Research Replicability Guide

Project: Labour Force Survey (LFS) 2024 — Monthly to Annual Pooling and Multinomial Logit Analysis

Author: *Galiba Zahid*

Last updated: *2025-10-27*

Purpose

This document provides a conceptual overview of how to reproduce the analysis using the 2024 *Labour Force Survey (LFS)*. It outlines the analytical workflow, including:

- Obtaining and preparing the microdata
- Pooling monthly files into a single annual dataset
- Constructing the key analytical variables
- Estimating the multinomial logistic regression model and generating predicted probabilities

All steps are described conceptually rather than through executable code, allowing other researchers to replicate the analysis using **Stata** or comparable statistical software.

Data Collection and Preparation

Downloading the Data

1. Download the publicly available **2024 Labour Force Survey (LFS) microdata** from Statistics Canada: https://www150.statcan.gc.ca/n1/pub/71m0001x/2021001/hist/2024-CSV.zip
2. Extract the ZIP file into a single folder on your computer.
   The folder will contain twelve monthly CSV files and a codebook describing the variables.

File Naming and Organization

Each file corresponds to one month of 2024 and follows a consistent naming convention:
pub0124.csv, pub0224.csv, pub0324.csv, …, pub1224.csv
Store all twelve files together in one folder (e.g., C:/Users/Name/Downloads/2024-CSV/).

Data Pooling Procedure

Combine Monthly Data into a Single Annual Dataset

Each monthly file represents a separate sample of individuals. To create an annual dataset:

1. Import each of the twelve-monthly files in turn.
2. For each dataset, create a variable identifying the month (e.g., 01 for January, 02 for February, etc.).
3. Stack or append all months vertically into one dataset so that every respondent from every month is included.
4. Save the combined dataset as lfs_2024pooled.dta (or in CSV format if preferred).

This step produces a single annual dataset containing all 2024 LFS respondents.

Variable Construction

All derived variables are created from existing variables in the microdata to simplify analysis and enable cross-group comparisons.

1 Family Type (famtype_simple)

Respondents are classified into four family-type categories using marital status (marstat) and the age of youngest child (agyownk).

Category	Definition	Conditions
1	Married/common-law, no children	marstat in {1,2} and agyownk is missing (no children)
2	Married/common-law, with children	marstat in {1,2} and agyownk in {1,2,3,4}
3	Single, no children	marstat in {3,4,5,6} and agyownk is missing
4	Single, with children	marstat in {3,4,5,6} and agyownk in {1,2,3,4}

This classification distinguishes between partnered and single respondents and between those with and without dependent children.

---

2 Age Group (age3)

A three-category variable simplifies the original 12-category age variable (age_12):

Category	Label	Condition
1	Under 25 years	age_12 in {1, 2}
2	25–54 years	age_12 in {3, 4, 5, 6, 7, 8}
3	55 years and over	age_12 in {9, 10, 11, 12}

---

3 Immigrant Status (imm)

The immigrant variable distinguishes immigrants from non-immigrants:

Category	Label	Condition
0	Non-immigrant	immig == 3
1	Immigrant	immig in {1, 2}

---

4 Gender × Immigrant Composite (genderimm)

A four-category composite variable is created by combining gender (original variable) and immigrant status (recoded as per 4.3):

Category	Label	Condition
1	Male non-immigrant	gender == 1 and imm == 0
2	Female non-immigrant	gender == 2 and imm == 0
3	Male immigrant	gender == 1 and imm == 1
4	Female immigrant	gender == 2 and imm == 1

---

5 Education (BA)

Education is simplified into a binary variable:

Category	Label	Condition
0	Below Bachelor’s degree	educ in {0, 1, 2}
1	Bachelor’s degree or higher	educ in {3, 4, 5, 6}

Model Estimation

Multinomial Logit Model

The dependent variable is whylefto — respondents' main reason for leaving their last job.
The model estimates the likelihood of each outcome as a function of demographic and family characteristics, controlling for province, age, and education.
The model is weighted using the LFS person weight variable finalwt.
Model is estimated as follows:

> Model 1: uses the combined genderimm × famtype_simple and age4, BA, and prov (original variable) as controls and applied person level weights.

Predicted Probabilities

Marginal Effects

Predicted probabilities are calculated for each combination of genderimm × famtype_simple across all outcomes of whylefto (coded 0–5).
Each outcome's results are saved separately and then merged into one dataset, pp_all.dta, which contains:

- _m1 = genderimm group
- _m2 = family type group
- _margin = predicted probability
- _se_margin = standard error
- _ci_lb and _ci_ub = lower and upper 95% confidence limits
- outcome = reason for leaving job

Output Files

File	Description
lfs_2024pooled.dta	Combined 2024 microdata across all months
pp_out0.dta – pp_out5.dta	Predicted probabilities for each whylefto category
pp_all.dta	Final merged dataset containing all predicted probabilities and confidence intervals

---

Visualization in R

Required Inputs and Software

The visualization step uses the file pp_all.dta, which contains the predicted probabilities generated in Stata.

The analysis requires R (version 4.x or higher) with the following packages installed:

haven, dplyr, tidyr, ggplot2, scales, RColorBrewer, stringr, ggtext, and grid.

File Setup

The file pp_all.dta must be stored in an accessible directory (e.g., C:/Users/Name/Folder/pp_all.dta).

Structure of pp_all.dta

The dataset contains one observation per combination of gender–immigrant group, family-type group, and outcome.

Key variables include:

Variable	Description
_m1, _m2	Factor codes for gender–immigrant and family-type groups
_margin	Predicted probability
_se_margin	Standard error of predicted probability
_ci_lb, _ci_ub	95 percent confidence interval bounds
outcome	Reason for leaving job (0–5)

These variables are converted to labelled factors and renamed to:

genderimm, famtype_simple, prob, se, ci_lo, ci_hi, and outcome.

Facet Order

The visualization displays six panels arranged from top left to bottom right in the following order:

1. Laid off
2. School
3. Personal / family responsibilities

4. Illness / disability
5. Retired
6. Other reasons

This ordering is fixed through factor-level specification to ensure facet layout matches the order of discussion (not required)

Tile Grid Structure

Each panel presents a 4 × 4 matrix of predicted probabilities.

Rows correspond to gender–immigrant groups

(Male non-immigrant, Female non-immigrant, Male immigrant, Female immigrant).

Columns correspond to family-type groups

(M/CL no kids, M/CL with kids, Single no kids, Single with kids).

Reference Group and Significance Masking

All comparisons are computed relative to a reference subgroup defined as:

Male non-immigrant × Married or Common-Law with no children.

For each outcome, the procedure:

1. Calculates the difference between each cell's predicted probability and the reference cell's probability.
2. Computes a z-statistic for this difference using combined standard errors.
3. Applies a two-sided p-value test; cells with $p \geq 0.05$ are deemed statistically indistinguishable from the reference.

Tiles failing to reach $p < 0.05$ are rendered white with grey text; those meeting the threshold are colored with black text.

This identifies which subgroups differ significantly from the reference within each outcome panel.

Color Scale and Binning

Predicted probabilities are grouped into ten deciles (0–10 %, 10–20 %, … , 90–100 %).

A ten-step YlGnBu palette encodes these bins from light (low) to dark (high).

The legend displays the lower range of deciles (0–70 %) to maintain a compact layout.
Fixed binning ensures consistent interpretation across all panels.

Legend and Caption Formatting

Invisible placeholder keys maintain a stable legend even when certain decile bins contain no observations.
Each tile shows its predicted probability (e.g., 12.3 %).
Black labels denote statistically significant differences; grey labels denote non-significant differences.
A caption beneath the plot documents data source, model, weighting, and significance threshold.

Exporting the Figure

The completed heatmap is exported as a high-resolution image (300 dpi) titled
lfsexit.jpeg and saved in (C:/Users/Name/Folder/
Any valid raster format (e.g., .png, .jpeg) may be used by adjusting the file extension in the save command.
The output directory must exist before saving.

Quality Checks

Reproducible output should display:

- Six panels in the facet order listed in §8.4;
- A 4 × 4 tile grid per panel;
- White tiles for non-significant differences and colored tiles for significant ones;
- A legend labeled "Predicted probabilities" with decile intervals;
- A readable, correctly sized image file that opens without error.

Interpretation

Color intensity reflects the magnitude of the predicted probability.
Tile color and text style jointly indicate statistical significance relative to the reference group.
Comparisons apply within outcomes (each reason for leaving employment) and not across different outcome panels.