economic activity in Canada totalled $51.6 billion in 2016, or 2.5% of gross domestic product (GDP).

Underground economy activity creates unfair competition for those who meet their tax responsibilities by allowing businesses who cheat to offer lower prices for similar goods and services. Use of statistical analytic measurements for risk assessment is helpful for detecting potential fraud, and it could be an exciting tool for auditors or accountants to use to enhance their detection of tax compliance risk.

Within North America, companies are categorized according to the North American Industry Classification System (NAICS). NAICS was created during the North America Free Trade Agreement, by the official statistical departments of the United States, Canada (Statistic Canada), and Mexico. With different classifications, companies might show different characteristics, so classification methods are important for detecting potential fraud.

In this research, I use a non-confidential audit dataset which has been collected by Auditor Office of India between tax year 2015-2016. The dataset contains 9 classifications of companies which include Agriculture, Irrigation, Animal Husbandry, Buildings and Roads, Public Health, Forest, Corporate, Fisheries, and Industries.

The cleaned dataset has 766 observations, two explanatory variables, one classification variable, and one outcome variable. The two explanatory variables are Para A value and Para B value, classification variable is Sector Score, and outcome variable is a binary variable: Risk. The explanation of selected variables is shown below:

- Para A value: Discrepancy found in the planned-expenditure of inspection and summary report-A in risks
- Para B value: Discrepancy found in the unplanned-expenditure of inspection and summary report-B in risks
- Sector Score: classification variable which contains 9 sectors
- Risk: binary variable with 0 (non-risk audit file) and 1 (at-risk audit file)
- Money Value: Amount of money involved in misstatements in the past audits
- History: Average historical loss suffered by firm in the last 10 years
- Number: Historical discrepancy score
- District Score: Historical risk score of a district in the last 10 years
- Loss: Amount of loss suffered by the firm last year

2 Methodologies:

2.1 Odds Ratio:

The odds is the ratio between the probability of 'success (1)' and 'failure (0)', and it describes if the probability of success has advantage when against the probability of failure.

$\text{Odds1} = \frac{\pi_{1|1}}{1-\pi_{1|1}}$, where $\pi_{1|1}$is the probability of success of s1

$\text{Odds2} = \frac{\pi_{1|2}}{1-\pi_{1|2}}$, where $\pi_{1|2}$ is the probability of success of s2

To interpret Odds1 and Odds2 within different sectors, we need to separate results into three parts:

- If Odd = 1, the probability of success and failure are equal
- If Odd < 1, the probability of success is smaller than the probability of failure
- If Odd > 1, the probability of success is greater than the probability of failure

The Odds Ratio is the ratio between two odds:

$$\theta = \frac{Odd\underline{S1}}{Odd\underline{S2}} = \frac{\frac{\pi_{1|1}}{1-\pi_{1|1}}}{\frac{\pi_{1|2}}{1-\pi_{1|2}}} = \frac{\pi_{1|1}/\pi_{0|1}}{\pi_{1|2}/\pi_{0|2}} = \frac{\pi_{1|1}\pi_{0|2}}{\pi_{0|1}\pi_{1|2}}$$

The odds ratio between sectors shows the relationship between odds, which may be interpreted as:

- If θ = 1, the Odds for S1 (sector 1) is equal with S2 (sector 2)
- If θ > 1, the Odds for S1 (sector 1) is greater than S2 (sector 2)
- If θ < 1, the Odds for S1 (sector 1) is smaller than S2 (sector 2)

The odds ratio may be estimated as:

$\hat{\theta} = \frac{n_{11}}{n_{12}}\frac{n_{02}}{n_{01}}$, where

n_{11} is the number of success for S1 (sector 1)

n_{01} is the number of failures for S1 (sector 1)

n_{12} is the number of success for S2 (sector 2)

n_{02} is the number of failures for S2 (sector 2)

The variance of $\log\hat{\theta}$ may be estimated as,

$$\hat{V}\left(\log\hat{\theta}\right) = \frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{12}} + \frac{1}{n_{02}}$$

Then a (1-α)100% confidence Interval for $\log\theta$ may be obtained as,

$\log\hat{\theta} \pm Z_{\alpha/2}\sqrt{\hat{V}\left(\log\hat{\theta}\right)}$, where α is significance level

To interpret the confidence intervals for $\log\theta$, we need to interpret these with different levels of significance:

- For α = 0.05, $Z_{\propto/2} = Z_{0.05/2} = Z_{0.025}$, and we have 95% confidence that the CI contains the true value of $\log\theta$
- For α = 0.1, $Z_{\propto/2} = Z_{0.1/2} = Z_{0.05}$, and we have 90% confidence that the CI contains the true value of $\log\theta$
- For α = 0.01, $Z_{\propto/2} = Z_{0.01/2} = Z_{0.005}$, and we have 99% confidence that the CI contains the true value of $\log\theta$

The smaller the significance level, the more robust the result will be. The most commonly recommended significance level is 0.05.

2.2 Correlation Test (Test of Independence):

To test if the predictors are statistically independent, we need to calculate their pairwise correlations and run hypothesis test to compare it with null hypothesis which is usually zero.

The equation of correlation coefficient:

$$r_{xy} = \frac{\sum x_i y_i - n\overline{xy}}{n\sqrt{Var(x)Var(y)}}, \text{where } Var(x) = \sum (x_i - \overline{x})^2 \ \ and \ Var(y) = \sum (y_i - \overline{y})^2$$

For hypothesis test procedure:

With the sample size over 30, we will use the critical value from Z-table instead of T-table.

First, we need to assign a null hypothesis for parameters as:

$$H_0: \rho_{xy} = 0, which\ indicates\ there\ is\ no\ linear\ relationship\ between\ x\ and\ y$$

Second, we need to assign an alternative hypothesis:

$$H_a: \rho_{xy} \neq 0, which\ indicates\ there\ is\ linear\ relationship\ between\ x\ and\ y$$

Third, we need to establish the test statistics and calculate the P-value for testing if there is any significant findings.

Test Statistics:

$$Z_{cal} = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xy}^2}{n-2}}} \sim Z_{1-\propto/2} or\ Z_{\propto/2}, where\ \propto is\ the\ level\ of\ significance$$

We will use P-value to determine if there are any significant findings, and P-value is calculated by

$$P - Value = P(|Z| \geq Z_0), with\ Z_0 = |Z_{cal}|$$

$$P - Value \begin{cases} \geq \propto, this\ means\ the\ testing\ result\ is\ insignificant\ and\ x, y\ are\ independent \\ < \propto, this\ means\ the\ testing\ result\ is\ significant\ and\ x, y\ are\ dependent \end{cases}$$

2.3 Logistic Regression Model (with three explanatory variables):

The logistic regression is a modeling technique for describing a binary response variable, and this method is used to estimate the probability of success π based on the logistic model.

$$\text{LOGIT}(\pi) = \text{Log}\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where $x_1, x_2, and\ x_3$ are continuous predictor variables.

In addition, the model can be written in another form:

$$\pi = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2++\beta_3 x_3)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2++\beta_3 x_3)}.$$

The probability of success π may be estimated from:

$$\text{LOGIT}(\hat{\pi}) = \text{Log}\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Here $\underline{\hat{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]$ is the vector of parameter estimates of logistic regression model. For estimating the parameters, we use the maximum likelihood estimators (MLE) for β, and they are obtained by maximizing the likelihood

$$L\left(\underline{\beta}\right) = \prod_{i=1}^{n} \pi_i^{m_i} (1-\pi_i)^{n_i-m_i},$$

with respect to β. This is equivalent to solving the score equations $\frac{\partial(l(\beta))}{\partial\beta} = 0$ with respect to β.

The matrix of covariance of the MLE<u>s</u> is: $\hat{V}(\hat{\beta}) = \left[-l''(\hat{\beta})\right]^{-1}$

For estimating a (1-α)100% confidence interval of the regression coefficient β_j, we need to use both the estimate and its variance, as the confidence interval for β_j is given by

$\widehat{\beta_J} \pm Z_{\alpha/2}\sqrt{\hat{V}(\widehat{\beta_J})}$, where $\hat{V}(\widehat{\beta_J})$ is j^{th} diagonal element of $\hat{V}(\hat{\beta})$

2.4 Logistic Regression Model (with three explanatory variables and one classification variable):

For the classification variable, we basically set each classification as a binary indicator variable. If the classification variable has m levels, then we set m-1 indicator variables, as:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$ indicates the values of a classification variable with 4 levels

So, in Statistical Analytic Software (SAS) program, the *Proc Logistic* commend will automatically set the last level into 0.

Recall the last section, our logistic model will become:

$$\text{LOGIT}(\pi) = \text{Log}\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3$$

- $z_1 = \begin{cases} 1, & if\ the\ response\ corresponds\ to\ the\ first\ level\ within\ classification \\ 0, & otherwise \end{cases}$
- $z_2 = \begin{cases} 1, & if\ the\ response\ corresponds\ to\ the\ second\ level\ within\ classification \\ 0, & otherwise \end{cases}$
- $z_3 = \begin{cases} 1, & if\ the\ response\ corresponds\ to\ the\ third\ level\ within\ classification \\ 0, & otherwise \end{cases}$

The binary response variable y is defined as

- $y = \begin{cases} 1, & if\ the\ audit\ file\ is\ at\ risk \\ 0, & if\ the\ audit\ file\ is\ not\ at\ risk \end{cases}$

The analysis procedure will be the same as discussed earlier.

We use Risk Assessment Algorithm to decide if the audit is in risk or not. The Audit Risk (AR) can be explained with two conditions:

$$AR_{ave} = \begin{cases} > 1, Lable\ it\ as\ Fraud\ (y = 1) \\ \leq 1, label\ it\ as\ No\ Fraud\ (y = 0) \end{cases}$$

And we have

$$AR = Para_A * W_{Para_a} + Para_B * W_{Para_B} + Number * W_{num} + MV * W_{MV} + Sector_{Score} * W_{SS} + History * W_{His} + District * W_{Dis} + Loss * W_L$$

Where W_i is weighted vector of those variables

2.5 Wald Test in Logistic Regression Modeling:

The Wald Test is a parametric statistical test which was found by statistician Abraham Wald. This test is detecting whether the resulting value of the vector for parameters is equal to some postulated value (Null Hypothesis) or not. The Wald Test is described below.

Consider the null hypothesis:

$$H_0: \underline{\beta} = \underline{\beta}_0\ , where\ \underline{\beta}_0\ is\ a\ vector\ of\ parameters\ with\ initial\ guess$$

Normally, we set the original guess to the vector of zeros.

The alternative hypothesis may be one or two-sided:

$$H_a: \underline{\beta} \begin{cases} < \underline{\beta}_0 \\ \neq \underline{\beta}_0 \\ > \underline{\beta}_0 \end{cases}$$

We calculate the test statistics:

$$X_{cal}^2 = \left(\underline{\hat{\beta}} - \underline{\beta_0}\right) \widehat{V^{-1}} \left(\underline{\hat{\beta}}\right) (\underline{\hat{\beta}} - \underline{\beta_0})^{-1}$$

Under H_0 , X_{cal}^2 follows an approximate Chi-Square distribution with P degrees of freedom, which P is the dimension of β.

Test Results with α=0.05:

$$P - Value \begin{cases} \geq \propto, this\ means\ the\ testing\ result\ is\ not\ significant \\ < \propto, this\ means\ the\ testing\ result\ is\ significanct \end{cases}$$

If the Wald Test result is not significant, then all parameters, which we are using in the modeling, are not having significant impacts when we generate regression model.

If the Wald Test result is significant, then one or more parameters are having significant impact in the model.

2.6 Likelihood Ratio Test Between Models:

Within this research project, Likelihood ratio tests (LRTs) will be used to compare two logistic models. The original model will only contain the continuous variables, and the alternative model will contain not only the continuous variables but also categorical variable. This test will be testing if the categorical variable is significant or not. Within the methodology part, we will be using this method to test if the categorical variable is significant with modeling. The following part will introduce the procedure of testing.

Recall the MLE result in logistic regression modeling, where we have:

$$L\left(\underline{\beta}\right) = \prod_{i=1}^{n} \pi_i^{m_i} (1 - \pi_i)^{n_i - m_i}$$

Using above equation to drive two different Maximum Likelihoods for both original guess and alternatives, such as $L_1(\underline{\hat{\beta}})$ and $L_2(\underline{\hat{\beta}})$, we define the likelihood ratio as

$$LR = -2\log_e\left(\frac{L\left(\underline{\widehat{\beta_1}}\right)}{L\left(\underline{\widehat{\beta_2}}\right)}\right) = -2(\log_e(L_1) - \log_e(L_2)) = 2\log_e(L_2) - 2\log_e(L_1)$$

We use LR ~ $X_{p-1,1-\propto}^2$, where p is number of Levels of the categorical variable.

Thus the LR indicates the difference between two model, and we know that the LRT can be computed as a difference in the deviance for the two models. To show the difference of the two models, hypothesis testing is necessary to test if there are any significant findings.

Within the hypothesis test procedure, LRs are used to compare with the critical value such as $X_{p-1,1-\propto}^2$. There will be two scenarios for the comparison results:

- If $LR > X_{p-1,1-\propto}^2$, the results are significant, and two models are having significant difference in the deviance.
- If $LR \leq X_{p-1,1-\propto}^2$, the results are non-significant, and two models are not having significant difference in the deviance.

2.7 Logistic Regression Model (With Single Sector):

The purpose of building separate models for each sector is to find if the coefficients of the explanatory variables are different for different sectors. The modeling procedure is similar to that in the third section, and the only difference in this part is using one sector of records from the dataset instead of the whole data.

The logistic regression model function represents as:

$$\text{LOGIT}(\pi_i) = \text{Log}\left(\frac{\pi_{i|m}}{1-\pi_{i|m}}\right) = L_i = \beta_{0m} + \beta_{1m}x_1 + \beta_{2m}x_2 + \beta_{3m}x_3,$$

$x_1, x_2, and\ x_3$ are continuous random variables for m^{th} sector

$$\text{LOGIT}(\pi_j) = \text{Log}\left(\frac{\pi_{j|n}}{1-\pi_{j|n}}\right) = L_j = \beta_{0n} + \beta_{1n}x_1 + \beta_{2n}x_2 + \beta_{3n}x_3,$$

$x_1, x_2, and\ x_3$ are continuous random variables for n^{th} sector

Based on the MLE estimations from one sector-based regression model, we will be able to set hypothesis tests to see if there are any significant differences. The comparison procedure will be showing on the next part.

2.8 Method of Comparison between Parameters with Different Sectors:

To compare if there is significant difference between the coefficient of β_i in sector m and the coefficient of β_i in sector n at level α, we need to find both estimations and its variations. As resulting of pairwise comparison test, it will be also based on the result of Confidence Interval of:

$\beta_{im} - \beta_{in}$, where i indicates the parameter, $m\ and\ n$ indicate the sectors

To set the procedure of comparison test, we will set the null hypothesis $H_0: \beta_{im} - \beta_{in} = 0$ against the alternatives $H_a: \beta_{im} - \beta_{in} \begin{cases} > 0 \\ \neq 0 \\ < 0 \end{cases}$

We can find the confidence interval for $(\beta_{im} - \beta_{in})$ as

$(\beta_{im} - \beta_{in}) \pm t\alpha/_2\sqrt{\frac{s_{im}^2}{n_m} + \frac{s_{in}^2}{n_n}}$, where t-value has $n_m + n_n - 2$ Degree of Freedom (DF)

To interpret the result of the hypothesis test, we will separate into three parts:

- If the resulting Confidence Interval contains 0 at α significant levels, then there is no significant difference between β_{im} and β_{in}.
- If both upper bound and lower bound are above 0 at α significant levels, then there is positive significant difference between β_{im} and β_{in}.
- If both upper bound and lower bound are below 0 at α significant levels, then there is negative significant difference between β_{im} and β_{in}.

3 Project Data Analysis:

Table 1 below presents summary statistics of fraudulent rate for each sector.

3.1 Analysis for Table:

Sector_score	Sector_Name	0 (Non-risk Audit File)		1 (Risk Audit File)		Total
		N	%	N	%	N
55.57	Agriculture	174	87.0%	26	13.0%	200
3.89	Irrigation	48	42.1%	66	57.9%	114
1.85	Animal Husbandry	54	56.8%	41	43.2%	95
2.72	Buildings and Roads	29	35.4%	53	64.6%	82
3.41	Public Health	17	22.4%	59	77.6%	76
2.37	Forest	38	51.4%	36	48.6%	74
1.99	Corporate	34	72.3%	13	27.7%	47
21.61	Fisheries	39	95.1%	2	4.9%	41
59.85	Industries	34	91.9%	3	8.1%	37
Total		467	61.0%	299	39.0%	766

Table 1: Summary Statistics of Audit Data for Different Sectors

Based on the table above, Agriculture sector has the most numbers in our dataset, and has relatively high non-risk proportion (87%). Last two sectors, which are Fisheries and Industries, have the highest proportion of non-risk, respectively with 95.1% and 91.9%.

Public Health and Buildings and Roads respectively have 76 and 82 audit files. Those two sectors are having the highest proportions of risking, respectively with 77.6% and 64.6%.

Among all sectors under study, Public Health has the highest risk with a fraudulent rate of 77.6%.

3.2 Logistic Regression Modeling with a classification variable:

Correlation Matrix (Continuous Variables):

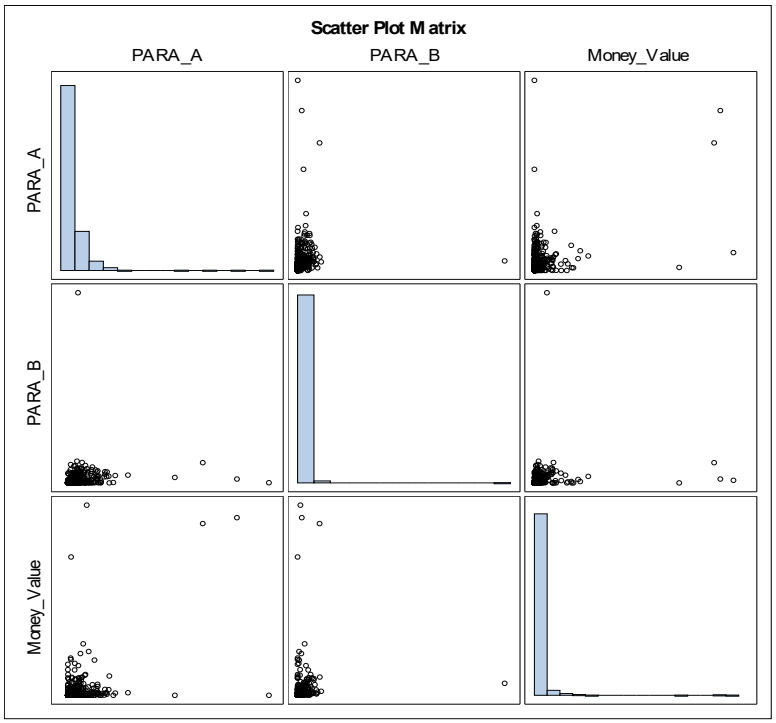

Figure 1: ScatterPlot and histogram for Para_A, Para_B, and Money_Value

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	PARA_A	**PARA_B**	**Money_Value**
PARA_A **PARA_A**	1.00000 766	0.16356 <.0001 766	0.43931 <.0001 766
PARA_B **PARA_B**	0.16356 <.0001 766	1.00000 766	0.13090 0.0003 766
Money_Value **Money_Value**	0.43931 <.0001 766	0.13090 0.0003 766	1.00000 766

Table 2: correlation Test

The correlation tests between continuous variables (Para_A, Para_B, and Money_Value) are significant. These significant results indicate that there is enough evidence to show that the correlations of (Para_A, Para_B), (Para_B, Money_Value), and (Para_A, Money_Value) are not zero. As the result shown, we have:

- Corr(Para_A, Para_B) = 0.16356 with P-value < 0.0001
- Corr(Para_A, Money_Value) = 0.43931 with P-value < 0.0001
- Corr(Para_B, Money_Value) = 0.1309 with P-value = 0.0003

However, as Scatter plot matrix shown, all paired plots are not showing any linear relations. So, with the following analysis of logistic regression modeling, we will still be using all three variables even all correlations are showing statistically significance as their results.

Logistic Regression Model Results (With Sector Variable):

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1025.764	358.360
SC	1030.404	414.039
-2 Log L	1023.764	334.360

Table 3: Likelihood Under different models

The above SAS Output shows $-2 * Log(Likelihood) = 334.360$, which will be used for running a likelihood ratio test.

Estimated Covariance Matrix												
Parameter	Intercept	Sector_Agriculture	Sector_Animal_Husbandry	Sector_Buildings_and_Roads	Sector_Corporate	Sector_Fisheries	Sector_Forest	Sector_Industries	Sector_Irrigation	PARA_A	PARA_B	Money_Value
Intercept	0.133615	-0.09787	-0.08747	-0.03649	-0.07261	0.474058	-0.09688	0.020137	-0.0726	-0.01021	-0.00635	-0.00877
Sector_Agriculture	-0.09787	0.175133	0.075332	0.040913	0.046973	-0.45273	0.078959	-0.04555	0.071391	0.004883	-0.00617	0.00693
Sector_Animal_Husbandry	-0.08747	0.075332	0.189199	0.040118	0.0578	-0.46575	0.078722	-0.05372	0.052288	-0.01421	0.006277	0.006733
Sector_Buildings_and_Roads	-0.03649	0.040913	0.040118	0.422784	0.008628	-0.46696	0.031763	-0.07614	0.029131	-0.00297	-0.00597	0.000744
Sector_Corporate	-0.07261	0.046973	0.0578	0.008628	0.308187	-0.47286	0.053357	-0.0345	0.034863	0.000778	0.008427	-0.00068
Sector_Fisheries	0.474058	-0.45273	-0.46575	-0.46696	-0.47286	3.811062	-0.46297	-0.52334	-0.44441	-0.00159	-0.01153	-0.0123
Sector_Forest	-0.09688	0.078959	0.078722	0.031763	0.053357	-0.46297	0.184871	-0.03753	0.059698	0.002226	0.001964	0.005015
Sector_Industries	0.020137	-0.04555	-0.05372	-0.07614	-0.0345	-0.52334	-0.03753	0.906876	-0.04302	0.00865	0.007491	-0.01041
Sector_Irrigation	-0.0726	0.071391	0.052288	0.029131	0.034863	-0.44441	0.059698	-0.04302	0.247437	0.005146	-0.00703	0.00119
PARA_A	-0.01021	0.004883	-0.01421	-0.00297	0.000778	-0.00159	0.002226	0.00865	0.005146	0.008353	-0.00125	0.000149
PARA_B	-0.00635	-0.00617	0.006277	-0.00597	0.008427	-0.01153	0.001964	0.007491	-0.00703	-0.00125	0.007252	-0.00031
Money_Value	-0.00877	0.00693	0.006733	0.000744	-0.00068	-0.0123	0.005015	-0.01041	0.00119	0.000149	-0.00031	0.003964

Table 4: Variance-Covariance Matrix of Parameter estimates

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	689.4034	11	<.0001
Score	258.7593	11	<.0001
Wald	122.9344	11	<.0001

Table 5: Global Tests Results

At this part, the SAS program showed the Wald test result, and based on the formula which showed on methodology part, the wald score is calculated based on the variance-covariance matrix. The Wald result shows there are enough evidence to show that one or more parameter coefficients of parameters are non-zero.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Sector_Name	8	19.6453	0.0118
PARA_A	1	26.6203	<.0001
PARA_B	1	29.7701	<.0001
Money_Value	1	37.5310	<.0001

Table 6: Factor Effects

From the above SAS output, we see significant results separately of each factor, and all factors are showing significant effect in the model.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.5229	0.3654	92.9532	<.0001
Sector_Name	Agriculture	1	0.0922	0.4184	0.0486	0.8255
Sector_Name	Animal Husbandry	1	1.1306	0.4348	6.7597	0.0093
Sector_Name	Buildings and Roads	1	-0.3467	0.6500	0.2845	0.5937
Sector_Name	Corporate	1	0.5792	0.5550	1.0888	0.2967
Sector_Name	Fisheries	1	-2.9056	1.9508	2.2183	0.1364
Sector_Name	Forest	1	1.5497	0.4298	12.9987	0.0003
Sector_Name	Industries	1	-0.7395	0.9522	0.6032	0.4374
Sector_Name	Irrigation	1	0.1297	0.4973	0.0680	0.7943
PARA_A		1	0.4714	0.0914	26.6203	<.0001
PARA_B		1	0.4645	0.0851	29.7701	<.0001
Money_Value		1	0.3857	0.0630	37.5310	<.0001

Table 7: Maximum likelihood estimates for audit data analysis

Based on the MLE results from SAS Outputs, Intercept, Animal Husbandry sector, Forest sector, ParaA, ParaB, and Money_Value have significant Coefficients as shown in the results.

For Intercept: the estimate is -3.5229 with P-Value < 0.0001, which is a significant result. It means when other parameters are zero (indicates the last sector Public Health), the odds for audit files been at risk are quite low with:

$$\text{OddInt} = \exp(-3.5229) = 0.03$$

For the Non-significant Sectors (Agriculture, Buildings and Roads, Corporate, Fisheries, Industries, and Irrigation sectors): the estimations of the indicator fields with above classifications are resulting as Non-significant. It means the changes of Intercept part will not have significant difference between Non-Significant sectors and Public Health sector.

For the Significant Sectors (Animal Husbandry, and Forest sectors): the estimates of the indicator fields with above sectors are resulting as highly significant. It means the changes of intercept part will have significant difference between those sectors and Public Health sector.

For Para_A: Para_A indicates the Discrepancy found in the planned-expenditure of inspection. The result is showing significance with 0.4714. The value indicates that for every unit change in Para_A, the logit will change 0.4714 unit where $logit = \ln({}^{p}/_{1-p})$.

For Para_B: Para_B indicates the Discrepancy found in the unplanned-expenditure of inspection. The result is showing significance with 0.4645. The value indicates that for every unit change in Para_B, the logit will change 0. 4645 unit.

For Money_Value: Money_Value indicates the amount of money involved in misstatements in the past audits. The result is showing significance with 0.3857. The value shows that for every unit of change in Money Value, the logit will change 0.3857 unit.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Sector_Name Agriculture vs Public Health	0.658	0.119	3.644
Sector_Name Animal Husbandry vs Public Health	1.859	0.343	10.074
Sector_Name Buildings and Roads vs Public Health	0.424	0.055	3.282
Sector_Name Corporate vs Public Health	1.071	0.164	7.014
Sector_Name Fisheries vs Public Health	0.033	<0.001	3.195
Sector_Name Forest vs Public Health	2.827	0.509	15.694
Sector_Name Industries vs Public Health	0.287	0.022	3.687
Sector_Name Irrigation vs Public Health	0.683	0.110	4.244
PARA_A	1.602	1.340	1.917
PARA_B	1.591	1.347	1.880
Money_Value	1.471	1.300	1.664

Table 8: Odds Ratio Results

As above Odds Ratio Analysis from SAS Outputs, it does not have significance finding when comparing first eight sectors with the last sector (Public Health).

For Para_A, Para_B, and Money_Value: the outputs are shown as significance, and both of with upper-bounds and lower-bound are greater than 1. The results are showing that:

- For each unit of increase in Para_A, the odds of risk is 1.602 times than before. The 95% confidence interval is (1.34,1.917)
- For each unit of increase in Para_B, the odds of risk is 1.591 times than before. The 95% confidence interval is (1.347, 1.88)
- For each unit of increase in Money_Value, the odds of risk is 1.471 times than before. The 95% confidence interval is (1.3, 1.664)

Partition for the Hosmer and Lemeshow Test					
		Risk = 1		Risk = 0	
Group	Total	Observed	Expected	Observed	Expected
1	77	0	0.80	77	76.20
2	77	4	2.51	73	74.49
3	77	1	3.43	76	73.57
4	77	6	4.64	71	72.36
5	77	13	7.55	64	69.45
6	77	14	15.45	63	61.55
7	77	35	41.28	42	35.72
8	77	77	74.33	0	2.67
9	20	20	20.00	0	0.00
10	129	129	129.00	0	0.00

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
13.3037	8	0.1018

Table 9: Goodness-of-Fit Result

For Goodness-Of-Fit test, the SAS automatically set into 10 boxes sorting by Continuous Variables (Para_A, Para_B, and Money_Value). We use Chi-Square tests to compare observed value and expected value with 10-2=8 degree of freedom. The outputs are showing non-significance corresponding with P-value = 0.1018. The results represented the regression model is not rejecting the null hypothesis, and the model is fit.

3.3 Logistic Regression Model Results (Without Categorical Variable):

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1026.754	366.513
SC	1031.395	385.078
-2 Log L	1024.754	358.513

Table 10: Likelihood score (W/O sectors)

The above SAS Output calculated $-2 * Log(Likelihood) = 358.513$, which will be used for running likelihood ratio test.

Estimated Covariance Matrix				
Parameter	Intercept	PARA_A	PARA_B	Money_Value
Intercept	0.045299	-0.01016	-0.00568	-0.00436
PARA_A	-0.01016	0.006862	-0.00023	0.000449
PARA_B	-0.00568	-0.00023	0.005647	-0.00018
Money_Value	-0.00436	0.000449	-0.00018	0.003506

Table 11: Variance-Covariance Matrix (W/O sectors)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	666.2414	3	<.0001
Score	144.5256	3	<.0001
Wald	121.5779	3	<.0001

Table 12: Global tests (W/O sectors)

At this part, the SAS program showed the Wald test result, it says there are enough evidence to show that one or more parameter coefficients are non-zero.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.1280	0.2128	215.9909	<.0001
PARA_A	1	0.5980	0.0828	52.1224	<.0001
PARA_B	1	0.4144	0.0751	30.4100	<.0001
Money_Value	1	0.3793	0.0592	41.0358	<.0001

Table 13: MLE Parameter Coefficients Results (W/O sectors)

Based on the MLE results from SAS Outputs, Intercept, ParaA, ParaB, and Money_Value have significant coefficients as shown in the results.

For Intercept: the estimate is -3.128 with P-Value < 0.0001, which is a significant result. It means when other parameters are zero (indicates the last sector Public Health), the odds for audit files been at risk are quite low with:

$$\text{OddInt} = \exp(-3.1280) = 0.0438.$$

For Para_A: Para_A indicates the Discrepancy found in the planned-expenditure of inspection. The result is showing significance with 0.5980. The value indicates that for every unit change in Para_A, the logit will change 0.5980 unit where $logit = \ln(^{p}/_{1-p})$.

For Para_B: Para_B indicates the Discrepancy found in the unplanned-expenditure of inspection. The result is showing significance with 0.4144. The value indicates that for every unit change in Para_B, the logit will change 0. 4144 unit.

For Money_Value: Money_Value indicates the amount of money involved in misstatements in the past audits. The result is showing significance with 0.3793. The value shows that for every unit of change in Money Value, the logit will change 0.3793 unit.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
PARA_A	1.819	1.546	2.139
PARA_B	1.513	1.306	1.754
Money_Value	1.461	1.301	1.641

Table 14: Odds Ratio Results (W/O sectors)

As above Odds Ratio Analysis from SAS Outputs, it does not have significance finding when comparing first eight sectors with the last sector (Public Health).

For Para_A, Para_B, and Money_Value: the outputs are shown as significance, and both of with upper-bounds and lower-bound are greater than 1. The results are showing that:

- For each unit of increase in Para_A, the odd of risk is 1.819 times than before. The 95% confidence interval is (1.546, 2.139)
- For each unit of increase in Para_B, the odd of risk is 1.513 times than before. The 95% confidence interval is (1.306, 1.754)
- For each unit of increase in Money_Value, the odd of risk is 1.461 times than before. The 95% confidence interval is (1.301, 1.641)

Partition for the Hosmer and Lemeshow Test					
		Risk = 1		Risk = 0	
Group	Total	Observed	Expected	Observed	Expected
1	77	6	3.25	71	73.75
2	77	4	3.80	73	73.20
3	77	7	4.70	70	72.30
4	77	7	5.68	70	71.32
5	77	5	7.12	72	69.88
6	77	12	11.40	65	65.60
7	77	31	38.74	46	38.26
8	77	77	74.31	0	2.69
9	32	32	32.00	0	0.00
10	118	118	118.00	0	0.00

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.6043	8	0.2251

Table 15: Goodness-of-Fit Result (W/O sectors)

For Goodness-Of-Fit test, the SAS automatically set into 10 boxes sorting by Continuous Variables (Para_A, Para_B, and Money_Value). Using Chi-Square tests to compare observed value and expected value with 10-2=8 degree of freedom. The outputs are showing non-significance corresponding with P-value = 0.2251. The results represented the regression model is not rejecting the null hypothesis, and the model is fit.

3.4 Likelihood Ratio Test (Compare Two Models):

For comparing two models, we have found Likelihood estimator for both models with categorical variable and without categorical variable in SAS outputs.

- $-2 * Log(L_2) = 334.360$, which is alternative guess with categorical variable
- $-2 * Log(L_1) = 358.513$, which is original guess without categorical variable

Using above two results, we have $LR = 2 * \log_e(L_2) - 2 * \log_e(L_1) = 358.513 - 334.360 = 24.153$, and we use the calculated result to compare with critical value as $X^2_{p-q,1-\propto} = X^2_{8,0.95} = 2.73$.

We have $LR = 24.153 > 2.73 = X^2_{8,0.95}$ which is rejecting the null hypothesis, so the result is significant.

In conclusion, two models are having significant difference in the deviance, and categorical variable (sectors variable) is affecting the model significantly.

4 Project Modeling by Sectors:

4.1 Logistic Regression Modeling (Agriculture Sector):

Outputs and Interpretations:

This modeling is based on single sector Agriculture, and there are 200 samples used in this regression modeling.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.7619	0.5054	55.3962	<.0001
PARA_A	1	0.4337	0.4614	0.8836	0.3472
PARA_B	1	0.6006	0.1570	14.6374	0.0001
Money_Value	1	0.5802	0.2875	4.0726	0.0436

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PARA_A	1.543	0.625	3.811
PARA_B	1.823	1.340	2.480
Money_Value	1.786	1.017	3.138

Table 16: Agriculture Sector MLE estimates

MLE Result:

For Intercept: the estimate is -3.7619 with P-Value < 0.0001, which is a significant result. It means that when other parameters are zero, the odds for audit files been at risk are quite low with:

$$\text{OddInt} = \exp(-3.7619) = 0.0232.$$

For Para_A: Para_A indicates the Discrepancy found in the planned-expenditure of inspection. The result is showing non-significance with parameter coefficient as 0.4337, and its p-value=0.3472.

For Para_B: Para_B indicates the Discrepancy found in the unplanned-expenditure of inspection. The result is showing significance with parameter coefficient as 0.6006, and its p-value=0.0001. The result indicates that for every unit change in Para_B, the logit will change 0.5902 unit.

For Money_Value: Money_Value indicates the amount of money involved in misstatements in the past audits. The result is showing significance with parameter coefficient as 0.5802, and its p-value=0.0436. The result indicates that for every unit change in Para_B, the logit will change 0.5802 unit.

Odds Ratio Results:

For Para_A: the result shows insignificance with the lower bound is less than 1. The result is showing that for each unit of increasing in Para_A, the odds of risk is not going to change significantly.

For Para_B: the result shows strongly significance with both lower and upper bounds are greater than 1. The result is showing that for each unit of increasing in Para_B, the odds of risk is 1.823 more times more than before.

For Money_Value: the result shows significance both lower and upper bounds are greater than 1. The result is showing that for each unit of increasing in Money_Value, the odds of risk is 1.786 times more than before.

Goodness-Of-Fit Test:

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
12.6039	8	0.1262

Table 17: Agriculture Sector Goodness-of-Fit test result

For Goodness-Of-Fit test, the SAS automatically set into 10 boxes sorting by Continuous Variables (Para_A, Para_B, and Money_value). We use Chi-Square tests to compare observed value and expected value with 10-2=8 degree of freedom. The result is non-significant which means the model is fitted.

4.2 Logistic Regression Modeling (Irrigation Sector):

Outputs and Interpretations:

This modeling is based on single sector Irrigation, and there are 114 samples used in this regression modeling.

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-3.0066	0.6097	24.3154	<.0001
PARA_A	1	0.5283	0.2586	4.1726	0.0411
PARA_B	1	0.3402	0.1564	4.7294	0.0297
Money_Value	1	0.3006	0.1243	5.8519	0.0156

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
PARA_A	1.696	1.022	2.815
PARA_B	1.405	1.034	1.909
Money_Value	1.351	1.059	1.723

Table 18: Irrigation Sector MLE result

MLE Result:

For Intercept: the estimate is -3.0066 with P-Value < 0.0001, which is a significant result. It means that when other parameters are zero, the odds for audit files been at risk are quite low with:

$$\text{OddInt} = \exp(-3.0066) = 0.0495.$$

For Para_A: Para_A indicates the Discrepancy found in the planned-expenditure of inspection. The result is showing significance with 0.5283 if we set significant level into 0.05. The value indicates that for every unit change in Para_A, the logit will change 0.5283 unit where $logit = \ln(^{p}/_{1-p})$.

For Para_B: Para_B indicates the Discrepancy found in the unplanned-expenditure of inspection. The result is showing significance with 0.3402. The value indicates that for every unit change in Para_B. the logit will change 0. 3402 unit.

For Money_Value: Money_Value indicates the amount of money involved in misstatements in the past audits. The result is showing significance with parameter coefficient as 0.3006. The result indicates that for every unit change in Money_Value, the logit will change 0. 3006 unit.

Odds Ratio Results:

For Para_A: the result shows significance, and it is showing that for each unit of increasing in Para_A, the odds of risk is 1.696 times than before averagely.

For Para_B: the result shows significance, and it is showing that for each unit of increasing in Para_B, the odds of risk is 1.405 times than before averagely.

For Money_Value: the result shows significance, and it is showing that for each unit of increasing in Money_Value, the odds of risk is 1.351 times than before averagely.

Goodness-Of-Fit Test:

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
6.2358	7	0.5125

Table 19: Irrigation Sector Goodness-of-Fit result

For Goodness-Of-Fit test, the SAS automatically set into 9 boxes sorting by Continuous Variables (Para_A, Para_B, and Money_Value). We use Chi-Square tests to compare observed value and expected value with 9-2=7 degree of freedom. The result is insignificant with P-Value=0.5125 which means the model is strongly fitted.

4.3 Logistic Regression Modeling (Animal Husbandry Sector):

Outputs and Interpretations:

This modeling is based on single sector Animal Husbandry, and there are 95 samples used in this regression modeling.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.1397	0.4461	23.0102	<.0001
PARA_A	1	0.4603	0.1043	19.4674	<.0001
PARA_B	1	0.0703	0.0769	0.8351	0.3608
Money_Value	1	0.3855	0.2200	3.0716	0.0797

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PARA_A	1.584	1.291	1.944
PARA_B	1.073	0.923	1.247
Money_Value	1.470	0.955	2.263

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
20.8842	8	0.0075

Table 20: Animal Husbandry Sector results

Even through, we have significant result in Intercept and Para_A terms as shown in SAS outputs, but this result might not be useful because we have significant result in goodness of fit test which means the model is rejecting null hypothesis and the model is not fit.

4.4 Logistic Regression Modeling (Buildings and Roads Sector):

Outputs and Interpretations:

This modeling is based on single sector Buildings and Roads, and there are 82 samples used in this regression modeling.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.6662	2.0098	5.3901	0.0203
PARA_A	1	0.6698	1.0567	0.4018	0.5262
PARA_B	1	0.8329	0.5000	2.7749	0.0958
Money_Value	1	0.2460	0.1335	3.3959	0.0654

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PARA_A	1.954	0.246	15.502
PARA_B	2.300	0.863	6.129
Money_Value	1.279	0.984	1.661

Table 21: Buildings and Roads Sector results

Modeling technique is failed in Buildings and Roads Sector because coefficients of continuous variables are not significant.

4.5 Logistic Regression Modeling (Public Health Sector):

Outputs and Interpretations:

This modeling is based on single sector Public Health, and there are 76 samples used in this regression modeling.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.1532	1.4407	4.7904	0.0286
PARA_A	1	0.2828	0.9307	0.0923	0.7612
PARA_B	1	0.8847	4.3792	0.0408	0.8399
Money_Value	1	1.1171	2.6179	0.1821	0.6696

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PARA_A	1.327	0.214	8.224
PARA_B	2.422	<0.001	>999.999
Money_Value	3.056	0.018	517.034

Table 22: Public Health Sector results

Modeling technique is failed in Public Health Sector because coefficients of continuous variables are not significant.

4.6 Logistic Regression Modeling (Forest Sector):

Outputs and Interpretations:

This modeling is based on single sector Forest, and there are 74 samples used in this regression modeling.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3583	0.4736	8.2252	0.0041
PARA_A	1	0.00719	0.4144	0.0003	0.9862
PARA_B	1	0.1797	0.1860	0.9337	0.3339
Money_Value	1	0.5544	0.2362	5.5105	0.0189

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PARA_A	1.007	0.447	2.269
PARA_B	1.197	0.831	1.723
Money_Value	1.741	1.096	2.766

Table 23: Forest Sector MLE results

MLE Result:

For Intercept: the estimate is -1.3583 with P-Value = 0.0041 which is significant result. It means when other parameters are zero, the probability of audit files been at risk is less than the probability of non-audit files:

$$\text{OddInt} = \exp(-1.3583) = 0.2571$$

For Para_A and Para_B: the coefficients for these two parameters are not significant which means there is not enough evidence to say Para_A and Para_B have significant relations with logit.

For Money_Value: Money_Value indicates the amount of money involved in misstatements in the past audits. The result shows significance with parameter coefficient as 0.5544. The result indicates that for every unit change in Money_Value, the logit will change 0. 5544 unit.

Odds Ratio Results:

The results for both Para_A and Para_B are insignificance, but significance with Money_Value.

Goodness-Of-Fit Test:

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
3.1868	8	0.9221

Table 24: Forest Sector Goodness-of-Fit Test result – Optional

For Goodness-Of-Fit test, the SAS automatically set into 10 boxes sorting by Continuous Variables (Para_A, Para_B, and Money_Value), and the result is non-significance which means the model is fit.

4.7 Logistic Regression Modeling (Corporate, Fisheries, and Industries Sectors):

There are warnings for using logistic regression modeling techniques in Corporate, Fisheries, and Industries sectors. The reasons for coursing those errors are varied, and I think one of the reasons is high non-risking rate of audits as:

- 47 samples in sector 1.99 Corporate with 72.3% as its non-risking rate
- 41 samples in sector 21.61 Fisheries with 95.1% as its non-risking rate
- 37 samples in sector 59.85 Industries with 91.9% as its non-risking rate

The warning messages are:

- Warning: The maximum likelihood estimate may not exist.
- Warning: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

4.8 Comparison Test between Parameters within Models:

Based on the last part, we have concluded that Agriculture (55.57), Irrigation (3.89), and Forest (2.37) sectors are able to be fitted into logistic regression model with different coefficient of parameters. The rest of sectors whether their models are not fit, or they have non-significant results within both parameters.

Within this part, we would like to test if there are any significant differences between coefficients of the same parameters.

Recall the formula of logistic regression model:

$$Logit(\pi) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3$$

Agriculture (55.57):

$$Logit(\pi) = -3.7619 + 0.4337 * Para_A + 0.6006 * Para_B + 0.5802 * Money\_Value,$$

$$where\ Std(\widehat{Int}) = 0.5054, Std(\widehat{\beta_1}) = 0.4614, Std(\widehat{\beta_2}) = 0.1570, and\ Std(\widehat{\beta_3}) = 0.2875$$

Irrigation (3.89):

$$Logit(\pi) = -3.0066 + 0.5283 * Para_A + 0.3402 * Para_B + 0.3006 * Money\_value,$$

$$where\ Std(\widehat{Int}) = 0.6097, Std(\widehat{\beta_1}) = 0.2586, Std(\widehat{\beta_2}) = 0.1564, and\ Std(\widehat{\beta_3}) = 0.1243$$

Forest (2.37):

$$Logit(\pi) = -1.3583 + 0.0072 * Para_A + 0.1797 * Para_B + 0.5544 * Money\_Value,$$

$$where\ Std(\widehat{Int}) = 0.4736, Std(\widehat{\beta_1}) = 0.4144, Std(\widehat{\beta_2}) = 0.1860, and\ Std(\widehat{\beta_3}) = 0.2362$$

With α=0.05, the estimation of Z-Score is 1.96, and the idea for comparing two Confidence Intervals is to find upper bounds and lower bounds for three sectors and check if there are any intersection parts between sectors.

Agriculture (55.57) VS Irrigation (3.89) VS Forest (2.37):

Interpret Part:

- Agriculture (55.57): $-3.7619 \pm 1.96 * 0.5054 => (-4.7525, -2.7713)$
- Irrigation (3.89): $-3.0066 \pm 1.96 * 0.6097 => (-4.2016, -1.8116)$
- Forest (2.37): $-1.3583 \pm 1.96 * 0.4736 => (-2.2866, -0.4300)$

As above calculation shown, the intercept estimation in agriculture sector is significantly lower than the estimation of Forest sector, and it because the lower bound of interpret within Forest sector is less than the upper bound of interpret confidence interval within Agriculture sector. The result means when other parameters are zero, the probability of audit files been at risk within Agriculture sector is significantly higher than Forest Sector.

With comparing Agriculture sector and Irrigation sector, and Irrigation sector and Forest sector, we can see there are having non-significant comparison results. It means when other parameters are aero the probability of audit files been at risk are not having significant difference between sectors.

Para_A Part:

- Agriculture (55.57): $0.4337 \pm 1.96 * 0.4614 => (-0.4706, 1.3380)$
- Irrigation (3.89): $0.5283 \pm 1.96 * 0.2586 => (0.0214, 1.0352)$
- Forest (2.37): $0.0072 \pm 1.96 * 0.4144 => (-0.8050, 0.8194)$

As above calculation shown, the Para_A estimations with comparing Agriculture, Irrigation and Forest sector, we can see there are non-significant comparison results. It means for every single unit of change in Para_A, the change of logit will not have any difference between Agriculture, Irrigation and Forest sectors.

Para_B Part:

- Agriculture (55.57): $0.6006 \pm 1.96 * 0.1570 => (0.2929, 0.9083)$
- Irrigation (3.89): $0.3402 \pm 1.96 * 0.1564 => (0.0337, 0.6467)$
- Forest (2.37): $0.1797 \pm 1.96 * 0.1860 => (-0.1849, 0.5443)$

As above calculation shown, the Para_B estimations with comparing Agriculture, Irrigation and Forest sector, we can see there are non-significant comparison results. It means for every single unit of change in Para_B, the change of logit will not have any difference between Agriculture, Irrigation and Forest sectors.

Money_Value Part:

- Agriculture (55.57): $0.5802 \pm 1.96 * 0.2875 => (0.0167, 1.1437)$
- Irrigation (3.89): $0.3006 \pm 1.96 * 0.1243 => (0.0570, 0.5442)$
- Forest (2.37): $0.5544 \pm 1.96 * 0.2362 => (0.0915, 1.0174)$

As above calculation shown, the Money_Value estimations with comparing Agriculture, Irrigation and Forest sector, we can see there are non-significant comparison results. It means for every single unit of change in Money_Value, the change of logit will not have any difference between Agriculture, Irrigation and Forest sectors.

5 Conclusions:

5.1 Summary and Findings:

Using statistical methods to detect potential fraud is becoming more and more important at present, and in this research project, we have found some significant findings by using logistic regression modeling techniques.

Due to the dataset we selected, the Agriculture sector has the most records, but it has a relatively high non-risk proportion (87%). Public Health and Buildings and Roads have 76 and 82 audit files respectively, and those two sectors have the highest proportions of risk, with 77.6% and 64.6% respectively.

The results of correlation tests between three continuous variables (Para_A, Para_B, and Money_Value) are significant. The results indicate there is enough evidence to show that all pairwise correlations are not equal to zero. However, three continuous parameters are kept in the modeling analysis, because all of the pairwise correlations are smaller than 0.5.

In our logistic regression modeling with classification variable results, we have a few statistically significant findings:

- First, we found some significant coefficients for sector indicator variables, and they are in the Animal Husbandry and Forest sectors. The results indicate that there are essential differences when compared with the last sector, which is Public Health. (The definition of indicator variable is shown in the methodology description)
- Second, we have found significant results in intercept, Para_A, Para_B, and Money_Value. The results indicate that the intercept, Para_A, Para_B, and Money_Value variables are essential to the model, and they are significantly affecting the change of logit.
- Third, in odds ratio results, the unit changes of Para_A, Para_B, and Money_Value are showing a significant positive rate which means both coefficients are greater than 1. These results shown whenever one unit of changes in these three continuous variables, the odds of audits in risk will change correspondingly with the point estimates.
- Fourth, the goodness of fit test of this model is not significant.

In the results of logistic regression modeling without classification variable, we have found:

- For Intercept, Para_A, Para_B, and Money_Value parts, we have found the results of parameter coefficients are highly significant with P-value < 0.0001. Therefore, the variables are affecting the model
- The odds ratio results of the three parameters are all significant. Whenever one unit of changes in these three continuous variables, the odd of risking audit will change correspondingly with the point estimates.

During the analysis process, Variance-Covriance matrix of coefficients for parameters are computed by using the SAS program, and these results are used for Wald Test. Both of the Wald-Test results are significant; they indicate that one or more coefficients for each parameter are not zero.

In our results of logistic regression modeling with singular sectors, we have found that Agriculture, Irrigation, and Forest sectors are able to be fitted into the model and the test of fit are non-significant.

However, we have significant estimated results within Animal Husbandry sector, but the test of fitness rejects the null hypothesis which means the model is highly rejecting the data. In addition, Buildings and Roads, Public Health, Corporate, Fisheries, and Industries sectors have non-significant results in parameter estimations (for $\beta_1, \beta_2, and\ \beta_3$) which means the Para_A, Para_B, and Money_Value are not influential variables with in our linear model.

Results of the pairwise comparison test between parameters and the estimation of interpret in Agriculture sector is significantly lower than the estimation of Forest sector. However, the pairwise comparison tests of continuous variables are showing non-significant as their results, and the results indicate that there is no essential change of Para_A, Para_B, and Money_Value's coefficients within this sector's modeling analysis.

5.2 Further Research Spotlight:

Logistic regression modeling is becoming extremely valuable for risk assessment. Due to the analysis we did in this honor research project, we have found both significant and non-significant results, and we tried to find the linear relationship between logit and predictor variables.

As a further research opportunity, I suggest that we try to find more explanatory variables and include them into our model and try build the model not only based on monomial relations, but also Binomial, trinomial, or even higher dimensions. In addition, we can also use transformation techniques to transform the explanatory variables into another format (Ex: Log(x), Exp(x), and Sin(x)) so that the residuals can be reduced.

6 Bibliography:

I. Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li *Applied Linear Statistical Models 5th edition*

II. Richard A. Johnson, Dean W. Wichern *Applied Multivariate Statistical Analysis 6th edition* Chapter 6

III. Robert V. Hogg, Joseph W. McKean, Allen T. Craig *Introduction to Mathematical Statistics 7th edition* chapter 6&8

IV. George Casella, Roger L. Berger *Statistical Inference* chapter 7&8&9

V. Statistic Canada introduction of North American Industry Classification System (NAICS) https://www.statcan.gc.ca/eng/subjects/standard/naics/2017/v3/introduction

VI. OECD *Compliance Risk Management: Managing and Improving Tax Compliance 2004*

7 Appendix (SAS Code):

Code:

```
PROC IMPORT
DATAFILE='/folders/myshortcuts/SAS_university_edition_folder/honor project/audit_risk.xlsx'
DBMS=XLSX OUT=audit_risk;
SHEET='audit_risk'; GETNAMES=YES;
RUN;

proc tabulate data= audit_risk missing order=formatted;
class Sector_score;
table Sector_score, n;
run;

proc format;
value sector_fmt
    55.57          = 'Agriculture'
    3.89           = 'Irrigation'
    1.85           = 'Animal Husbandry'
    2.72           = 'Buildings and Roads'
    3.41           = 'Public Health'
    2.37           = 'Forest'
    1.99           = 'Corporate'
    21.61          = 'Fisheries'
    59.85          = 'Industries'
;
run;
```

```
data Honor_Proj;
set work.audit_risk;

where Sector_score not in (2.34,15.56,2.36,17.68);

keep Sector_score PARA_A PARA_B TOTAL History Money_Value Risk
;
run;

data Honor_Proj_V2;
set work.honor_proj;

length Sector_Name $20.;
*format Sector_Name sector_fmt.;
Sector_Name = put(Sector_score,sector_fmt.);

if Money_Value = . then Money_Value = 0;

if Sector_Name = 'Agriculture'              then Agriculture = 1;          else Agriculture = 0;
if Sector_Name = 'Irrigation'               then Irrigation  = 1;          else Irrigation = 0;
if Sector_Name = 'Animal Husbandry'    then Animal_Husbandry = 1;   else Animal_Husbandry = 0;
if Sector_Name = 'Buildings and Roads'  then Buildings_Roads = 1;     else Buildings_Roads = 0;
if Sector_Name = 'Forest'                   then Forest = 1 ;              else Forest = 0;
if Sector_Name = 'Corporate'                then Corporate = 1;            else Corporate = 0;
if Sector_Name = 'Fisheries'                then Fisheries = 1;            else Fisheries = 0;
if Sector_Name = 'Industries'               then Industries = 1;           else Industries = 0;

run;
```

```
proc corr data=Honor_Proj_V2 pearson plots=matrix(histogram);
var PARA_A PARA_B Money_Value;
run;

proc calis data=Honor_Proj_V2 pcorr;
mstruct var =
PARA_A PARA_B Money_Value Agriculture
Irrigation Animal_Husbandry Buildings_Roads Forest
Corporate Fisheries Industries
;
run;

proc tabulate data = Honor_Proj_V2 missing order=formatted;
class risk Sector_Name Sector_score;

table Sector_score*Sector_Name all, risk n;

run;

/*
model with categorical variable, and also display variance-covariance matrix
*/

proc logistic data=honor_proj_V2 OUTEST = Cov_Matrix descending alpha=0.05;
class Sector_Name;
model risk = Sector_Name PARA_A PARA_B Money_Value/lackfit;
run; quit;
```

```
/*
model without categorical variable
*/

proc logistic data=honor_proj_V2 descending alpha=0.05;
model risk = PARA_A PARA_B Money_Value/lackfit;
run; quit;


%macro Sector_Analysis (sector,name);

proc logistic data=honor_proj_V2 descending alpha=0.05;
title "sector score is &sector. with &name. classification";
where sector_score eq &sector;
model risk = PARA_A PARA_B Money_Value/lackfit;
run; quit;

%mend Sector_Analysis;

%Sector_Analysis (55.57,Agriculture);
%Sector_Analysis (3.89,Irrigation);
%Sector_Analysis (1.85,Animal Husbandry);
%Sector_Analysis (2.72,Buildings and Roads);
%Sector_Analysis (3.41,Public Health);
%Sector_Analysis (2.37,Forest);
%Sector_Analysis (1.99,Corporate);
%Sector_Analysis (21.61,Fisheries);
%Sector_Analysis (59.85,Industries);
```