The Big Short (Interest): Closing the Loopholes in the Dividend-Withholding Tax

Instructions for replication

Elisa Casi, Evelina Gavrilova, David Murphy and Floris Zoutman

1. Overview

This package is structured in the following way:

Folders:

**0\_raw\_data** – contains the raw data, subject to data availability. See below for more information.

**1\_transformation\_code** – code which transforms the data from raw to analysis

**2\_final\_data** – data used for the analysis

**3\_analysis\_code**  - code used for the analysis

**4\_output** – folder for output

There are three streams of data, used in the paper. For easier overview, each data stream is denoted by an initial in the naming of the file, e.g.

Figures\_3\_8\_m.do contain the code for producing figures 3 & 8 with the m-markit data.

M – Markit, now S&P Global

C – Compustat

T - Tax

1. Data Availability Statement

These are the data used in the paper:

1. Data from S&P Global, (at the time known as IHS Markit) was downloaded on August 6th 2020. This data is internally known as DXOpen\_BuysideAnalyticsEU\_Standard, and was provided through a download provided by IHS Markit per the terms of our contract. The data we analyze in the paper ranges from January 1st 2010 to December 31st 2019. This data is proprietary and cannot be shared.
2. Data from Compustat Securities Daily was downloaded through the Wharton Research Data Services (WRDS) portal per the contract between our institution, NHH Norwegian School of Economics, and WRDS. We accessed the data on 13 February 2021. This data is proprietary and cannot be shared.
3. Data from Compustat Annual Fundamentals was downloaded through the Wharton Research Data Services (WRDS) portal per the contract between our institution, NHH Norwegian School of Economics, and WRDS. We accessed the data on 16 December 2022. This data is proprietary and cannot be shared.
4. Tax Revenue Data was received separately from the tax authorities/statistical agencies in Denmark (on 5 October 2021), in Finland (on 23 December 2021), in Norway (on 19 November 2021) and in Sweden (on 16 November 2021). Data was converted to annual aggregates (when appropriate) and converted into millions as displayed in Table C.3. Raw data as displayed in table 3C is saved 4\_output. For running the synthetic DID analysis, we converted the data from local currency to US Dollars. We use world bank annual exchange rate from here: https://data.worldbank.org/indicator/PA.NUS.FCRF?end=2019&locations=DK&start=2010. Exchange rate data and raw dividend withholding tax revenue is contained in the folder **0\_raw\_data.**
5. Exchange rate data from the IMF on 22 November 2021. This data is contained in the folder **0\_raw\_data.**

The first three data sources are proprietary data and can therefore not be shared. However, the data is available through academic subscription, and in practice many researchers will have access to this data either directly or through their institution.

1. Instructions

Open the folder **1\_transformation\_code** and run the master do-file. Comments provided within.

Open the folder **3\_analysis\_code** and run the master do-file. Comments provided within.

Variables are either labeled or commented in the do-file, when used first in the transformation code and in the analysis code, separately.

1. Format and software versions

[Floris inserts his Stata setup]

**For Figures R\_plots\_ridgeline\_map.R**

R version4.3.0 (2023-04-21)gginnards\_ 0.2.0.1

stargazer\_ 5.2.3

haven\_ 2.5.4

ggplot2\_ 3.5.1

synthdid\_ 0.0.9

**For Figure10\_Table5\_T.R:**

R version 4.3.0 (2023-04-21)

gginnards\_0.2.0

stargazer\_5.2.3

haven\_2.5.2

ggplot2\_3.4.2

synthdid\_0.0.9

Basic pseudocode to take 20+GB source files and query/concatenate Markit and Compustat together with prefixes:  
  
// 1. Ingest raw .txt files into Athena via external tables

for each source in [“Compustat”, “Markit”]:

// upload the source’s .txt files to S3

upload\_txt\_files\_to\_s3(source, s3\_path=“s3://bucket/raw/{source}/”)

// define an Athena external table over the uploaded .txt files

create\_athena\_table(

table\_name = source.lower() + "\_raw",

s3\_location = “s3://bucket/raw/{source}/”,

format = "TEXT",

schema = {...} // columns as in the .txt files

)

// 2. Query and transform by country

results = []

for each country\_code in LIST\_OF\_COUNTRIES:

// pull Compustat rows for this country

comp = athena\_query(

"SELECT \* FROM compustat\_raw WHERE country\_code = '{country\_code}'"

)

// prefix all data columns with “c\_” (leave country\_code untouched)

comp\_prefixed = add\_prefix\_to\_columns(comp, prefix="c\_", exclude=["country\_code"])

// pull Markit rows for this country

mark = athena\_query(

"SELECT \* FROM markit\_raw WHERE country\_code = '{country\_code}'"

)

// prefix all data columns with “m\_”

mark\_prefixed = add\_prefix\_to\_columns(mark, prefix="m\_", exclude=["country\_code"])

// merge the two sources on the country\_code key

merged = merge\_datasets(

left = comp\_prefixed,

right = mark\_prefixed,

on = ["country\_code"],

how = "outer"

)

results.append(merged)

// 3. Concatenate all country‐level frames into one

alltodos = concatenate\_dataframes(results, axis=0)

// 4. Persist the result as an Athena table

create\_athena\_table\_as\_select(

table\_name = "alltodos\_processed",

database = "finance\_db",

query = "SELECT \* FROM UNNEST(ARRAY[alltodos])"

// or use a CTAS:

// "CREATE TABLE finance\_db.alltodos\_processed WITH (format = 'PARQUET') AS SELECT \* FROM (...)"

)

// End of preprocessing pseudocode for alltodos.do equivalence