Life Satisfaction and Sociodemographic Social Characteristics in Canada Validation and Replication Results

SUMMARY

This replication reproduces the linear regression analysis predicting life satisfaction from Aboriginal identity, sex, and marital status using the 2017 General Social Survey (Cycle 31 – Family). The analysis was conducted in RStudio (2025.09.0) using a single master script, as no separate program file was provided. All data were accessible via the CAnD3 course site, and the code executed without error. The replication successfully reproduced the author's reported outputs, including regression coefficients, standard errors, t-values, and model fit indices. The residual diagnostics were acceptable and revealed no major violations of model assumptions.

A minor methodological issue was identified: the marital status variable (MARSTAT) was coded as a numeric variable instead of a categorical factor, representing a conceptual model mis-specification. While this did not affect the computational reproducibility of the results, it compromises the interpretive validity of the model.

Data description

Data Sources

From: Statistics Canada, Social and Aboriginal Division

Data: General Social Survey - Family, Cycle 31 (2017) – Family, Social Support, and Life Satisfaction.

README only links to a general Statistics Canada webpage; access path for the actual data used is not described. The GSS 2017 PUMF was retrieved via the CAnD3 course LMS portal.

To download codebook:

https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816#a4

Analysis Data Files

Replication package included only a master R script:

- RRWM_master_script.R – Single master script for data cleaning, model estimation, and output generation
- No dataset provided, but retrieved from CAnD3 course portal.

No separate program file was included; replication followed the workflow and commands contained directly in the master script.

The script outputs three files:

1. RRWM_assumption_result_Xin.pdf – Regression diagnostic plots
2. RRWM_regression_result_Xin.txt – Model summary and ANOVA results
3. RRWM_model_forest.pdf – Coefficient plot

Data checks

INSTRUCTIONS: When data are present, run checks:

- File format and readability
    - dataset gss-12M0025-E-2017-c-31_F1.csv was read in R using read.csv().
    - The file is in CSV format, which is an archive-ready, open, and non-proprietary format suitable for long-term preservation.
    - The script initially referenced an absolute file path, but the file could be accessed using a relative path without issue.
    - No encoding or delimiter problems occurred, and all variables were read correctly.

- Data structure and content
    - The dataset corresponds to the 2017 General Social Survey (GSS), Cycle 31, Public Use Microdata File.
    - Lanlan's master script selected four variables for analysis:
        - SLM_01 – Overall life satisfaction
        - AMB_01 – Aboriginal identity
        - SEX – Sex of respondent
        - MARSTAT – Marital status
    - The structure (str(gss2017)) confirmed all variables imported successfully as numeric or integer fields.

- Variable labels
    - CSV files do not store variable labels, only column names.
    - Labels and code definitions for each variable are available separately in the official Statistics Canada documentation and were manually reconstructed in the R script during recoding.

- PII and disclosure control
    - The dataset is a publicly released, anonymized file provided by Statistics Canada.
    - It contains no direct identifiers (names, addresses, postal codes, contact information) and no fine-grained geographic detail.
    - No derived variables were created that could lead to re-identification.
    - Therefore, no personally identifiable information (PII) is present, and disclosure risk is negligible.

Code description

There is one R script (RRWM_Xin.R) serving as both master and only program file.

The GitHub repository also contains:

- README.md – brief description with dataset web link
- RRWM_assumption_result_Xin.pdf – regression-diagnostic figures
- RRWM_model_forest.pdf – coefficient (forest) figure
- RRWM_regression_result_Xin.txt – text output of regression and ANOVA
- gss-12M0025-E-2017-c-31.pdf – dataset codebook

No additional programs or modular scripts were provided.

Programs and Purpose

- RRWM_Xin.R (master script, 128 lines)
    - Imports data, subsets variables, cleans and recodes, estimates linear model, generates figures, and saves text/graphic outputs.
    - Creates one analysis object (gss_clean) in memory but does not write an analysis dataset to disk.

Data Preparation Code

- Lines 1-5: Load packages (haven, dplyr) and import the CSV file gss-12M0025-E-2017-c-31_F1.csv.
- Lines 8-11: Subset the dataset to four variables (SLM_01, AMB_01, SEX, MARSTAT), saved as gss_subset.
- Lines 14-16: Rename variables (SLM_01=feeling_lf, AMB_01 = Aboriginal ID).
- Lines 19-25: Clean the subset by filtering valid codes:

    - feeling_lf 0-10
    - aboriginal_id and SEX 1-2
    - MARSTAT 1-6
    - Produces the cleaned dataset gss_clean.
- Lines 75-90: Recoding/labeling after analysis – categorical variables relabeled for readability (SEX, MARSTAT, aboriginal_id).
- Lines 91-102: Reorder factor levels for MARSTAT (using forcats::fct_relevel()), stored as gss_plot.

Descriptive Summaries

- Lines 26-47: skim() and tbl_summary() produce descriptive statistics of gss_clean (displayed in console only, not exported).

Model Fitting and Output Generation

- Lines 50–60: Fit the multiple linear regression model predicting life satisfaction (feeling_lf) from Aboriginal identity (aboriginal_id), sex (SEX), and marital status (MARSTAT).
- The fitted model (lm_model) is used throughout the remainder of the script for diagnostic plotting, model summaries, ANOVA, and estimated marginal means (emmeans).
- Diagnostic plots are produced immediately afterward with plot(lm_model), generating four standard residual plots.

Model Results

- Lines 70–73: Save model summary and ANOVA results to a text file using sink().
- Output file: RRWM_regression_result_Xin.txt.

Figures

- Figure 1: Regression Diagnostics
    - Lines 52-68: plot(lm_model) produces four diagnostic panels (Residuals vs Fitted, Q–Q, Scale-Location, Residuals vs Leverage).
    - Lines 65-68: pdf("RRWM_assumption_result_Xin.pdf") … dev.off() to save the plots.
    - Output file: RRWM_assumption_result_Xin.pdf.

Stated Requirements

No stated requirements were noted in the README or within the RRWM_Xin.R script. Lanlan did not specify required software versions, package dependencies, or computational resources.

Missing Requirements

- Computational Requirements not specified.
- Time Requirements not specified

Computing Environment of the Replicator

No computing environment was stated. The following configuration reflects the environment used by the replicator to execute and verify the code.

- Operating System: Windows 11 Pro (64-bit, x64-based processor)
- Hardware:
  - Processor: 13th Gen Intel(R) Core (TM) i7-1370P @ 1.90 GHz
  - Installed RAM: 32.0 GB (31.7 GB usable)
  - System Type: 64-bit operating system, x64-based processor
- Software:
  - RStudio: 2025.09.0 Build 387 ("Cucumberleaf Sunflower" Release)
  - R Version: 4.4.1 (64-bit)
  - R Packages Loaded as noted in R script:
    - haven, dplyr, skimr, gtsummary, emmeans, forcats, ggplot2, broom
    - All packages were installed from CRAN using install.packages() commands included in the script.

Replication steps

1. Downloaded the master script and output files from the GitHub repository provided.
2. The dataset (gss-12M0025-E-2017-c-31_F1.csv) was not included, so I downloaded it separately from the CAnD3 Learning Management System (LMS), which provides access to the Statistics Canada General Social Survey (Cycle 31 – Family).
3. Because no runnable program file was available, I re-created the analysis from the annotated master script, following the structure and variable references indicated.
4. Imported the dataset into RStudio (2025.09.0 Build 387) using R version 4.4.1 (64-bit).
5. Installed and loaded the necessary packages (haven, dplyr, skimr, gtsummary, emmeans, forcats, ggplot2, broom) from CRAN.
6. Subsetted the GSS dataset to include four variables: SLM_01, AMB_01, SEX, and MARSTAT.

7. Renamed variables to descriptive names (feeling_lf, aboriginal_id) and filtered valid codes for each variable to remove out-of-range or missing responses.
8. Recoded categorical variables (SEX, MARSTAT, aboriginal_id) into labeled factor variables to improve readability in outputs.
9. Produced descriptive statistics using skim() and tbl_summary() to summarize sample characteristics.
10. Fitted a linear regression model predicting life satisfaction (feeling_lf) from Aboriginal identity, sex, and marital status.
11. Generated regression diagnostics using plot(lm_model) and exported the four diagnostic plots (Residuals vs Fitted, Q-Q, Scale-Location, Residuals vs Leverage) into a single PDF.
12. Exported the model summary and ANOVA table to a text file (RRWM_regression_result_Xin.txt).
13. Created a coefficient (forest) plot using broom::tidy() and ggplot2 to display model estimates, saved as RRWM_model_forest.pdf.
14. Total runtime was under one minute on my Windows 11 Pro (13th Gen Intel i7, 32 GB RAM) environment.

Findings

The replicated model produced results identical to those in Table 1 and Table 2, confirming the reproducibility of the analysis. Treating *life satisfaction* as a continuous outcome is methodologically appropriate for this dataset, as 0–10 satisfaction scales are routinely analyzed as continuous measures in social and population health research.

However, the marital status variable (MARSTAT) was not recoded as a categorical factor, meaning it was treated numerically (1–6). This coding choice incorrectly imposes a linear order across categories such as “Married,” “Widowed,” “Separated,” and “Single,” which have no inherent numeric progression. As shown in Tables 1 and 2, this resulted in a statistically significant coefficient ($B = -0.173$, $p < .001$) that reflects a numeric trend rather than meaningful group differences. As such, the model cannot be interpreted.

Tables

Lanlan’s Results (Table 1 & 2)

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.4879 -0.5512  0.2037  1.3766  2.6995

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.015076   0.123178  65.069  < 2e-16 ***
aboriginal_id  0.250696   0.058521   4.284 1.85e-05 ***
SEX            0.072144   0.024966   2.890  0.00386 **
MARSTAT       -0.172895   0.005961 -29.004  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.597 on 16515 degrees of freedom
Multiple R-squared:  0.05061,   Adjusted R-squared:  0.05044
F-statistic: 293.5 on 3 and 16515 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: feeling_lf
                 Df Sum Sq Mean Sq  F value    Pr(>F)
aboriginal_id     1     81   80.71  31.6619 1.865e-08 ***
SEX               1     19   19.36   7.5961  0.005856 **
MARSTAT           1   2144 2144.27 841.2100 < 2.2e-16 ***
Residuals     16515  42097    2.55
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

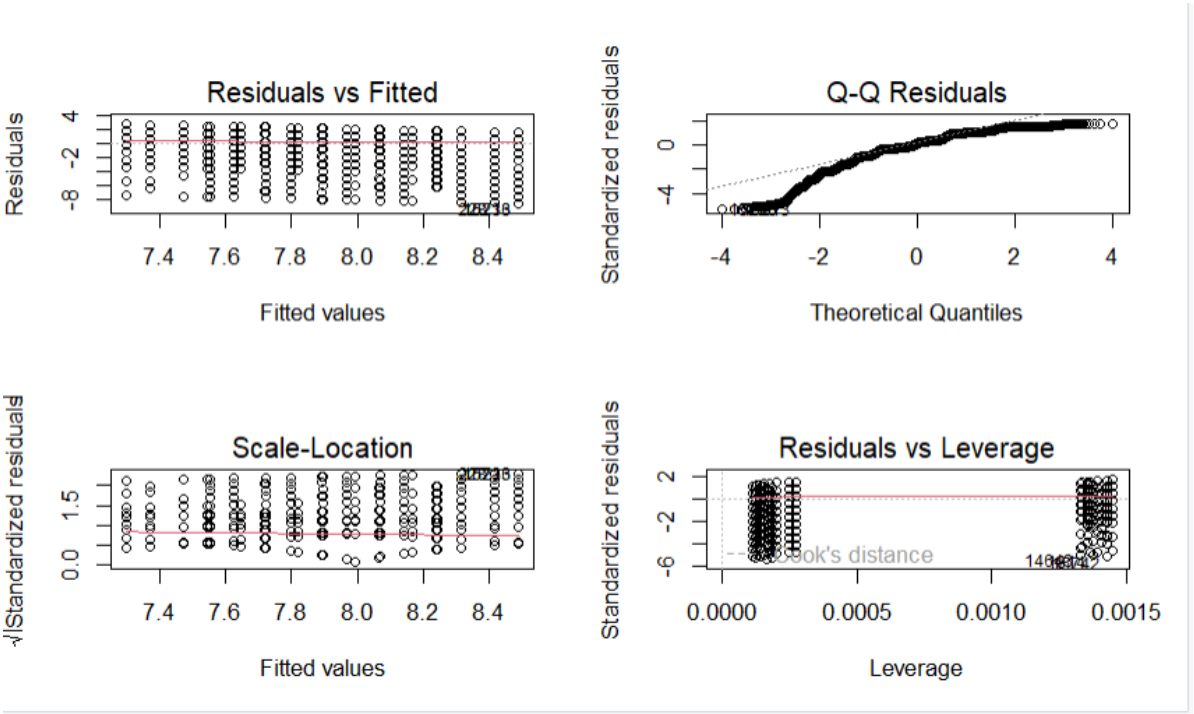
My Reproduced Results

```
Call:
lm(formula = feeling_lf ~ aboriginal_id + SEX + MARSTAT, data = gss_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4879 -0.5512  0.2037  1.3766  2.6995

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.015076   0.123178  65.069  < 2e-16 ***
aboriginal_id  0.250696   0.058521   4.284 1.85e-05 ***
SEX            0.072144   0.024966   2.890  0.00386 **
MARSTAT       -0.172895   0.005961 -29.004  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.597 on 16515 degrees of freedom
Multiple R-squared:  0.05061,   Adjusted R-squared:  0.05044
F-statistic: 293.5 on 3 and 16515 DF,  p-value: < 2.2e-16

> anova(lm_model)
Analysis of Variance Table

Response: feeling_lf
                Df Sum Sq Mean Sq  F value    Pr(>F)
aboriginal_id    1     81   80.71  31.6619 1.865e-08 ***
SEX              1     19   19.36   7.5961  0.005856 **
MARSTAT          1   2144 2144.27 841.2100 < 2.2e-16 ***
Residuals    16515  42097    2.55
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

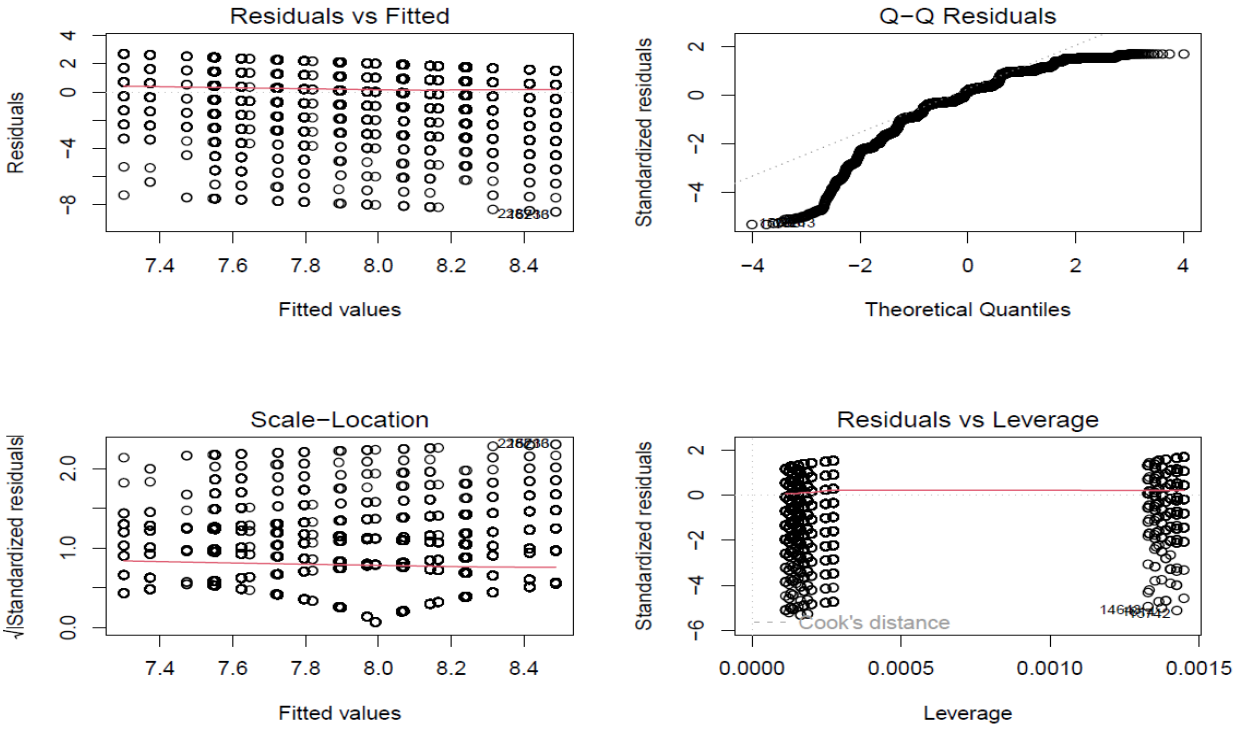
Figures

The figures I recreated were identical to Lanlan's.

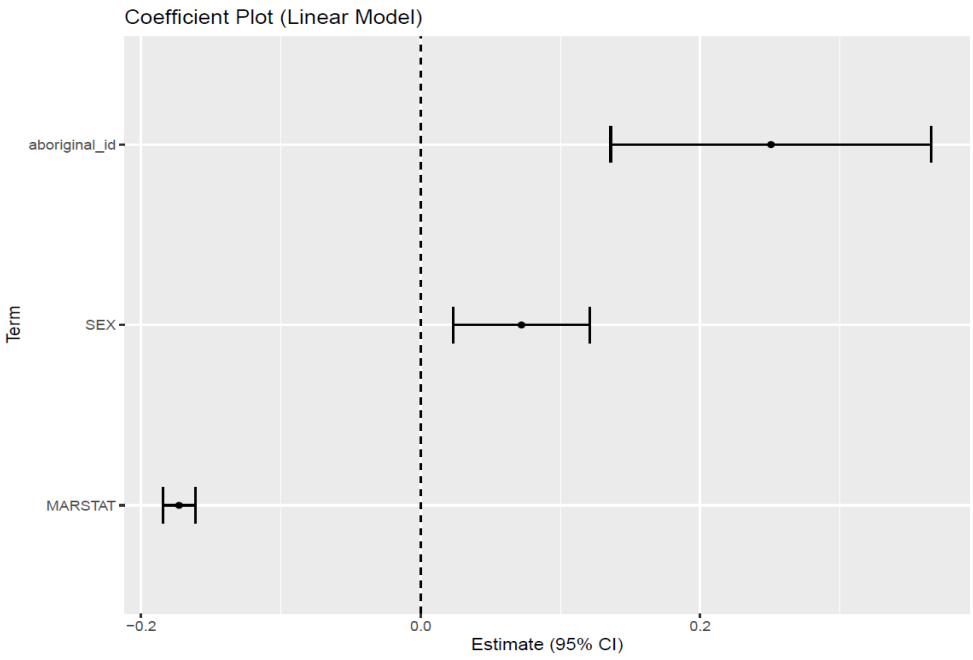
Lanlan's Assumption Checks



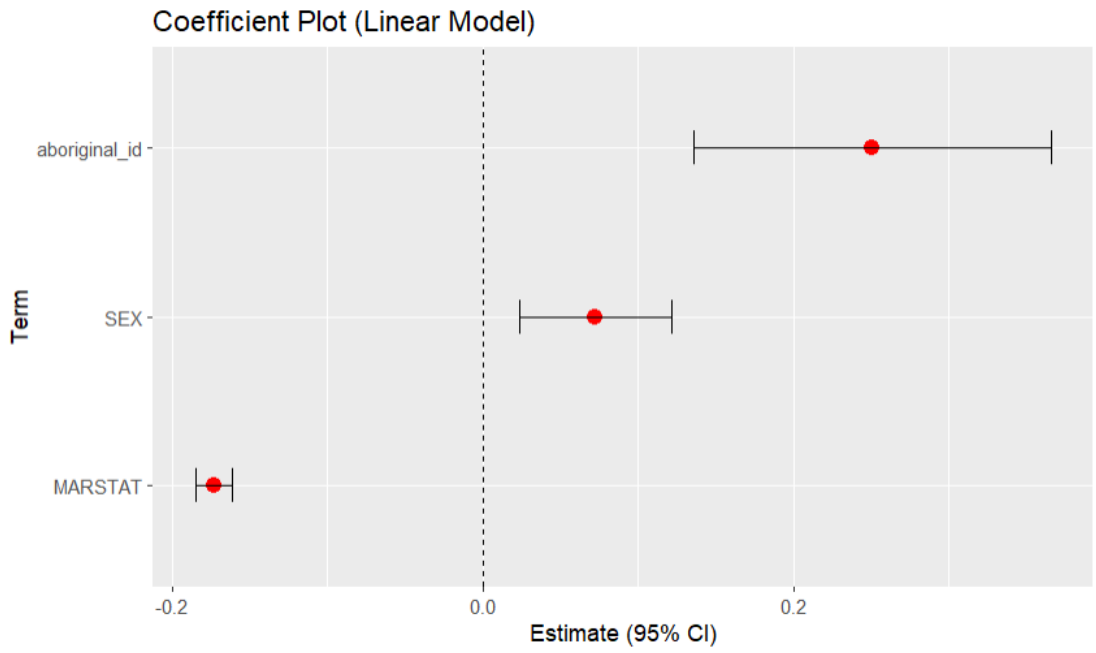
My Reproduced Assumption Checks



Lanlan's Forest Plot



My Reproduced Forest Plot



In-Text Numbers

[X] There are no in-text numbers, or all in-text numbers stem from tables and figures.

[ ] There are in-text numbers, but they are not identified in the code

Classification

The replication successfully reproduced all reported numerical results based on the provided master script. No separate program file was included in the submission; therefore, the replication followed the structure and annotations in the master script to reproduce the analysis. All coefficient estimates, standard errors, t-values, p-values, and model fit statistics were identical to the original output, indicating computational reproducibility.

However, one methodological issue was identified: the marital status variable (MARSTAT) was treated as a numeric variable rather than a categorical factor. Although this did not affect the reproducibility of the numerical outputs, it represents a model specification error under accepted methodological standards, as the predictor should have been factor-coded to allow meaningful group comparisons. Accordingly, this replication is classified as a full reproduction with minor issues, where the issue concerns variable treatment and program completeness rather than code execution or data consistency.

- full reproduction
- [X] full reproduction with minor issues
- partial reproduction (see above)
- not able to reproduce most or all of the results (reasons see above)