Assignment 5.1: Building a Hazard Model for Default

Submission Details

- Submit through T-square dropbox
- The assignment is due on Nov 21^{st}.
- You have to submit ONLY
  - SAS program
  - Output in PDF format
  - You don't need to submit any datasets
  - Short and concise explanation (1-3 lines) for the specification used in the regressions (economic rationale for each variable and expected sign). For example, I chose *leverage* as one of the explanatory variables because it determines the default boundary and higher leverage is expected to result in a higher (+) likelihood of default.
  - Do not run the regressions first, see the sign and then input this sign in the writeup. Provide an economic explanation for why the variable should matter for default prediction and how (positive or negative) it would effect the default likelihood.

Assignment Tasks

1. Come up with a list of possible covariates that should matter for default prediction. Some possible sources are
    - Lecture on credit scoring
    - Chava and Jarrow (2004), Chava, Stefanescu and Turnbull (2012)
    - Variables computed in assignments 2, assignment 3 and assignment 4
2. Compute these explanatory variables
3. Do a *in-sample* estimation and prediction. Follow these steps
    - Use the entire time period 1962-2014 for estimation
    - Run a PROC LOGISTIC model with DESCENDING option with bankruptcy as the LHS variable and the variables in step 1 as explanatory variables
    - Present the output and fit statistics for the model (output of proc logistic).
4. Do a *out-of-sample* prediction. Follow these steps
    - Divide the sample into in-sample estimation period (1962-1990) and out of sample forecasting period (1991-2014)
    - Estimate the model with 1962-1990 data
    - Forecast default for 1991-2014 time period using the estimates from 1962-1990 time period and explanatory variable data from 1991-2014

- Rank the default probabilities into deciles (10 groups). Use PROC RANK
- Compute the number (and percentage of defaults) in each of the 10 groups during 1991-2014 time period.
- A good model is one that has the majority of defaults in decile 1 or 2 and very few in other deciles
- You can use other metrics like ROC curve etc., but you don't need to compute those measures

5. iterate steps 1-4 till you get a model with good out of sample performance

Data Extraction

- You can use the same datasets that you have used in previous assignment - CRSP and COMPUSTAT data
- bankruptcy data from 1964 onwards in .dta format. Use

```
proc import
datafile=Q:\Data-ReadOnly\SurvivalAnalysis\BR1964_2014.dta
out=BR1964_2014 dbms = dta replace;
run;
```

- In case you can't use the Stata .dta dataset, data in CSV format is also loaded on Q drive
- merge them with the bankruptcy data to create a dataset that I outlined in the class. multiple observations per firm per year, with an indicator variable that takes one in the year of bankruptcy and zero other wise
- make sure that the accounting and market data is lagged (so that there is no look ahead bias)

Data Analysis

Steps in the assignment

1. First, make a list of variables you expect to matter for default prediction.

2. pre-process CRSP (DSF) data to create variables that you require and at the frequency that you require (think whether you really need daily data)

3. DSF (CRSP) data

   - some of the key variables required: CUSIP, DATE, PRC, SHROUT (SHOUT = SHOUT*1000), RET
   - You can construct YEAR variable as YEAR = YEAR(DATE);
   - market capitalization E = ABS(PRC) * SHROUT
   - The data is DAILY. You need to first compute the standard deviation of equity returns (RET) based on the last one year.
   - You can use the PROC SQL to compute the cumulative annual annual return and standard deviation for each firm (CUSIP) for each YEAR

```
PROC SQL;
CREATE TABLE dsf_A4 AS
SELECT CUSIP, YEAR(date) as year,
EXP(SUM(LOG(1+ret)))-1 AS annret,
STD(ret)*SQRT(250) as sigmae
```

```
FROM DSF
GROUP BY cusip, year;
QUIT;
```

- the above code collapses the DAILY data to ANNUAL data
- Note: for the assignment you would need only ANNUAL data from this point on (once the daily standard deviation is computed)
- After the above steps you should have a dataset with CUSIP, YEAR, this year's cumulative return (ANNRET), this year's annualized daily standard deviation (SIGMAE)
- Note that you would need to LAG the cumulative annual return and equity standard deviation. An easy way to do would be to say YEAR = YEAR+1 so for example, standard deviation computed from returns of 2008 is assigned to observations in 2009.

4. pre-process COMPUSTAT (FUNDA) data and create the required variables (think whether you require 1000+ variables)

5. merge CRSP, COMPUSTAT and bankruptcy data making sure that there is no look ahead bias (explanatory variables are available to the market at the time of estimation)

    - Linking DSF (CRSP) and FUNDA (COMPUSTAT)
    - Use CUSIP which is the unique firm identifier in both the datasets.
    - (not required for the assignment but you can try cusip = substr(cusip,1,6) to match on the first 6 characters of the cusip for a

match that gives larger coverage.)

- Make sure that the FUNDA data is lagged appropriately (so that it is available to the market at the time of estimation).
- So after lagging, you can merge with DSF data using CUSIP and YEAR
- Don't forget to SORT the data before you merge the data

6. Compute the required variables
7. Compute the required statistics
8. Use PROC LOGISTIC to model default occurrence
9. Save the results in PDF format. I don't want to see hundreds of pages of output.
10. Check the LOG file to see if there are any errors in your code
11. Upload the SAS program and results to tsquare

SAS commands that may be useful for the Data Analysis

As I mentioned in the class, SAS provides multiple methods to perform any given task. Some of the following commands may be useful (look up the examples from the UCLA SAS web site link that I posted on Tsquare)

- PROC LOGISTIC
- PROC MODEL
- PROC RANK

- DATA step
- PROC SORT
- PROC FREQ
- PROC UNIVARIATE
- PROC MEANS
- PROC PRINT
- PROC CORR