Income Gap Between Visible Minorities and Non-visible Minorities

Maxwell Garrett

October 19th 2020

Abstract

Using data from the 2017 General Social Survey performed by Statistics Canada, we look into the relationship between the income of a Canadian individual and the individual being a visible minority. We create a logistic regression model of income in relation to visible minority status and education of the respondents. We show through this model that an individual being a visible minority decreases their likelihood of earning an income greater than or equal to $50,000 while controlling for education.

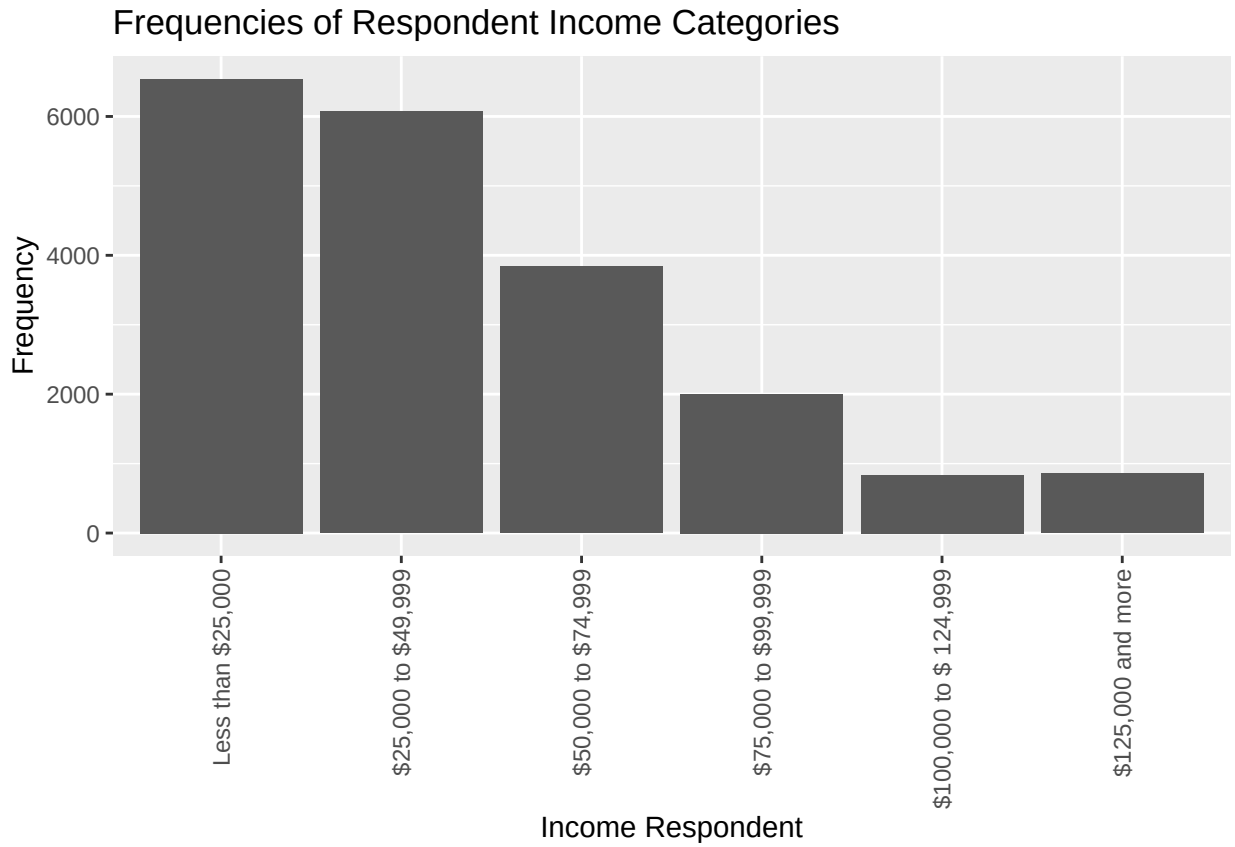
Introduction

In this report, we analyze data from the GSS (General Social Survey) which was collected in 2017 (Statistics Canada 2017). The data provides detailed demographic information on each respondent which is analyzed to see relationships between education, visible minority status, and likelihood of having an income greater than or equal to $50,000. We analyze the relationships between these factors with the expectation that we will discover how visible minority status relates to income while controlling for education. This is an important relationship to observe as ideally the income received by an individual should be independent of whether an individual is a visible minority. We performed a logistic regression analysis and found that an individual who is a visible minority was less likely to earn an income greater than or equal to $50,000 when compared to an individual who was not a visible minority. This result is important as it means there is potential biases in society resulting in an individual not being paid a fair wage for their work.

Data

The data selected for this report is from the General Social Survey collected in 2017 (cycle 31: Families). The data was collected through a stratified random sample (no replacement) of households in Canada (excluding Yukon, Northwest Territories, and Nunavut) through the use of calling telephone numbers attached to a given household (Beaupré 2020). Both cellular and landline phone numbers were included in the random sample (Beaupré 2020). The strata were based on geographic areas in each province (Beaupré 2020). The target population of the survey was all individuals in Canada above the age of fifteen not including institutionalized individuals or residences of Yukon, Northwest Territories, and Nunavut (Beaupré 2020). The frame population was all individuals with phone numbers available to Statistics Canada in the above target population (Beaupré 2020). The sampled population was 20,602 individuals spread across the different strata (Beaupré 2020). The response rate from selected respondents was approximately 52% (Beaupré 2020). If individuals did not consent to be interviewed on first phone call, they were recontacted up to two more times to attempt to be interviewed (Beaupré 2020). An issue with this data collection is that we do not include individuals who for some reason do not have a phone number. This potentially excludes unhoused individuals, and those who live in remote areas. Another issue within this data collection is the lack of recording of reasons respondents provided for not responding. This data could be important as reasons such as lack of time could be associated with those of a specific sub-population which would result in these sub-populations being underrepresented in the survey data.

Figure 1: Bar graph of respondent income categories



The variables used in the following analysis were income level of the respondent, visible minority status, education level of the respondent, and age. Income level of the respondent was used to create a new binary variable, inc_status, indicating if an individual had an income greater than or equal to $50,000 or less than $50,000. Originally the income variable had a category for each income bracket in $25,000 dollar increments up to $125,000 (Statistics Canada 2017). This variable describes the respondent's income which is helpful for the analysis as we are looking for links between the respondent's education and their specific income. The variable inc_status was created to allow for a logistic regression with a binary response variable to be performed, the response variable being inc_status. The value $50,000 was chosen as approximately half of the incomes recorded were above $50,000 while approximately half below, this distribution can be seen in figure 1. The variable visible minority status, vis_minority, was also used in the analysis. This variable was recorded with responses being visible minority, not a visible minority, or don't know (Statistics Canada 2017). The variable was modified by removing all missing values which were designated by the response "Don't know". These unknown values were excluded as we are not sure the visible minority of these individuals and we would not be able to make a relationship between these individuals income and visible minority status. The next variable used was education, this variable was not modified for analysis. This variable indicates the education level that the respondent has received with the following categories "Less than high school diploma or its equivalent", "High school diploma or a high school equivalency certificate", "College, CEGEP or other non-university certificate or diploma", "University certificate or diploma below the bachelor's level", "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)", "Trade certificate or diploma", "University certificate, diploma or degree above the bachelor degree" (Statistics Canada 2017). This variable has the strength of providing many different classifications for an individual's education which would help facilitate accurate responses by participants. The last variable analyzed was age of the respondent as we thought that this could be associated with the income of a respondent, and controlling for age would have allowed us to better isolate the relationship between income and visible minority status. The age variable was not used

in the final model. All four variables chosen to be analyzed have low missing data rates, at 2.2% for visible minority data, 1.6% for education data, 0% for income of the respondent data, and 0% for age data (see figure 6). This made these variables ideal to analyze as they are representative of the sample data.

Model

The relationship between the respondent's income, respondent's education, and visible minority status was modeled using a logistic regression model. The respondent's income was represented using a binary variable indicating if the income was greater than or equal to $50,000 or less than $50,000. This was necessary as the logistic regression model requires that the response variable, respondent's income in this case, is a binary variable. We can write the relationship using the following notation: $inc\_status \sim vis\_minority + education$. This formula indicates that *inc_status*, the binary variable created, is the response variable. In the logistic regression this means, we will be modeling the odds of an individuals income being greater than or equal to $50,000. The variables *vis_minority* and *education* are our predictor variables that will be used to predict the odds of the response value being above or equal to $50,000. This means we can input values of these two variables into the model to get an output odds of the response variable being a value. Our model also takes into account the survey structure used with the target population set as $30,302,287$ and the survey sample size of $20,148$.

The feature of visible minority status was included as we want to view the relationship between a respondent being a visible minority and their income. The feature of education was included to allow us to control for education when viewing the probability of their income. This helps explain variation in income that is not only due to visible minority status. A strength of this model having few variables included is that we limit the number of missing values encountered. As variables would be added into this model, we would likely experience an increase in rows that have a single missing value resulting in a whole row's data being ignored. A model with age included as a predictor along with the above predictors was also produced as an alternative model and can be seen in the results section. This alternative model with age was rejected as age has a high P-value indicating that it did not have a statistically significant relationship with the respondent's income when controlling for the other predictors.

Results

Figure 2: Logistic regression model with formula $inc\_status \sim vis\_minority + education$

term	estimate	std.error	statistic	p.value
(Intercept)	-2.0451	0.0583	-35.1028	0
vis_minorityVisible minority	-0.5596	0.0486	-11.5194	0
educationHigh school diploma or a high school equivalency certificate	1.0372	0.0670	15.4834	0
educationCollege, CEGEP or other non-university certificate or di...	1.5918	0.0660	24.1341	0
educationUniversity certificate or diploma below the bachelor's level	2.0459	0.0950	21.5275	0
educationBachelor's degree (e.g. B.A., B.Sc., LL.B.)	2.4187	0.0675	35.8365	0
educationTrade certificate or diploma	1.6409	0.0790	20.7687	0
educationUniversity certificate, diploma or degree above the bach...	2.9568	0.0784	37.7102	0

Our logistic regression model is specified in figure 2 with its coefficients. Below, we will outline the meaning of the coefficients for each variable's terms. We will begin with the visible minority coefficient as this is of

interest.

We can see that the coefficient for the visible minority category is equal to -0.55956, indicating that the expected odds of having an income above $50,000 for an individual who is a visible minority is $e^(-0.56294) \approx 0.6$ times the odds for an individual who is not a visible minority, controlling for education level. In other words, the predicted probability of an individual having an income above $50,000, controlling for education level, decreases when an individual is a visible minority, and this difference is statistically significant (p-value < 0.05).

Below are the coefficients for the different education categories.

The coefficient for the "High school diploma or a high school equivalency certificate" category is equal to 1.03716, indicating that the expected odds of having an income above $50,000 for an individual who has a high school diploma or equivalent is $e^(1.03716) \approx 2.8$ times the odds for an individual who has an education of less than a high school level, controlling for visible minority status. In other words, the predicted probability of an individual having an income above $50,000, controlling for visible minority status, is higher when an individual has a high school diploma (or equivalency) compared to when an individual has less than a high school education, and this difference is statistically significant (p-value < 0.05).

The coefficient for the "College, CEGEP or other non-university certificate or di.." category is equal to 1.59177, indicating that the expected odds of having an income above $50,000 for an individual who has one of the specified certificates is $e^(1.59177) \approx 4.9$ times the odds for an individual who has an education of less than a high school level, controlling for visible minority status. In other words, the predicted probability of an individual having an income above $50,000, controlling for visible minority status, is higher when an individual has one of the specified certificates compared to when an individual has less than a high school education, and this difference is statistically significant (p-value < 0.05).

The coefficient for the "University certificate or diploma below the bachelor's level" category is equal to 2.04587, indicating that the expected odds of having an income above $50,000 for an individual who has one of the specified certificates is $e^(2.04587) \approx 7.7$ times the odds for an individual who has an education of less than a high school level, controlling for visible minority status. In other words, the predicted probability of an individual having an income above $50,000, controlling for visible minority status, is higher when an individual has one of the specified certificates compared to when an individual has less than a high school education, and this difference is statistically significant (p-value < 0.05).

The coefficient for the "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" category is equal to 2.41869, indicating that the expected odds of having an income above $50,000 for an individual who has a bachelor's degree is $e^(2.41869) \approx 11.2$ times the odds for an individual who has an education of less than a high school level, controlling for visible minority status. In other words, the predicted probability of an individual having an income above $50,000, controlling for visible minority status, is higher when an individual has a bachelor's degree compared to when an individual has less than a high school education, and this difference is statistically significant (p-value < 0.05).

The coefficient for the "Trade certificate or diploma" category is equal to 1.64091, indicating that the expected odds of having an income above $50,000 for an individual who has a trade certificate or diploma is $e^(1.64091) \approx 5.2$ times the odds for an individual who has an education of less than a high school level, controlling for visible minority status. In other words, the predicted probability of an individual having an income above $50,000, controlling for visible minority status, is higher when an individual has a trade certificate or diploma compared to when an individual has less than a high school education, and this difference is statistically significant (p-value < 0.05).

The coefficient for the "University certificate, diploma or degree above the bach..." category is equal to 2.95684, indicating that the expected odds of having an income above $50,000 for an individual who has an education above a bachelor's degree is $e^(2.95684) \approx 19.2$ times the odds for an individual who has an education of less than a high school level, controlling for visible minority status. In other words, the predicted probability of an individual having an income above $50,000, controlling for visible minority status, is higher when an individual has an education above a bachelor's degree compared to when an individual has less than a high school education, and this difference is statistically significant (p-value < 0.05).

Figure 3: Logistic regression model with formula $inc\_status \sim vis\_minority + education + age$

term	estimate	std.error	statistic	p.value
(Intercept)	-2.0408	0.0755	-27.0199	0.0000
vis_minorityVisible minority	-0.5602	0.0493	-11.3513	0.0000
educationHigh school diploma or a high school equivalency certificate	1.0368	0.0670	15.4746	0.0000
educationCollege, CEGEP or other non-university certificate or di...	1.5913	0.0659	24.1451	0.0000
educationUniversity certificate or diploma below the bachelor's level	2.0457	0.0950	21.5360	0.0000
educationBachelor's degree (e.g. B.A., B.Sc., LL.B.)	2.4181	0.0675	35.8257	0.0000
educationTrade certificate or diploma	1.6405	0.0789	20.7877	0.0000
educationUniversity certificate, diploma or degree above the bach...	2.9565	0.0783	37.7407	0.0000
age	-0.0001	0.0009	-0.0837	0.9333

The model specified in figure 3 is the model produced that includes age as a predictor as well as visible minority status and education level. The response variable is also income category being above or below $50,000. The goal of this model is to see if age is an important predictor to include. The coefficient for age's P-value in this model is approximately 93%. Since this probability is above the significance level of 5%, we can say that the coefficient for age is not statistically significant. For this reason, our final model does not include age as a predictor as it's relationship with the income status variable was not statistically significant.

Figure 4: Logistic regression model table controlling for education

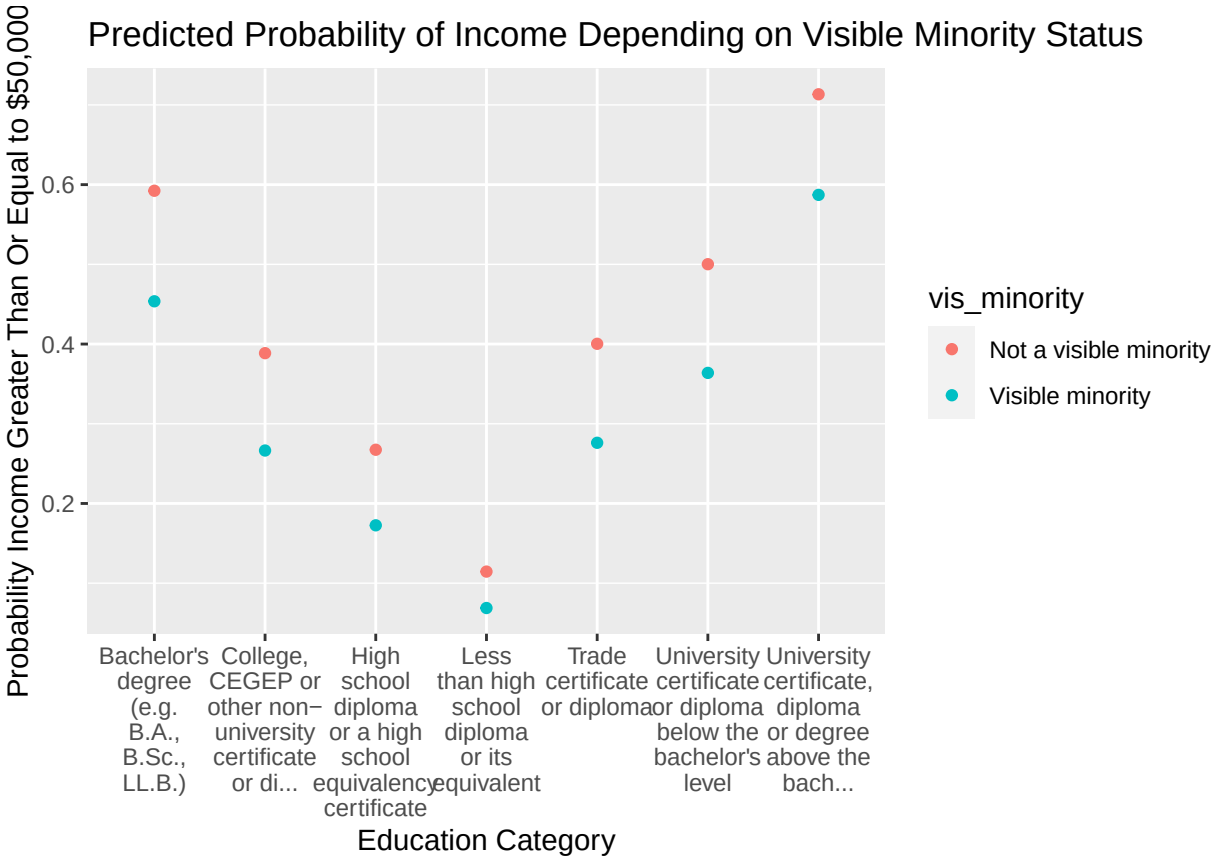
Visible Minority Status	Education	P(Inc.>=$50,000)
Not a visible minority	Less than high school diploma or its equivalent	0.11454676
Not a visible minority	High school diploma or a high school equivalency certificate	0.26738060
Not a visible minority	College, CEGEP or other non-university certificate or di...	0.38856487
Not a visible minority	University certificate or diploma below the bachelor's level	0.50018826
Not a visible minority	Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	0.59232220
Not a visible minority	Trade certificate or diploma	0.40030208
Not a visible minority	University certificate, diploma or degree above the bach...	0.71335332
Visible minority	Less than high school diploma or its equivalent	0.06883807
Visible minority	High school diploma or a high school equivalency certificate	0.17257123
Visible minority	College, CEGEP or other non-university certificate or di...	0.26641105
Visible minority	University certificate or diploma below the bachelor's level	0.36382361
Visible minority	Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	0.45363703
Visible minority	Trade certificate or diploma	0.27612475
Visible minority	University certificate, diploma or degree above the bach...	0.58714291

The above table (figure 4), displays all the potential values for the visible minority and education variables as well as the predicted probability of income being greater than or equal to $50,000 given two input values. For each value of education, we can see that the visible minority status modified the predicted probability.

The predicted probabilities above are calculated using the model specified in figure 2. We will go through these differences below.

An individual with "High school diploma or a high school equivalency certificate" who is a visible minority is approximately 4.57% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority. An individual with "High school diploma or a high school equivalency certificate" who is a visible minority is approximately 9.48% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority. An individual with "College, CEGEP or other non-university certificate or di..." who is a visible minority is approximately 12.21% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority. An individual with "University certificate or diploma below the bachelor's level" who is a visible minority is approximately 13.64% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority. An individual with "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" who is a visible minority is approximately 13.87% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority. An individual with "Trade certificate or diploma" who is a visible minority is approximately 12.42% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority. An individual with "University certificate, diploma or degree above the bach..." who is a visible minority is approximately 12.62% less likely to have an income greater than or equal to $50,000 compared to a non-visible minority.

Figure 5: Logistic regression model plotted controlling for education



In the above plot (figure 5), we visually illustrate the differences between the probability of an income greater than or equal to $50,000 depending on whether an individual is a visible minority while controlling for education. We can see that for all education categories, a visible minority has a significant lower probability of having an income greater than or equal to $50,000.

Figure 6: Percent of missing values for variables used.

term	missing_perc
vis_minority	0.0220367
education	0.0165518
income_respondent	0.0000000
age	0.0000000
occupation	0.3541889

The above table (figure 6) shows the percent of missing values for each variable that was used in the final analysis. We can see that for each variable the percent of values that were missing was quite low with the highest proportion of missing values for a variable being 2.2%. We count a value as missing if it was answered as "Don't know" or was recorded as a N/A value in the data. The occupation variable has a high missing data rate (35%) and therefore was not selected to be used in this analysis.

Discussion

Through our analysis above, we've come to some conclusions about the relationship between visible minority status and the probability of income being greater or equal to $50,000. To begin, we can see in the results that education did have a statistically significant relationship with income while controlling for education. Education is used to explain some of the variation we see in income that might not be related to visible minority status. We saw that the expected odds of having an income above $50,000 for an individual who is a visible minority is $e^(- 0.56294) \approx 0.6$ times the odds for an individual who is not a visible minority, when controlling for education level. This result indicates that an individual who is a visible minority was less likely to earn above or equal to $50,000 dollars compared to an individual who is not a visible minority. This highlights the inequalities in income we expected in our research. Ideally we would have hoped an individual with similar education level would have a non-statistically significant income difference compared to an individual who is not a visible minority. We can further see in the results that the income disparity is much higher for some education categories compared to others. For example, individuals with a bachelor's degree who are a visible minority are 13.87% less likely to receive an income above or equal to $50,000 dollars compared to an individual who is not a visible minority with a bachelor's degree. We see throughout all education categories that there is a significant income disparity with the probability of a visible minority earning above or equal to $50,000 dollars being less than that of their non-visible minority counterpart.

Weaknesses

There are several weakness to our analysis. One weakness is in how the income categories were created for analysis. The income categories were reduced from six separate categories to two to allow the creation of a binary variable. This was necessary to use a logistic regression model on the data as it requires a binary variable. Reducing from six categories to two categories resulted in lost data from the categories available which could have been helpful for analysis and possibly shown more important trends. In future, a mulinomial logistic regression could be performed which is a regression that allows for more than two categories in the response variable ("Multinomial Logistic Regression" n.d.). Performing this analysis would provide us more insight into the relationship between visible minority status and different income levels. Another weakness is the use of only two predictor variables in the model. There could potentially be other predictor variables that are important in the relationship described by the model. These values could change the coefficient or significance of the visible minority variable. This could be solved by performing variable selection in future, which would allow us to systematically chose which variables out of all available are significant in the relationship ("Variable Selection Methods" n.d.).

Next Steps

There are many potential next steps to this study. The first next step would likely be to perform analysis on this data with more variables considered. This would help explain more of the variation we see in income, and make the variation that is explained by visible minority status clearer. Another step would be to perform an analysis that includes all the income categories available. This would likely be a multinomial logistic regression as mentioned in the weaknesses section. The inclusion of all income categories would give us more detail on the relationships between visible minority status and the other income brackets available. Another step would be to perform a follow-up survey specifically on our areas of interest. The follow-up survey would be ideal if it could be distributed in a similar manner as this one, as this survey had a large amount of data spread out across Canada. We would be interested in having income reported as an exact number instead of a bracket in this survey, as this would give much more data to analyze when looking for relationships between predictors and income. It would also be helpful to retrieve more occupation data as that could be helpful in controlling more of the factors related to income. The GSS survey unfortunately had many missing values for occupation (see figure 6) which made this variable unideal to use in analysis.

References

Alexander, Rohan, and Sam Caetano. 2020. "GSS Cleaning Code for 2017 Data."

Beaupré, Pascale. 2020. *Cycle 31 : Families Public Use Microdata File Documentation and User's Guide*. Statistics Canada. http://www.chass.utoronto.ca/.

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9 (1): 1–19.

———. 2010. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley; Sons.

———. 2020. "Survey: Analysis of Complex Survey Samples."

"Multinomial Logistic Regression." n.d. Accessed October 19, 2020. https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2020. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Statistics Canada. 2017. "2017 General Social Survey: Families Cycle 31." http://www.chass.utoronto.ca/.

"Variable Selection Methods." n.d. Accessed October 19, 2020. https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.