From missing to imputed values and regressions

What have you done until now?

1. From the file excel **Cbcr data.xls** you select your data (before starting: check that data are unfiltered)

2. Copy and paste your sample of data in a new sheet (called it i.e. BASE_2)

3. Eliminate the observations with missing values in profits, number of employees, tangible assets, related party revenues

4. Calculate the effective tax rate (ETR) as ratio between tax accrued and profit before tax

5. Select all the observations with positive profits (>0) and effective tax rates, equal to zero or positive, but smaller than 50% (0.5). If you have tax accrued equal to zero and then ETR=0, you must leave the observation, not to eliminate

6. Transform in natural log the variables: Profit before tax, Number of employees, Tangible assets, Related party revenues

7. Run the linear and not-linear regressions with control variables (the dependent variable is ln(profits) and ETR and control variables are the independent variables, add ETR square in the non-linear regression)

8. Apply U-test

What have you done?

1. From the file excel **Cbcr data.xls** select your data

2. Copy and paste your sample of data in a new sheet (called it BASE_2)

3. **Eliminate the observations with missing values** in profits, number of employees, tangible assets, related party revenues

4. Calculate the effective tax rate (ETR) as ratio between tax accrued and profit before tax

5. Select all the observations with positive profits (>0) and effective tax rates, equal to zero or positive, but smaller than 50% (0.5). If you have tax accrued equal to zero and then ETR=0, you must leave the observation, not to eliminate

6. Transform in natural log the variables: Profit before tax, Number of employees, Tangible assets, Related party revenues

7. Run the linear and not-linear regressions with control variables (the dependent variable is ln(profits) and ETR and control variables are the independent variables, add ETR square in the non-linear regression)

8. Apply U-test

Missing values: complete cases

You have applied listwise deletion, that is, you use only complete cases (CC) analysis, ignoring all cases with missing values.

This procedure does not lead to biased estimates, but may significantly reduce the sample sizes for analysis.

Missing values: **Imputation**

- Missing value imputation is the process of replacing missing data points in a dataset with estimated or arbitrary values to create a complete dataset.
- Common methods include simple imputation (using the mean, median, or mode) and multivariate imputation (using algorithms like k-nearest neighbors or regression to estimate values based on other variables).

Missing values: **Imputation**

- Regression imputation uses a regression model (like linear regression) to predict and fill missing values in a dataset, using other correlated variables, providing better estimates than mean/median imputation

Missing values: **Imputation**

How Regression Imputation Works

1. **Identify Variable:** Choose the variable with missing data (e.g., `X3`) and variables with strong correlations to it (e.g., `X1`, `X2`).
2. **Build Model:** Train a regression model (e.g., `X3 = b0 + b1*X1 + b2*X2`) on the complete cases.
3. **Predict & Fill:** Use the trained model to predict the missing `X3` values based on their corresponding `X1` and `X2` values.

Missing values: **Imputation**

- Variables with missing data: Profit before tax, Tax paid, Number of employees, Tangible assets, Tax accrued, Related party revenues

- Variables correlated from the jurisdiction country: Gross domestic product (GDP) per capita, Population, Corporate Income Tax (CIT)

- Imputed values: Estimates the missing values of the variable based on the values of the correlated variables

Missing values:complete cases regression

- from the excel file Cbcr_data imputation example,
- sheet original dataset includes 1395 observations
- some variables have missing values
- eliminating the missing values from the original dataset, the sheet dataset with missing values CC includes 883 observations (about 63%)

Missing values: complete cases regression

	beta	*Standard error*	*Stat t*	*p-value*
Constant	7.151	0.301	23.759	0.000
ETR	-4.650	1.169	-3.976	0.000
square of ETR	6.569	2.545	2.581	0.010
ln_employees	0.352	0.030	11.738	0.000
ln_tangible_assets	0.325	0.020	16.006	0.000
ln_related_revenues	0.170	0.018	9.718	0.000
Number of observations	883			
R-square adjusted	0.752			
F (p-value)	0.000			

Missing values: **Imputation**

- For all the observations (observed and missing), collect data for GDP per capita, population and corporate income tax (CIT) related to the jurisdiction country and year
- Save the sheet as txt file with name imputation.txt
- Imputed values: Estimates the missing values of the variable based on the values of the correlated variables using the MATLAB code

Example for profit before tax

```
%% Load data
opts = detectImportOptions('imputation.txt', ...
'Delimiter', '\t',...
'FileType', 'text');
data = readtable('imputation.txt', opts);


%% Identify missing and non-missing
missingIdx = isnan(data.profit_before_tax);
observedIdx = ~missingIdx;


% Training data for regression
GDP_train = data.ln_GDP_pop(observedIdx);
POP_train = data.ln_pop(observedIdx);
CIT_train = data.CIT(observedIdx);
profit_before_tax_train = data.profit_before_tax(observedIdx);
```

```
%% Fit linear regression model: ~ GDP + POP
mdl= fitlm([GDP_train POP_train CIT_train], profit_before_tax_train)


%% Predict missing BUR values
GDP_missing = data.ln_GDP_pop(missingIdx);
POP_missing = data.ln_pop(missingIdx);
CIT_missing = data.CIT(missingIdx);


profit_before_tax_pred = predict(mdl, [GDP_missing POP_missing CIT_missing]);


% %% Enforce positivity
% tax_pred(tax_pred <= 0) = 0.0001;


%% Insert imputed values back into table
data.profit_before_tax(missingIdx) = profit_before_tax_pred;


%% Save updated dataset to Excel
outputFile = 'imputed_profit_before_tax.xlsx';
writetable(data, outputFile);


fprintf('Imputation completed. Saved to %s\n', outputFile);
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
```

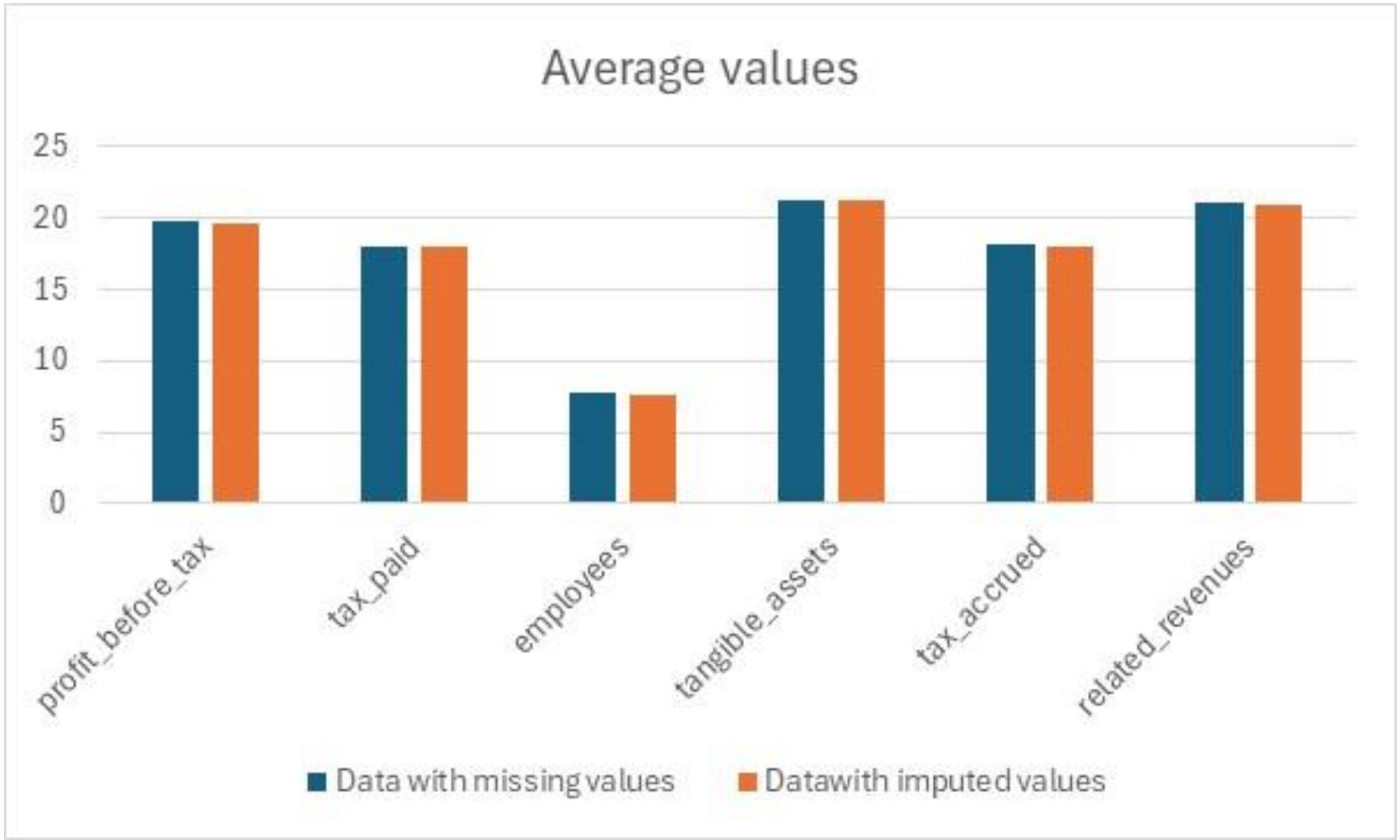
	Estimate	SE	tStat	pValue
(Intercept)	-1.5961e+09	6.0356e+08	-2.6445	0.0083078
x1	1.054e+08	3.371e+07	3.1267	0.0018188
x2	3.1514e+07	2.5942e+07	1.2148	0.22473
x3	1.4444e+09	6.6336e+08	2.1775	0.029678

```
Number of observations: 1008, Error degrees of freedom: 1004
Root Mean Squared Error: 1.17e+09
R-squared: 0.0158,  Adjusted R-Squared: 0.0128
F-statistic vs. constant model: 5.37, p-value = 0.00114



Imputation completed. Saved to imputed_profit_before_tax.xlsx
```

Comparison of the missing and imputed values (in ln)



Missing values: **Imputation**

- Imputed values: Estimates the missing values of the variable based on the values of the correlated variables using the MATLAB code
- Create a new sheet with imputed data
- Run the regression with the imputed data and apply the U-test

Missing values: imputation regression

	beta	*Standard error*	*Stat t*	*p-value*
Constant	7.334	0.229	32.042	0.000
ETR	-4.618	0.941	-4.907	0.000
square of ETR	6.576	2.044	3.218	0.001
ln_employees	0.365	0.025	14.682	0.000
ln_tangible_assets	0.292	0.015	18.910	0.000
ln_related_revenues	0.189	0.013	14.534	0.000
Number of observations	1291			
R-square adjusted	0.782			
F (p-value)	0.000			

Your next steps

1) Starting from your sample with missing values, for each jurisdiction country and year, data collection:

- GDP per capita (https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)
- Size of population (https://data.worldbank.org/indicator/SP.POP.TOTL)
- Corporate income tax (https://taxfoundation.org/data/all/global/corporate-tax-rates-by-country-2024/)

Your next steps

2) Save your sheet as imputation.txt file with missing values and data on GDP per capita (in natural log), population (in natural log) and corporate income tax (CIT) for all the observations

3) Run the Matlab file for each variable with missing values (profits, number of employees, tangible assets, tax accrued, related party revenues)

4) Construct the imputed dataset from the excel file with imputed data

Your next steps

5) Calculate the effective tax rate (ETR) as ratio between tax accrued and profit before tax

6) Select all the observations with positive profits (>0) and effective tax rates, equal to zero or positive, but smaller than 50% (0.5). If you have tax accrued equal to zero and then ETR=0, you must leave the observation, not to eliminate

7) Transform in natural log the variables: Profit before tax, Number of employees, Tangible assets, Related party revenues

Your next steps

8) Run the linear and not-linear regressions with control variables (the dependent variable is ln(profits) and ETR and control variables are the independent variables, add ETR square in the non-linear regression)

9) Apply U-test

10) Compare your results using the imputed data with the results using missing values

Your presentation must investigate the following 2 aspects:

- Existence of Nonlinearity in your data sample
- Difference in regression results between complete case (CC) and imputed data