

Using Survey Data to Understand the Health Needs of Difficult-to-Reach Populations: Evidence from a Community Survey Regarding the Individual and Contextual Correlates of Sex Life Happiness Amongst European MSM

This file explains how to replicate the analysis for the paper “Using Survey Data to Understand the Health Needs of Difficult-to-Reach Populations: Evidence from a Community Survey Regarding the Individual and Contextual Correlates of Sex Life Happiness Amongst European MSM.”

Replication

The replication package consists of three elements: (1) combining the datasets and generating the variables for analysis from the raw data, (2) generating the four tables (one descriptive statistics table and three regression analyses tables) in the paper, and (3) exporting the tables into word and excel and making presentational modifications.

The paper uses sources from seven different datasets. However, the bulk of the data comes from just one dataset (EMIS 2017), while the other six datasets only contribute one variable each for the analysis.

The EMIS 2017 dataset is restricted because of its public health focus and privacy considerations, while all other datasets are publicly available. Details, of how to request access to the EMIS dataset is available below.

Once permission has been granted to access the relevant observations and variables from the EMIS 2017 wave, the script (Replication.do) can be run and will automatically incorporate the other dataset values into the raw dataset. The rest of the replication.do. file will then enable the replication of the four tables in the paper.

License for Code

The code is licensed under a Creative Commons license. See LICENSE for details.

Computational requirements

Software Requirements • Stata (used version 16) • All scripts take about 20 minutes to run in total.

Detailed Instructions

To run the script, a user only needs to change the directory of the replication.do file and run the code. The step-by-step preparation and analysis of the augmented EMIS2017 dataset is as follows:

Data Preparation and Augmentation (Replication.do lines x-x)

The first part of the replication.do file (lines 1-25) removes observations given our substantive question of interest and based on the recommendations of the EMIS Manual (see p.3 of the EMIS Manual: emis-project.eu/wp-content/uploads/2025/10/EMIS-2017_Variable-Manual.pdf).

The second part of this process (lines 25-569 and 1072-1091) involves transforming and renaming variables.

The third part of this process (lines 571-1067) involves augmenting the baseline (EMIS2017) dataset with data from the other six datasets. This is done manually (as there are only 43 observations at the country-level). This process can be verified by opening the datasets provided (see files Freedom, GDP, Health, Population and PowerIndex) and comparing relevant values - country and year (2017) between the original dataset and the lines 537-898).

Table 1 Descriptive Statistics

Lines 1134-1167 replicate Table 1 (also available as Tab1 in the Output files) which provides descriptive statistics of all the variables used in the subsequent analysis.

It is important to note that the code generates a Table with column percentages shown. However, in the final paper row percentages are shown- this was calculated manually from the output. The number of observations in each cell remains the same in both the code output and the final paper- making verification of the results easy.

For example, for the first variable in Table 1 'Alcohol consumption in previous 4 weeks' the STATA output is:

	Satisfied with sex life?			
	No		Yes	
	N	%	N	%
Alcohol consumption in the last 4 weeks				
Yes	8,410	(20.1%)	10,269	(14.8%)
No	33,429	(79.9%)	59,348	(85.2%)
Total	41,839	(100%)	69,617	(100%)

While in the final paper we present row total after manually transforming this output:

	Satisfied with sex life?					
	No		Yes		Total	
	N	%	N	%	N	%
Alcohol consumption in the last 4 weeks						
Yes	8,410	(45.0%)	10,269	(55.0%)	18,679	(100%)
No	33,429	(36.0%)	59,348	(64%)	92,777	(100%)

For ease of verification the Table provides the full question that participants were asked (e.g. ‘Alcohol consumption in previous 4 weeks’ each question is easily mapped onto a variable by (a) its position in the Tables is the same (e.g. this is the first variable in Table 1 generated by the STATA output and in the Table found in the paper), and (b) the variable name in the STATA output is intuitive (in this case ‘alcohol’).

Table 2 Baseline Regression Model

Lines 1169-1184 replicate Table 2 which provides regression analysis. The ‘Tab2’ document in the outputs file replicates these models.

Note: the post-estimation statistics are not in the exported table, but, after each regression model the likelihood ratio test is at the bottom of the regression table, the AIC and BIC are estimated by the ‘estat IC’ command and the intraclass correlation by the ‘estat ICC’ command.

Table 3 Subgroup Analysis

Lines 1185-1237 replicate Table 3 which provides regression analysis. The 'Tab2' document in the outputs file replicates these models. The 'Tab3' document in the outputs file replicates these models.

Note: the post-estimation statistics are not in the exported table, but, after each regression model the likelihood ratio test is at the bottom of the regression table, the AIC and BIC are estimated by the 'estat IC' command and the intraclass correlation by the 'estat ICC' command.

Table 4 Mediation Analysis

Lines 1239-1286 replicate Table (although most results are not reported in the paper). Table 4 in the paper reports only the indirect, direct, and total effect of the variable of interest. For the first model in Table 4, the coefficients of interest are calculated, post regression, by line 1251 (direct effect), lines 1253-1256 (indirect effect), and lines 1259-1262 (total effect). For the second model in Table 4, the coefficients of interest are calculated, post regression, by line 1275 (direct effect), lines 1277-1280 (indirect effect), and lines 1283-1286 (total effect). Tab4 in the Output file provides all the regression coefficients (not reported in the paper).

Note: the post-estimation statistics are not in the exported table, but, after each regression model the likelihood ratio test is at the bottom of the regression table, the AIC and BIC are estimated by the 'estat IC' command and the intraclass correlation by the 'estat ICC' command.

Note: the direct effect is the exponential of the beta coefficient of 'nofreedom' in the second stage the regression. For ease I have added a 'display exp(beta coefficient)' in the code.

Data Availability Statement

We list the raw data sources here and detail how they can be accessed in order to facilitate replication of the paper.

ALL DATA was last accessed on the 4th of February 2026.

EMIS 2017

Most of the data used in our analysis consists of confidential data from the '2017 European MSM Internet Survey' (EMIS). EMIS was undertaken by Sigma Research at the London School of Hygiene and Tropical Medicine (LSHTM) in association with the Robert Koch Institute (RKI) in

Berlin. The official website for this dataset, including details of how permission to access the data can be requested, is available here: [EMIS-2017 & LAMIS – EMIS](#) . The EMIS dataset provides all the observations (n=111,456) at the individual respondent level undertaken in our analysis.

The replication.do file lines (13-18) identifies which subset of respondents are included in the subsequent analysis, while lines (18-911) transform some of the raw data into the variables used in the subsequent analysis.

The other datasets used in our analysis provide data at the country-level (all EMIS participants are nested in 43 countries).

Freedom in the World Country Territories and Status Dataset

Freedom House's dataset on regime type, in 2017, provides data on regime at the country level (n=43). Each country is given a score ranging from 2-14.

This data is publicly available at: [Country and Territory Ratings and Statuses FIW 1973-2023 .xlsx](#)

This dataset was downloaded in Excel (Freedom) and the relevant data (2017 values for each of the 43 countries in the EMIS dataset) were incorporated into the main dataset via the replication.do file (lines 629-683).

World Bank Open Data Portal

Four of the additional six datasets are from the World Bank Open Data Portal. These being: GDP per capita (in current USD), healthcare expenditure (% of GDP), population (thousands), and the urbanisation rate (% of total population).

Each dataset can be found here (accessed 18/01/2026):

GDP per capita (current USD, year 2017): [GDP per capita \(current US\\$\) | Data](#)

Healthcare expenditure (% of GDP, year 2017): [Current health expenditure \(% of GDP\) | Data](#)

Population (Thousands, year 2017): [Population, total | Data](#)

Urbanisation rate (% of population): [Urban population \(% of total population\) | Data](#)

These datasets were downloaded in Excel (GDP, Health, Population, and Urban) and the relevant data (2017 values for each of the 43 countries in the EMIS dataset) were incorporated into the main dataset via the replication.do file (lines 539-812).

Varieties of Democracies (V-DEM): LGBTQ+ Access to Power Index

The LGBTQ+ Access to Power Index is available as part of the Varieties of Democracy (V-Dem) database.

The data is publicly available here (accessed 18/01/2026): [Country-Year: V-Dem Full+Others v15 – V-Dem](#)

This datasets were downloaded in Excel (PowerIndex) and the relevant data (2017 values for each of the 43 countries in the EMIS dataset) were incorporated into the main dataset via the replication.do file (lines 814-855).