DEPARTMENT OF ECONOMICS AND RELATED STUDIES

ECONOMIC DATA ANALYSIS 1

SUMMATIVE ASSIGNMENT – SPRING/SUMMER 2023

The submission deadline for the Economic Data Analysis assignment is **Thursday 27 April 2023, by 11am** (UK time).

Introduction[1]

Unemployment is a problem for society, not only because of the lost output and wages, but also because it can have a direct impact on individuals' wellbeing and life satisfaction. If a society has a strong view that able people of working age should be in employment, then being unemployed could result in a fear of being stigmatized, a sense of shame and feelings of being inferior to others who are employed. All these would reduce an individual's life satisfaction.

Disutility from unemployment is a concept that we cannot measure directly, so instead we will use self-reported life satisfaction. This measure has its limitations but is widely used to quantify costs that we cannot observe, such as the effect of becoming chronically ill or other life-changing events.

We will be using an approach and data that is similar to the study 'Employment status and subjective well-being'[2], which used the European Values Study (EVS), a cross-country survey, to investigate life satisfaction between people of different employment status. The hypothesis used in that study was that unemployed people would on average be less satisfied with life than employed people, and that this relationship between employment status and life satisfaction would vary depending on social norms. There is evidence to support this hypothesis in some countries, and we will explore this relationship in EU countries.

One explanation for a relationship between employment status and reported life satisfaction is social norms regarding views towards working (termed a work ethic). If social norms are an

---

[1] Acknowledgement – Introduction, data and ideas are from CORE Doing Economics
[2] Stam, K., Sieben, I., Verbakel, E. and de Graaf, P.M., 2016. Employment status and subjective well-being: the role of the social norm to work. *Work, employment and society*, *30*(2), pp.309-333.

important determinant of life satisfaction, then we expect the gap in life satisfaction between employed and unemployed to be larger in countries with a stronger self-reported work ethic.

While the main focus of this project is on the (full-time) employed and the unemployed, we will also consider whether life satisfaction differs for other labour market groups, such as the retired. Norms of working may not be as strong for the elderly, so the lack of formal employment is likely to have less of an effect on life satisfaction compared with working-age people who are unemployed.

The assignment has three parts. The objective of the first part is to understand and summarise the data. The second part looks at the relationship between individual characteristics of the respondents, life satisfaction and work ethic. Finally, Part 3 tests our assumption that the gap in life satisfaction between employed, unemployed and retired differs across countries with different views on work ethic.

Before you start, please read the instructions on page 3 carefully.

Instructions

From the EDA VLE page (Assessment submission only) in *Summative Assignment – Instructions and Submission point:*

- Download the Excel file EDA23_summative_assignment_dataset.xlsx
- Download the paper by Stam et al (2015)
- Access the pre-recorded video: "Introduction to the EDA summative assignment". It presents the assignment and offers technical help for questions 1.5 and 2.2(b).

Each part is composed of questions that involve computational tasks or/and analytical answers. You are advised to use **Excel** (from Office 365) to complete the tasks. **Answer all questions.**

All the material from the EDA module remains available during the time of the assignment and you are allowed to use these resources.

If you think that there is a typo or an inconsistency in the assignment, please use the dedicated Discussion board "EDA Summative Assignment Forum" on the EDA VLE page (Assessment submission only). Please note that questions related to Excel tasks or concepts seen during the lectures will not be answered.

The marks for each question are **given in brackets**: they show the total for each part, the total for each question and the breakdown for sub-questions.

Respect the word limit indicated under each question. Markers will stop reading after the limit. The word count <u>does not include</u>: charts or tables copied from Excel, Excel formulas, mathematical expressions, captions and references to tables, figures or to the literature.

Provide your answers on a **report that respects professional standards:**

- Your answers have to be clearly indicated, expressed in a clear and appropriate manner.
- Use a standard font (Calibri, Arial or Times New Roman), size 12, font colour Black.
- Your charts should have a title, axis labels and a legend if necessary.
- Charts and tables that you copy from your Excel spreadsheet should be pasted as "Pictures". They should be legible and captioned.
- If you write an Excel formula on your report, briefly explain how it works: which function you use, and which variables or values the cells refer to.

Your final report should be compiled in a **single PDF document:**

- It is your responsibility to make sure that the PDF document is legible.
- You do not have to submit your Excel file.

Your report will be marked anonymously. Do not include your name or student number. **Only include your exam number** (starting with Y…).

This formally assessed assignment is marked as part of your Part I examinations and forms 100% of your final module mark for Economic Data Analysis (ECO00017C). Under no circumstances should you submit a report that you have worked on with another student, this is an **individual assignment for you to complete on your own.**

Submit your report by **Thursday 27 April 2023, 11am** (UK time) on the EDA VLE page (Assessment submission only) in *Summative Assignment - Instructions and Submission point.*

Part 1 – Understanding and summarising the data [50]

The objective is to understand the variables in the dataset, produce summary statistics and discuss potential limitations of the data in terms of accuracy and comparability.

Question 1.1 [15]

Open the Excel file EDA23_summative_assignment_dataset.xlsx

The "European Values Study (EVS)" collects data from a representative sample of individuals, from different European countries, in 1981-1984 (wave 1), 1990-1993 (wave 2), 1999-2001 (wave 3) and 2008-2010 (wave 4). Individuals are not followed over time: the sample is different for each wave. The 4 waves are available in separate tabs in the spreadsheet. The "Data dictionary" tab in your spreadsheet contains a description of all variables in the dataset. You are also encouraged to read the *Introduction* and *Data and measurement* sections of the paper by Stam et al. (2015).

Use this information as well as your knowledge in data analysis to answer the following questions.

(a) How would you characterise the design (cross-section, repeated cross-section, panel, time-series…) of the "European Values Study"?
[3] Justify your answer. Max 75 words

(b) Is the source of the data experimental or observational?
[3] Justify your answer. Max 75 words

(c) Life satisfaction and Employment are self-reported variables. What does "self-reported" mean? Explain why self-reported variables might create issues in terms of accuracy and comparability of the data.
[3] Max 75 words

(d) How would you interpret a value higher than 90 for the variable **percentile**? Briefly explain why the variable **percentile** is better suited than the absolute level of income (**Monthly household income**) to compare individuals at different points of the income distribution and across countries.
[3] Max 75 words

(e) The variable **Full employment** is constructed from the variable **Employment** with the following formula:
=IF(OR(R2="Full time", R2="Unemployed"), IF(R2="Full time",1,0),"")
Where R is the column for **Employment**
Explain how the formula works.
What is the value returned for an individual who only works part time?
[3] Max 75 words

Question 1.2 [10]

To identify the number of observations and produce basic descriptive statistics for our variables of interest, we will use Excel's PivotTable option.

Using only the Wave 4 data, make a Pivot Table showing the number of respondents (**Respondent number**, use *count)* as well as the average (*average*) and standard deviation (*StdDev*) of **Life satisfaction** and **Work ethic,** with **Country** as the row variable.

For Albania (first row), you should find 1,200 observations, an average life satisfaction of 6.47 with a standard deviation of 2.26.

- Paste your table into your report.
- How many countries are represented in Wave 4?
- How many observations (total number of respondents) are available in Wave 4?
- Are individuals generally happier in Iceland compared to Italy? Justify.
- Are unemployed individuals in Italy more likely to feel stigmatized compared to those living in Iceland? Justify.
  Max 75 words

Question 1.3 [7]

To get an idea of what the distribution of life satisfaction and work ethic across countries looks like, we will use box and whisker plots. We will also explore how the distributions changed between Wave 3 (1999-2001) and Wave 4 (2008-2010).

Using the Excel's PivotTable option, make a similar table as for question 1.2 but this time using data from Wave 3. Use **Country** as the row variable and show the average of **Life satisfaction** and **Work ethic**.

Copy the values from your two Pivot Tables (Wave 3 and Wave 4) to populate the following table in a separate spreadsheet. You do not need to paste this table into your report.

	Life satisfaction		Work Ethic	
Country	Wave 3	Wave 4	Wave 3	Wave 4
Albania	7.42	6.47	3.25	3.92
…	…	…	…	…
…	…	…	…	…

Table 1- Table for Question 1.3

Create two box and whisker plots that show the distribution of the average score of life satisfaction across countries, for Wave 3 and Wave 4. Do the same for the average score of work ethic.

- Paste your charts into your report.
- Comment on the overall distribution and the presence or not of outliers. Compare the mean and the median. What does the comparison tell us about the shape of the overall distribution? Has it changed over time?
  Max 100 words

Question 1.4 [7]

We now want to compare the change in life satisfaction, over time, for three specific countries: Denmark, Iceland and Italy.

Using the Excel's PivotTable option, create separate tables for Wave 1 (1980-1984), Wave 2 (1990-1993), Wave 3 (1999-2001) and Wave 4 (2008-2010). For each of the tables, use **Country** as the row variable, select only Iceland, Italy and Denmark and show the average of **Life satisfaction**.

Use the values from your Pivot Tables to populate the following table in a separate spreadsheet. You do not have to paste the table into your report.

	1981-1984	1990-1993	1999-2001	2008-2010
Denmark				
Iceland				
Italy				

Table 2- Table for Question 1.4

- Make a chart that compares the levels of life satisfaction in Denmark, Iceland and Italy over the four waves.
- Paste the chart into your report.
- Briefly present your chart and explain your choices in terms of data visualisation. You might want to refer to the work of Edward Tufte (2001) and Jonathan Schwabish (2014).
  Max 100 words

Question 1.5 [7]

We will now explore how social norms and attitudes towards work vary within each of our three countries (Denmark, Iceland and Italy) and over time.

(a) Using the Excel's PivotTable option, calculate the frequency distribution for the range of work ethic score in Denmark, Iceland and Italy, using the data from Wave 3.

Hint: you can express the frequency (*count*) as a "% of the column total" by accessing the option "Show Values As" in the Value Field Settings.

The first rows of your table should be similar to the following:

Count of Respondent number	Column Labels		
	⊟1999-2001		
Row Labels	Denmark	Iceland	Italy
1	0.23%	0.35%	0.21%
1.2	0.12%	0.35%	0.14%
1.4	0.23%	0.59%	0.14%

Figure 1 - First three rows of the Pivot table - Question 1.5 (Wave 3)

Create a similar Pivot Table for Wave 4.

- Paste your two tables into your report.
  [2]

(b) For each of our three countries, create a separate column chart that shows the frequency distribution of the work ethic score in the population, with bars of a distinct colour for Wave 3 (1999-2001) and Wave 4 (2008-2010).

Hint: copy and paste the values from your Pivot Tables on a separate spreadsheet before creating your column charts.

- Paste your three column charts into your report.
- Briefly compare the distributions of work ethic scores in Denmark, Iceland and Italy. Did the distributions change between 1999-2001 and 2008-2010? *[5] Max 100 words*

Question 1.6 [4]

We have removed from the dataset individuals who did not answer all questions, for example on their health, wellbeing, income or employment status. The non-response rate varied from 0.79% on wellbeing to almost 5% for variables like income.

- Define the concepts of sample error and sample bias. *[2] Max 50 words*

- Discuss whether these missing values are likely to cause sample error and sample bias in the data? *[2] Max 100 words*

Part 2 – Empirical relationships between individual characteristics, life satisfaction and work ethic [25]

We will now use correlation coefficients and scatterplots to look at the relationship between life satisfaction, work ethic and the other variables in our dataset.

Question 2.1 [8]

Using the data from Wave 4, create a table as shown below in Table 3 and calculate the required Pearson correlation coefficients. Display the values of the coefficients with 3 decimal places (3 d.p.). Refer to the "Data dictionary" tab to choose the relevant variables and to understand how they are constructed.

Hint: for employment status and gender, you will need to use the constructed variables **Full employment** and **Female**.

For the relationship between Age and Life Satisfaction you should obtain a correlation coefficient of -0.099.

Variables	Life satisfaction	Work ethic
Age	-0.099	
Education		
Full-time employment		
Gender		
Self-reported health		
Income		
Number of children		
Relative income		
Life satisfaction		
Work ethic		

Table 3- Pearson correlation coefficients - Question 2.1

- Paste your table into your report.
- Interpret the coefficients for Age, Education and Gender. For example, you can comment on the sign and magnitude of the coefficients.
- Explain whether the relationships implied by the coefficients are what you expected. For example, do you expect life satisfaction to increase or decrease with health, income etc.
  Max 125 words

Question 2.2 [10]

Next, we will look at the relationship between employment status and life satisfaction and investigate the hypothesis that this relationship varies with the average work ethic in a country.

(a) Using the data from Wave 4, create a Pivot Table showing the average life satisfaction according to employment status (selecting the Full time, Retired and Unemployed categories only), with **Country** as the row variable and **Employment** as the column variable.

- Paste your table into your report.
- Briefly, comment on any differences in average life satisfaction between the full-time employed, retired and unemployed.
  [4] Max 75 words

(b) Copy and paste the values from your Pivot Table on a separate spreadsheet and use these values to calculate, for each country, the difference in average life satisfaction between Full time and Unemployed (i.e. Full time minus Unemployed) and between Full time and Retired (i.e. Full time minus Retired). Add a column to your table with the data on average work ethic, for each country, based on the Pivot Table you produced for Question 1.2.

Make two separate scatterplots, one for the difference between Full time and Retired, the other one for the difference between Full time and Unemployed. Your scatterplots should show average work ethic on the horizontal axis and difference in life satisfaction on the vertical axis. You should have as many data points in each scatterplot as you have countries in your sample. Add a linear trend line to your scatterplots.

- Paste your two charts into your report.
- Comment on the relationship between work ethic and differences in life satisfaction.
- If we were to calculate the Pearson correlation coefficients between work ethic and differences in life satisfaction (for Full time minus Retired and Full time minus Unemployed), which sign and magnitude would you expect? Justify.
  [6] Max 100 words

Question 2.3 [7]

(a) Explain the difference between correlation and causation when looking at the empirical relationship between two variables.
[3] Max 100 words

(b) In Question 2.1, we estimated the relationship between self-assessed health and life satisfaction as well as income and life satisfaction. Explain the reasons why these relationships might not be causal.
[4] Max 150 words

Part 3 – Is there a statistically significant relationship between employment status and life satisfaction? [25]

The aim of this project was to look at the empirical relationship between employment and life satisfaction. When we calculate differences between groups, we collect evidence which may or may not support a hypothesis that life satisfaction is identical between different subgroups, especially when comparing full-time employed, retired and unemployed. Economists often call this "testing for statistical significance".

We are also interested in whether this relationship is different in countries with different social norms and especially attitudes towards work, or "work ethic". In particular, we will look at Iceland (low score of work ethic), Denmark (mid-range score of work ethic) and Italy (high score of work ethic), using the data from Wave 4.

Question 3.1 [18]

(a) Using the data from Wave 4, make a Pivot Table showing the average, standard deviation (use *StdDev*) and number of observations (use *count*) for **Life satisfaction**, with **Country** as a row variable (selecting only Denmark, Iceland and Italy), and **Employment** as a column variable (selecting only Full time, Retired and Unemployed).

- Paste your table into your report.
- Comment on any similarities or differences in the mean, distribution and number of observations across the Employment categories and across countries. *[5] Max 100 words*

Copy and paste the values of your Pivot Table on a new spreadsheet. Use these values to create two tables similar to Table 4: one for the difference between Full time and Retired (i.e. Full time minus Retired), one for the difference between Full time and Unemployed (i.e. Full time minus Unemployed).

Table 4 shows the values you should find for Iceland when looking at the difference between Full time and Retired.

Table 4 - Differences in means (Full time minus Retired) for Iceland, Denmark and Italy, critical t-value and 95% confidence interval width.

Country	Difference in means	t*	CI 95% (width)
Iceland	-0.243	2.013	0.547
Denmark			
Italy			

The column "Difference in means" is the difference between the average value of life satisfaction for Full-time employed and Retired individuals.

The column "t*" is the critical value from the t-Student distribution. Use the Excel function TINV, a probability of 0.05 and adopt a conservative approach to calculate the degrees of

freedom: take the smaller sample size between n_1 and n_2, minus 1. Where n_1 and n_2, are the number of observations respectively for Full time and Retired.

The column "CI 95% (width)" is the 95% confidence interval "width" of the difference in means: the distance from one end of the interval to the mean. Use the following formula:

$$CI\ 95\%\ width = \ t^* \times \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

Where t^* is the critical t value you calculated, sd_1 and sd_2 represent the standard deviation of life satisfaction respectively for Full time and Retired.

(b) After you created two tables similar to Table 4, one for Full time minus Retired, one for Full time minus Unemployed (see instructions above):

- Paste your two tables into your report.
- Write and describe the Excel formulas you have used to calculate the values presented in Table 4: difference in means, t* and CI95% for Iceland, Full-time minus Retired.
- Interpret the values presented in Table 4: difference in means, t* and CI95% for Iceland, Full time minus Retired.
  [8] Max 125 words

(c) Plot two column charts, one for Full time minus Retired, one for Full time minus Unemployed, showing the differences in means on the vertical axis and the three countries (Iceland, Denmark, Italy) on the horizontal axis. Add the confidence intervals from question 2.2(b) to your charts.

- Paste your two charts into your report.
- Interpret your findings. Is there evidence that life satisfaction is related to employment status? Does the relationship vary across countries?
  [5] Max 100 words

Question 3.2 [7]

The method we used to compare life satisfaction relied on making comparisons between people with different employment status, but a person's employment status is not entirely random. We cannot therefore make causal statements such as 'being unemployed causes life satisfaction to decrease'.

(a) Describe how we could use a natural experiment to assess better the effect of being unemployed on life satisfaction, and make some statements about causality.
  [3] Max 100 words

(b) If we were to use a Differences-in-Differences approach to estimate the causal effect of being unemployed on life satisfaction, what assumptions would we have to make to guarantee the internal and external validity of our results?
[4] Max 150 words

END OF THE ASSIGNMENT.