Analysis on the impact of in-person socialization on individuals' mental health conditions

Guemin Kim

December 22, 2020

Code and data supporting this analysis is available at: https://github.com/Guemin/Final_Project

Abstract

In such an unprecedented time caused by the COVID-19 pandemic, it has become more apparent that socializing plays a significant role in human's life. The main purpose of this analysis is to find and analyze the causal relationship between face-to-face socialization and individuals' mental health conditions by comparing the mental health scores for those who spend more time socializing in person than online and those who do not, by using propensity score matching. The 2015 General Social Survey(GSS) data obtained from the CHASS website is used to identify some potential factors affecting one's mental health conditions, as well as to find the relationship between in-person socialization and mental health scores of the respondents. The results from the propensity score analysis indicate that face-to-face socialization causes an individual's mental health score to increase.

Keywords - In-person Socializaing, Mental Health, Propensity Score Matching, Causal Inference, Observational Study

Introduction

Coronavirus pandemic has brought a devastating impact on individuals and communities across the world. In particular, due to the lockdowns and social distancing, the number of in-person meetings has reduced significantly, and many of them are either postponed or replaced with virtual gatherings instead.

In such unprecedented times caused by the COVID-19 crisis, the regulations and guidelines to limit social contact between people are crucial in order to slow down the spread of the virus; however, at the same time, there is a high cost associated with the restrictions on social distancing, especially for those who experience isolation and loneliness which could ultimately affect their mental health conditions. Although the virtual meeting has emerged as an alternative way to communicate with others, it is not enough to ease emotional distress and loneliness one might experience during the pandemic.

Hence, this paper will thoroughly investigate and analyze the impact of in-person socializing on individuals' mental health conditions by comparing the mental health scores for those who spend more time socializing in person than socializing using technology and those who do not. More specifically, propensity score matching will be used to assign the probability of preferring in-person socialization over technology-based socialization to each observation, then, the observations in treatment[1] group will be matched with the observations in control group based on the similar propensity scores, and finally, the effect of in-person socialization on individuals' mental health scores will be examined using a multiple linear regression model.

The General Social Survey conducted in 2015 will be used throughout the analysis; more information regarding the data, models and the techniques using propensity score is described in the Methodology section. Moreover,

[1]The treatment group in the analysis contains those who spend more time socializing in person than socializing using technology

the results from the propensity score matching and further discussions on the overall outcomes of the study will be provided in the Results and Discussion sections, respectively.

Methodology

Data

The data set chosen for the analysis contains responses for the 2015 General Social Survey(GSS) on Canadians' time use, obtained from the CHASS website. Since the main objective of the GSS is to monitor the living conditions and well-being of individuals across the ten provinces in Canada, the target population includes all individuals, 15 years old or older, living in the ten provinces of Canada excluding Yukon, Northwest Territories and Nunavut. Also, since the responses for the survey is collected via telephone calls, the frame population is those who have registered telephone numbers in use within the ten provinces and are on the lists available to Statistics Canada; therefore, the sample population is those who were reached and participated in the survey via telephone calls [2].

For this analysis, a new data set is created by cleaning and reformatting the 2015 GSS data. In the new data set, only the variables that are necessary[3] for the analysis are included, and the non-responses, as well as some variables with too many missing values, are removed from the data. Most importantly, throughout the data cleaning process, a new variable was created to separate individuals who spend more time socializing in-person than socializing using technology versus those who do not into treatment and control groups[4]. This is a list of variables included in the data:

Table 1: Variables in the data

Variables
id
age_group
sex
num_children
prefer_in_person
self_rated_mental_health
work_hrs
outdoor_sports
chores_dur
sleepdur
dur_at_home
self_development

As shown above, the cleaned data contains twelve variables that are selected from the 2015 GSS data. Since the purpose of the data cleaning process was to identify and include some potential predictor variables that could affect individuals' mental health scores, some potential factors - such as the amount of time spent working, sleeping, socializing and doing some other activities like outdoor sports or household chores - are included along with some basic demographic information.

[2] According to Statistics Canada, the overall response rate for the 2015 General Social Survey was 38.2%.

[3] What it means by "necessary" is that the variables included in the new data set have potentials to provide useful information in predicting the outcome of this study.

[4] Around 30.8% of the total respondents are placed in the treatment group and the other 69.2% in the control group. On the other hand, individuals who spend the same amount of time on in-person and technology-based socializations are placed into the control group.

One of the key features to notice in the data set is that there are many variables indicating the time use of the respondents on certain activities. These variables provide useful information regarding the daily routine of each respondent; however, unlike basic demographics such as age group or sex which do not change frequently, the amount of time spent on doing certain activities could be changed from day to day[5].

Another weakness of those time-use variables in the data set is that some of them contain too many responses indicated as "No time spent doing this activity"; for example, in the case of the time-variables called 'outdoor_sports' and 'self_development', 97.7% and 99.0% of the respondents respectively answered that they did not spend time doing those activities. Since it is unlikely that such variables will provide useful information in predicting the outcome, they should be removed from the data.

After removing the two variables, this is what the data looks like:

Table 2: Preview of the data

id	age_group	sex	num_children	prefer_in_person	work_hrs	chores_dur	sleepdur	dur_at_home	self_rated_mental_health
1	55 to 64 years	M	0	0	0	180	510	1180	3
2	55 to 64 years	M	0	1	0	570	420	1440	4
3	45 to 54 years	F	3	0	480	135	570	1350	4
4	65 to 74 years	F	0	1	20	20	510	455	4
5	15 to 24 years	M	2	0	530	0	435	870	3
6	15 to 24 years	M	1	0	0	60	635	1140	4

[Descriptions on the variables are provided in the footnote[6].]

Since the purpose of this study is to analyze the causal relationship between in-person socialization and mental health conditions, the outcome variable is the respondent's self-rated mental health score, and as the variable name suggests, it takes self-rated scores in a range of 1 to 5 from each respondent; the higher the score, the better the mental health condition.
On the other hand, the main predictor variable that is going to be used in the propensity score calculation is 'prefer_in_person'; this is a dummy variable with values recorded as either 0 or 1 depending on whether or not the respondent spent more time socializing in person than using technology.
Other potential factors such as age group, sex, number of children, work hours, sleep duration, duration at home, and the amount of time spent doing household chores will be the predictor variables for both 'prefer_in_person' and 'self_rated_mental_health' in the propensity score analysis.

[5]Especially, in such an uncertain time caused by the pandemic, the amount of time people spend doing outdoor activities could decrease drastically whereas the amount of time people spend at home increases significantly.

[6]Variable Description:
* age_group is divided into 7 different groups: "15 to 24 years", "25 to 34 years", "35 to 44 years", "45 to 54 years", "55 to 64 years", "65 to 74 years", and "75 years and older".
* sex indicates "M" for male and "F" for female.
* num_childen indicates the number of children in the respondent's household.
* prefer_in_person indicates whether or not the respondent spent more time on in-person socialization than on socalization using technology.
* each of the work_hrs, chores_dur, sleepdur, and dur_at_home indicates the amount of time(in minutes) respondent spent working, doing chores, sleeping and staying at home, respectively.
* self_rated_mental_health represents the mental health score rated by the respondent.

Before moving on to the next section, verifying whether there is any evidence of multicollinearity in the data is necessary. This is because, multicollinearity could increase the variance of the regression estimates, ultimately making them unstable and difficult to interpret.
One way to check for the multicollinearity between the predictor variables is by reviewing the correlation matrices:

Table 3: Correlations between the quantitative variables

	self_rated_mental_health	num_children	work_hrs	chores_dur	sleepdur	dur_at_home
self_rated_mental_health	1	-0.0143	0.0092	0.0205	-0.0371	-0.0251
num_children		1	0.1152	-0.0185	-0.0581	-0.068
work_hrs			1	-0.3451	-0.3197	-0.4994
chores_dur				1	-0.0259	0.363
sleepdur					1	0.3063
dur_at_home						1

The correlation matrix above(Table 3) contains the correlation coefficients between quantitative variables in the data. As it is shown in the matrix, none of the coefficients are close to 1 or -1, and this indicates that there is no evidence of the presence of multicollinearity in the data.

(Note: Visualization of the data is not provided since most of the variables in the data are either categorical or discrete variables.)

Model

As already mentioned in the introduction, the goal of this analysis is to study the causal relationship between in-person socialization and mental health conditions using propensity score matching. To be more specific, by using R software, a logistic regression model will be used to predict whether or not an individual prefers in-person socialization over technology-based socialization, then using the logistic model, the probability of being in the treatment group will be assigned to each observation. After assigning the propensity scores, the observations in the treatment group[7] will be matched with the observations in the control group based on the similar propensity scores, and finally, the effect of in-person socialization on individuals' mental health scores will be examined by using a multiple linear regression model.
Here, the propensity score matching technique is used, because the inference on the causal relationship between in-person socialization and mental health condition should be made using the observational data. In many of the cases, experiments, in which the observations are randomly assigned to treatment or control groups, are used to identify causal relationships between two variables. Accounting for this, the GSS data is not suitable for finding the causal relationship between in-person socialization and mental health condition; however, the propensity score matching allows for balancing the treatment and control groups, so that a causal inference can be made using the observational data.

Model selection

As mentioned in the Data section, some of the potential predictor variables are chosen from the original data set and they were included in the new data during the data cleaning process. However, since not all of the selected variables may affect mental health conditions, the significant variables that provide useful information in predicting the mental health score will be verified using the Akaike Information Criterion(AIC). In particular, a backward elimination method[8] with AIC will be used to find the optimal model that well predicts the mental health score.

[7]As mentioned in the previous section, the treatment group contains respondents who tend to spend more time socializing in person than using technology.

[8]The backward elimination will use the full model which contains all predictor variables available from the data and remove one predictor variable with the largest p-value at a time until the variables left in the model are not redundant. the full model contains all of the seven predictors in the data - age group, sex, number of children, duration at home, work hours, sleep duration, and the time spent doing chores.

[Note, as an alternative, other selection criterions such as corrected AIC, BIC, or adjusted R-squared could be used to find an optimal model, and they could result in different models.]

Following is the optimal model obtained from the backward elimination using AIC:

self_rated_mental_health = $\beta_0 + \beta_1 * X_{age\_group} + \beta_2 * X_{sex} + \beta_3 * X_{num\_children} + \beta_4 * X_{dur\_at\_home} + \beta_5 * X_{work\_hrs} + \beta_6 * X_{chores\_dur} + \beta_7 * X_{prefer\_in\_person}$

where β_0 is the intercept parameter and $\beta_1, \ ...., \ \beta_7$ are the slope parameters for the predictors.

As it is shown in the result of the backward elimination, the predictor variable that indicates the sleep duration of the respondent has been removed from the full model.

Before fitting the optimal model obtained by using AIC, the following logistic regression model, which predicts whether an individual prefers to socialize in person than using technology, will be fitted:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_{age\_group} + \beta_2 * X_{sex} + \beta_3 * X_{num\_children} + \beta_4 * X_{dur\_at\_home} + \beta_5 * X_{work\_hrs} + \beta_6 * X_{chores\_dur}$

where p is the probability that an individual would prefer to socialize in person than using techonology, $log(\frac{p}{1-p})$ is the corresponding log odds, β_0 is the intercept parameter and $\beta_1, \ ...., \ \beta_6$ are the slope parameters for the predictors.

Then, with the logistic regression model, the propensity scores will be calculated and they will be used to create matches between individuals in the treatment(who prefer in-person socializing) and control groups. Finally, using a reduced data set that only contains the matched-observations, the effect of in-person socialization on the mental health score will be examined by fitting the multiple regression model obtained by using AIC.

Model Diagnostics

To make sure that the fitted models are valid, the diagnostics of the models should be thoroughly discussed. In particular, for the logistic regression model, there are several assumptions that must be satisfied:
- First, the response variable in a logistic regression model must be binary.
- Second, the sample size must be large enough.
- Lastly, the multicollinearity among the predictor variables should not be high.

In the case of the logistic regression model in this analysis, the first two assumptions are satisfied; the response variable consists of two levels(either 0 or 1) and there are 16,669 observations in the data.
In order to check whether the last assumption is satisfied or not, the variance inflation factor will be calculated:

Table 4: Variance Inflation Factors for predictors

predictors	VIF
age_group	1.713328
sex	1.024954
num_children	1.381999
chores_dur	1.223239
dur_at_home	1.439308
work_hrs	1.489888

Since none of the predictors in Table 4 have VIF greater than the cutoff of 5, there is no evidence of multicollinearity between the predictor variables; hence, the last assumption is also satisfied.

Results

Table 5 below gives the result from the logistic regression model which predicts the probability that an individual would prefer in-person socialization over technology-based socialization:

Table 5: Logistic Regression Result

	Estimates
(Intercept)	**0.8171 *** (0.0916)**
as.factor(age_group)25 to 34 years	**0.2145 ** (0.0827)**
as.factor(age_group)35 to 44 years	**0.3066 *** (0.0825)**
as.factor(age_group)45 to 54 years	**0.3620 *** (0.0792)**
as.factor(age_group)55 to 64 years	**0.3845 *** (0.0754)**
as.factor(age_group)65 to 74 years	**0.2642 *** (0.0778)**
as.factor(age_group)75 years and older	**0.3007 *** (0.0845)**
as.factor(sex)M	**-0.1108 ** (0.0355)**
num_children	-0.0109 (0.0275)
chores_dur	**-0.0011 *** (0.0001)**
dur_at_home	**-0.0013 *** (0.0001)**
work_hrs	**-0.0026 *** (0.0001)**
Num obs.	16669

*** p < 0.001; ** p < 0.01; * p < 0.05.

As the table indicates, all variables except for the number of children in the respondent's household have smaller p-values than the significance level of 0.05. This provides evidence that the number of children in the respondent's household is not a significant variable whereas the other variables are significant in predicting whether or not an individual prefers to socialize in person rather than using technology.

The formula for the fitted logistic model is:

$log(\frac{p}{1-p}) = 0.8171 + 0.2145 * X_{age\_group_1} + 0.3066 * X_{age\_group_2} + 0.3620 * X_{age\_group_3} + 0.3845 * X_{age\_group_4} + 0.2642 * X_{age\_group_5} + 0.3007 * X_{age\_group_6} - 0.1108 * X_{sex_M} - 0.0109 * X_{num\_children} - 0.0011 * X_{chores\_dur} - 0.0013 * X_{dur\_at\_home} - 0.0026 * X_{work\_hrs}$

where $log(\frac{p}{1-p})$ indicates the log odds of being treated as an individual who prefers in-person socialization. [Detailed descriptions on the predictor variables are provided in the footnote[9].]

[9]* $X_{age\ group_1}$, ..., $X_{age\ group_6}$ represent the age groups of the respondent from "25 to 34 years" to "75 years and older".
* X_{sex_M} is a dummy variable which indicates whether the respondent is male or not.
* $X_{dur\_at\_home}$ indicates the amount of time the respondent spent at home.
* X_{work_hrs} indicates the work hours of the respondent.
* $X_{num\_children}$ indicates the number of children in the respondent's household.
* X_{chores_dur} indicates the amount of time the respondent spent doing household chores.

Moving on, here is the result of the propensity score regression:

Table 6: Propensity Score Regression Result

	Estimates
(Intercept)	**3.6596 *** (0.0472)**
as.factor(age_group)25 to 34 years	0.0048 (0.0447)
as.factor(age_group)35 to 44 years	-0.0606 (0.0445)
as.factor(age_group)45 to 54 years	-0.0476 (0.0430)
as.factor(age_group)55 to 64 years	**0.1474 *** (0.0406)**
as.factor(age_group)65 to 74 years	**0.3254 *** (0.0419)**
as.factor(age_group)75 years and older	**0.2441 *** (0.0455)**
as.factor(sex)M	**0.0507 ** (0.0193)**
dur_at_home	**-0.0001 *** (0.0000)**
work_hrs	**0.0001 * (0.0001)**
num_children	**0.0510 *** (0.0151)**
chores_dur	**0.0002 * (0.0001)**
prefer_in_person	**0.0571 ** (0.0188)**
Num.obs	10260
F statistic	16.9944
P-value	0.0000

*** p < 0.001; ** p < 0.01; * p < 0.05.

The fitted model based on the estimates provided in the table is:

$$\hat{self\_rated\_mental\_health} = 3.6596 + 0.0048 * X_{age\_group_1} - 0.0606 * X_{age\_group_2} - 0.0476 * X_{age\_group_3} + 0.1474 * X_{age\_group_4} + 0.3254 * X_{age\_group_5} + 0.2441 * X_{age\_group_6} + 0.0507 * X_{sex_M} - 0.0001 * X_{dur\_at\_home} + 0.0001 * X_{work\_hrs} - 0.0510 * X_{num\_children} + 0.0002 * X_{chores\_dur} + 0.0571 * X_{prefer\_in\_person}$$

[Detailed descriptions on the predictor variables are provided in the footnote[10]]

As it is shown in Table 6, there are some significant predictor variables with p-values less than 0.05 such as variables indicating the age groups "55 to 64 years", "65 to 74 years" and "75 years and older", male, duration at home, work hours, number of children in the household, amount of time spent doing household

[10]* $X_{age\ group_1}$, ..., $X_{age\ group_6}$ represent the age groups of the respondent from "25 to 34 years" to "75 years and older".
* X_{sex_M} is a dummy variable which indicates whether the respondent is male or not.
* $X_{dur\_at\_home}$ indicates the amount of time the respondent spent at home.
* X_{work_hrs} indicates the work hours of the respondent.
* $X_{num\_children}$ indicates the number of children in the respondent's household.
* X_{chores_dur} indicates the amount of time the respondent spent doing household chores.
* $X_{prefer\_in\_person}$ indicates whether the respondent prefers to socialize in person than online.

chores, and most importantly, the treatment variable indicating whether or not the respondent prefers to socialize in person rather than using technology.

Discussion

Using the 2015 General Social Survey data, the propensity score matching was implemented to examine the effect of in-person socialization on individuals' mental health conditions. Since this study is based on the observational data, the observations are not randomly assigned to the treatment group, and therefore, it may have selection bias. For this reason, the propensity score matching, which could eliminate a great portion of bias and balance the treatment and control groups, was implemented.

In the "Model" Section, a logistic regression model was fitted to predict the probability of preferring in-person socialization over technology-based socialization and it was used to create the propensity scores. The propensity scores created from the logistic model were used to create matches between the treatment and control groups, where the treatment group includes those who prefer to socialize in person than online and the control group contains those who spend more time socializing using technology or do not socialize at all. Then, each observation in the treatment group was matched with the observation in the control group based on similar propensity scores, and finally, the effect of in-person socialization on individuals' mental health conditions was analyzed using a multiple linear regression model.

Here is the average mental health scores before and after matching:

Table 7: Number of observations and mean mental health scores before and after matching

	Before Matching		After Matching	
prefer_in_person	Num.obs	Mental_health_score	Num.obs	Mental_health_score
0	11539	3.700061	5130	3.704678
1	5130	3.762183	5130	3.762183

Table 7 above shows that the numbers of observations are the same in the treatment and control groups after the matching; this is natural since the data set obtained after the one-to-one matching only contains observations that are matched. Another thing to notice in the table is that the average mental health score in the control group has increased after the matching. This indicates that the mental health scores for the treatment and control groups are more in line after the matching.
Overall, the difference in the average mental health scores between the treatment and control group provides evidence that those who spend more time socializing in person are more likely to have better mental health conditions than those who do not.

This finding is also supported by the results of the regression models in the previous section. To begin with the result from the logistic regression, the small p-values for the predictor variables indicate that they are significant variables in predicting the probability of preferring in-person socialization over technology-based socialization.
Furthermore, the estimates of the significant variables in Table 5 are interpreted as, for every additional unit increase in each predictor variable, the log odds of preferring in-person socialization over technology-based socialization will either increase or decrease, according to the corresponding estimate[11].
(Note that when the interpretation on the estimate of each predictor is made, other predictor variables are assumed to be constant.)

However, in a propensity score analysis, what is much more important than having a logistic regression model that well predicts whether an individual is in the treatment group or not, is including the predictor variables

[11] Only the estimates of the "statistically significant" variables are interpreted.

in the logistic regression model that are correlated with the outcome variable. This is the reason why the backward elimination method with AIC was used to find the optimal propensity score regression model[12].

As already mentioned in the "Result" section, the propensity score regression contains some significant predictor variables with p-value less than 0.05 - variables indicating the age groups "55 to 64 years", "65 to 74 years" and "75 years and older", male, duration at home, work hours, number of children in the household, amount of time spent doing household chores, and most importantly, the treatment variable indicating whether or not the respondent prefers to socialize in person rather than using technology.

Each of the significant variables in Table 6 is interpreted as follows:
* First, note that the interpretations on each predictor are made, assuming that the other predictor variables are constant.
- The mental health score is higher for those in the age groups "55 to 64 years", "65 to 74 years" and "75 years and older" by 0.1474, 0.3254, 0.2441 units respectively, assuming that other predictor variables are constant.
- The mental health score is higher for male respondents by 0.0507 units.
- The mental health score is decreased by 0.0001 units as the time spent at home increases by 1 minute.
- The mental health score is increased by 0.0001 units as the work hours increases by 1 minute.
- The mental health score is increased by 0.0510 units as the number of children increases by 1 unit.
- The mental health score is increased by 0.0002 units as the time spent doing household chores increases by 1 minute.
- The mental health score is higher for those who prefer in-person socialization over technology-based socialization by 0.0571 units.

Overall, from the result of the propensity score regression, it is evident that people who spend more time socializing in person are more likely to have better mental health conditions than those who spend more time socializing using technology or do not spend time on socializing at all. Furthermore, the potential factors such as age group, sex, number of children in the household, work hours, amount of time spent at home, and the amount of time spent doing household chores are also affecting one's mental health conditions.

Weaknesses

There is no such thing as a perfect model. Every model contains biases and has its weaknesses.
One of the weaknesses in this study is regarding the model selection.
In the "Model" section, an optimal model was selected as a propensity regression model that predicts the mental health score. Here, the backward elimination method with AIC was used to remove redundant predictors. However, relying too much on the result from a single selection method is unlikely to be "universally successful"(Brewer) since selection methods using AIC, BIC, corrected AIC or adjusted R-squared have different standards for choosing a model[13]. Therefore, as the next step, comparing the results from at least two selection methods - using any of the adjusted R-squared, AIC, corrected AIC, or BIC - is recommended.

Another potential weakness in the study is regarding the interpretation associated with the outcome variable, mental health score. In the 2015 GSS data, the mental health score is recorded as an integer value from a scale of 1 to 5. Based on these values, a propensity score regression was fitted and interpretations were made in the "Discussion" section; however, it might not be appropriate to give interpretations on the mental health score in the first place since it is ambiguous what a "1 unit increase in mental health score" actually means. Also, since the mental health score in the 2015 GSS is a self-rated score, the values could be recorded based on the different standards of the respondents, and therefore, they might not be a reliable measurement to analyze.

Lastly, separating the respondents into either the treatment or control groups could also contain some potential biases. When the treatment variable that indicates whether an individual spent more time on in-person socialization was created, only those who spent a greater amount of time on in-person socialization than technology-based socialization was assigned to the treatment group; this means that those who spent more time using technology to socialize or those who do not socialize at all are placed into the control group.

[12]The backward elimination method removes predictors that are redundant from the model.

[13]For example, a model selection strategy using the adjusted R-squared chooses a model that maximizes the adjusted R-squared whereas a strategy using AIC chooses a model which minimizes the AIC.

However, including those who do not socialize at all in this study may not be appropriate, because depending on which group they are assigned to (either to treatment or control groups), the result of the entire propensity score analysis, including the conclusion on the relationship between in-person socialization and mental health conditions, could be changed. Hence, as the next step, it is worth investigating whether the result of the study changes when those who do not socialize at all are removed from the data.

References

1. 2015 GSS data:
   "General Social Survey Main File, Cycle 29: 2015: Time Use." CHASS - Microdata Analysis and Subsetting with SDA, Statistics Canada, Oct. 2017, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda2+gss15m.

2. Data cleaning code:
   Alexander, Rohan, and Samantha-Jo Caetano. gss_cleaning.R, 7 Oct. 2020, Alexander, R., & Caetano, S. (2020, October 7). Gss_cleaning.R. Retrieved December 1, 2020, from q.utoronto.ca/courses/184060/files/9422740?module_item_id=1867317.

3. 2015 GSS data user guide - general information on the data set:
   "General Social Survey, 2015 Cycle 29: Time Use - Pubilc-Use Microdata File Documentation and User's Guide." CHASS Microdata Analysis and Subsetting with SDA, Statistics Canada, Oct. 2017, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss29/gss29/more_doc/PUMF%20GSS%202015.pdf.

4. Propensity Score matching using R:
   Alexander, Rohan. "Difference in Differences." Telling Stories With Data, 5 Nov. 2020, www.tellingstorieswithdata.com/06-03-matching_and_differences.html.

5. Creating a table using kable and kableExtra:
   Zhu, Hao. Create Awesome HTML Table with Knitr::Kable and KableExtra. 22 Oct. 2020, cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html.

6. Information on the Akaike Information Criterion(AIC):
   Sheather, Simon J. A Modern Approach to Regression with R. Springer, 2010.

7. Performance of different model selection methods:
   Brewer, Mark J., et al. The Relative Performance of AIC, AICC and BIC in the Presence of Unobserved Heterogeneity. 13 June 2016, besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12541.

8. Creating a regression table with Huxtable: Hugh-Jones, David. Regression Tables with Huxreg, 27 Oct. 2020, cran.r-project.org/web/packages/huxtable/vignettes/huxreg.html.

9. Information on Multicollinearity:
   "Tutorial on Detecting Multicollinearity with Example." EduPristine, 7 Feb. 2020, www.edupristine.com/blog/detecting-multicollinearity.