Dana DiVincenzo

Assignment 4: Data Management and Regression Analysis

Introduction

In this assignment, I was given two csv files which contained both audit fee information from the Audit Analytics database, and financial characteristics of firms from the Compustat Annual Industrial file. The goal of the assignment was to create a regression equation with audit fees as the dependant variable. In order to create a more accurate regression equation, I had to use explaratory data analysis to discover which independant variables would likely have a significant impact on audit fees. Additionally, I used data cleaning, management, and analysis to make educated guesses on which variables were not significant enough to include in my final regression equation.

Step 1 - Importing and Combining the Data Frames ¶

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.formula.api as sm

from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import r2_score
from sklearn.metrics import confusion_matrix
```

In [52]:
```
# Reading in the files
auditfiledf = pd.read_csv('D:/Desktop/dana/auditinfo.csv', encoding = "ISO-8859-1")
compustatfile = "D:/Desktop/dana/compustat.csv"
compustatdf = pd.read_csv(compustatfile)
```

In [53]:
```
# Combining the data frames into one
# After researching what the column headers mean, it seems that fyear is the same as
# FISCAL_YEAR, and tic is the same as BEST_EDGAR_TICKER
# The first step is to change the names of the column headers so they match in both file
s

auditfiledf = auditfiledf.rename(columns = {'FISCAL_YEAR':'fyear','BEST_EDGAR_TICKER':'t
ic'})

# Now that the column headers match, I can combine the data files
varsdf = pd.merge(compustatdf, auditfiledf, on=['tic', 'fyear'], how='outer')
```

Step 2 - Cleaning the Data

In the following section I clean the data, with the goal of creating a smallar data frame that only includes independant variables that significantly effect audit fees. To do this, I initially researched audit fees on Google Scholar to gain a general understanding of what an audit fee model was and the common determinants of audit fees. I then looked up the definition of all of the column headers from the data files given, as most of them were unfamiliar to me. Combining that information, I was able to make an educated guess on what variables to cut from my data frame right away. I then use computational analysis to continue eliminating independant variables from my data frame.

In [30]:
```
# I chose these variables because based on my reasearch they seemed like they would have the most influence on audit fees
newvarsdf = varsdf[["fyear", "act", "at", "ceq", "ebit", "ebitda", "invt", "lct", "pifo", "AUDIT_FEES"]]

# Here, I find out which columns contain high amounts of NaN values that could make my model less accurate
newvarsdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 708351 entries, 0 to 708350
Data columns (total 10 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   fyear       708216 non-null  float64
 1   act         74166 non-null   float64
 2   at          115898 non-null  float64
 3   ceq         115718 non-null  float64
 4   ebit        104713 non-null  float64
 5   ebitda      102389 non-null  float64
 6   invt        107337 non-null  float64
 7   lct         74391 non-null   float64
 8   pifo        22824 non-null   float64
 9   AUDIT_FEES  640710 non-null  float64
dtypes: float64(10)
memory usage: 59.4 MB
```

In [54]:
```
# pifo seems to have a much larger number of NaN values than the other variables, so I drop it
newvarsdf = newvarsdf.drop(["pifo"], axis = 1)
```

Step 3 - Creating a Correlation Table and Heatmap

To further analyze the possible affect of the independant variables on audit fees, I create a correlation table and heatmap for all of the variables below. The numbers in the table are the correlation coefficient for the associated variables. According to my research, a strong correlation coefficient usually has a value of .75 to 1, and a moderate correlation coefficient has a value of around .5 to .75. Based on these criteria and looking at the correlation table, no variables have a strong correlation to audit fees. However, this criteria is subjective, and the table still shows me which variables are likely to have a stronger effect on audit fees than others. Notably, fyear has the lowest correlation coefficient to audit fees, with a value of only .060952. The variables with the highest correlation coefficient are act, at, and ceq.

In [55]:
```
corr = newvarsdf.corr()
corr.style.background_gradient(cmap ='Reds')
```

Out[55]:

	fyear	act	at	ceq	ebit	ebitda	invt	lct	AUDIT_FEES
fyear	1.000000	0.004470	0.013979	0.030142	0.036066	0.042569	0.000913	0.009588	0.060952
act	0.004470	1.000000	0.878738	0.795528	0.736822	0.803530	0.761590	0.933275	0.714294
at	0.013979	0.878738	1.000000	0.602795	0.755036	0.695827	0.322583	0.895050	0.681888
ceq	0.030142	0.795528	0.602795	1.000000	0.611907	0.730549	0.289925	0.763414	0.685369
ebit	0.036066	0.736822	0.755036	0.611907	1.000000	0.964488	0.259888	0.685466	0.586250
ebitda	0.042569	0.803530	0.695827	0.730549	0.964488	1.000000	0.258360	0.787260	0.628417
invt	0.000913	0.761590	0.322583	0.289925	0.259888	0.258360	1.000000	0.783603	0.353143
lct	0.009588	0.933275	0.895050	0.763414	0.685466	0.787260	0.783603	1.000000	0.685861
AUDIT_FEES	0.060952	0.714294	0.681888	0.685369	0.586250	0.628417	0.353143	0.685861	1.000000

Step 4 - Test Regression Equations

Before creating my final regression equation, I wanted to create individual regression equations that test each independent variable, with audit fees as the dependant variable. This way, I could look at the R-squared value for each, and gain a better understanding of each individual relationship. The results of these regression tests are in the table below.

Independent Variable:	R-Squared Value:
fyear	.004
act	.510
at	.465
ceq	.470
ebit	.344
ebitda	.395
invt	.125
lct	.470

Shown in the table above, fyear had the smallest R-Squared value, which was .004. This low value means that financial year is responsible for very little of the variance in audit fees. I was not surprised to see this result considering the fact that fyear and audit fees had the lowest correlation coefficient in my heatmap earlier. Invt also produced a low R-Squared value of only .125. Although this is much higher than fyear's, it is still much lower than the R-Squared values produced by the rest of the variables. Therefore, I decided to drop both fyear and invt from my final regression equation.

Step 5 - The Final Regression Equation

Below, I create my multiple regression equation. After eliminating many of the possible independant variables, I ended up creating an equation with Y = AUDIT_FEES, X1 = act, X2 = at, X3 = ceq, X4 = ebit, X5 = ebitda, and X6 = lct.

In [51]:
```
result_all = sm.ols(formula = "AUDIT_FEES ~ act + at + ceq + ebit + ebitda + lct", data
= newvarsdf).fit()
result_all.summary()
```

Out[51]:

OLS Regression Results

Dep. Variable:	AUDIT_FEES	**R-squared:**	0.551
Model:	OLS	**Adj. R-squared:**	0.550
Method:	Least Squares	**F-statistic:**	8130.
Date:	Fri, 23 Oct 2020	**Prob (F-statistic):**	0.00
Time:	20:50:44	**Log-Likelihood:**	-6.4610e+05
No. Observations:	39830	**AIC:**	1.292e+06
Df Residuals:	39823	**BIC:**	1.292e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.454e+05	1.44e+04	65.792	0.000	9.17e+05	9.74e+05
act	311.6844	6.012	51.843	0.000	299.901	323.468
at	135.7273	2.586	52.483	0.000	130.658	140.796
ceq	-140.2066	4.453	-31.483	0.000	-148.935	-131.478
ebit	292.0862	28.966	10.084	0.000	235.313	348.860
ebitda	-353.1470	28.553	-12.368	0.000	-409.112	-297.182
lct	-68.9573	7.736	-8.914	0.000	-84.119	-53.795

Omnibus:	30897.803	**Durbin-Watson:**	0.535
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3871100.724
Skew:	2.989	**Prob(JB):**	0.00
Kurtosis:	50.925	**Cond. No.**	2.87e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.87e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Intepretation of Results

Shown by the Adjusted R-Squared value, we can determine that 55.0 percent of the variation in audit fees is due to act (current assets - total), at (assets - total), ceq (common ordinary equity), ebit (earnings before interest and taxes), ebitda (earnings before interest, taxes, depreciation, and amortization), and lct (current liabilities).

By looking at the coefficient column, we can see what one unit of change in each independent variable will do to audit fees. Increasing act by 1 unit will increase audit fees by 311.6844 units. Increasing at by 1 unit will increase audit fees by 135.7273 units. Increasing ceq by 1 unit will decrease audit fees by 140.2066 units. Increasing ebit by 1 unit will increase audit fees by 292.0862 units. Increasing ebitda by 1 unit will decrease audit fees by 353.147 units. Increasing lct by 1 unit will decrease audit fees by 68.9573 units. Therefore, act, at, and ebit have a positive relationship with audit fees, while ceq, ebitda, and lct have a negative relationship with audit fees.

Conclusion

For this assignment, I attempted to create the best possible regression model for audit fees, given audit fee information and financial characteristics of firms. Overall, I would say I was semi-successful in doing so. An Adjusted R-Squared value of .550 is not horrible, but I would guess that it is possible to get a higher value with a different model. However, I do believe that I was successful with exploratory data analysis, and with applying my new knowledge to the project. I also increased my experience with Pandas DataFrames, markdown tables, and with visualization, in my attempt to make a thorough and clean looking report in Jupyter Notebook.