Analyzing People's Stress Level Related to Their Work Status

Siyu Chen

December 21, 2020

Code and data can be accessed via github at https://github.com/cindyoy65/304 , licensed under MIT.

Abstract

This study is going to analyze the impact of sex, numbers of hours worked per week, self-rated job satisfaction and disability status on self-rated stress level, so it is able to predict the relationship between the stress level and other factors. We use the multiple linear regression model with propensity score to predict the relationship by using General Social Survey (Canadians at Work and Home) 2016 data. We find that the stress level is related to persons' income, numbers of hours worked and whether they feel isolated at work.

Keywords

Statistics, Multiple Linear Regression, Propensity Score Matching, Observational Study, Stress Level, Work Status

Introduction

People spend the most time of their lives working and it plays a crucial role in our daily lives. Working can be frustrating and make people stress. On the other hand, working can be inspiring and let people feel their sense of worth. How do working status affect self-rated stress level is an interesting topic, their relationship is complex but important. High stress levels can cause a variety of mental health issues. According to a search by the World Health Organization, Depression, anxiety and other mental health problems have a significant economic impact; the estimated cost to the global economy is US$ 1 trillion per year in lost productivity. Being able to predict workers' stress levels based on working status and conditions is crucial. For employers, it can make employees more productive and earn more profits. For the governments, it helps boost the economy and increase the utility level of citizens. This analysis tries to predict peoples' stress levels based on working status data, using multiple linear regression paired with propensity score matching.

This report investigated the relationship between individuals' self-rated stress levels and how well their works are for people aged 20 to 50. People of these age groups mainly have jobs and may struggle with working. In our model, we first build the multiple linear regression model with propensity score matching (PSM). The concept of PSM was first introduced by Rosenbaum and Rubin in 1983 (Thavaneswaran). In 1997, Heckman focused on selection bias and later developed difference-indifferences approach which has applications to PSM (Thavaneswaran). After that, PSM became popular all the way up to now. We apply propensity score matching to estimate the effect of the treatment group (job satisfaction) we select. Job satisfaction is used as the treatment group in the propensity score. The multiple linear regression model is a prediction method using multiple explanatory variables to estimate the value of the response variable. In our case, the stress level is the response variable and the number of hours of work, income, sex, whether feel isolated at work, job satisfaction and disability status are considered as the explanatory variables.

Data is collected by the 2016 General Social Survey (GSS) on Canadians at work and home. In the Methodology section (Section 2), a multiple linear regression model with propensity score is used to create a model to predict individuals' stress levels related to their work status. Then, the build of the model using backward elimination of AIC and using p-value will be shown in the Results section (Section 3), and in the Discussion section (Section 4), the conclusion and final steps will be presented. The final part is the Reference section (Section 5).

Methodology

Data

All data is collected by General Social Survey (Canadians at work and home) 2016.The frame population is the lists of telephone numbers in use available to Statistics Canada from various sources and the address register. The dataset has 10117 observations and after filter, it has 241 observations of 7 variables. The target population are all non-institutionalized persons over 15 years old who living in the 10 provinces of Canada. We selected the age groups from 20 to 49 years old, the mainly worked age population. The observation will be removed if one of the response is missing among five questions. The missing data may cause difficulty when we perform the future statistic method.

Table 1 provides the characteristics of our study populations which separated by job satisfaction.

Table 1: Baseline Characteristics Table for Job Satisfaction

	Very satisfied (N=51)	Satisfied (N=110)	Neither satisfied nor dissatisfied (N=40)	Dissafistied (N=29)	Very dissatisfied (N=11)	Total (N=241)
stress_lvl						
Not at all stressful	3 (5.9%)	3 (2.7%)	3 (7.5%)	0 (0%)	1 (9.1%)	10 (4.1%)

	Very satisfied (N=51)	Satisfied (N=110)	Neither satisfied nor dissatisfied (N=40)	Dissafistied (N=29)	Very dissatisfied (N=11)	Total (N=241)
Not very stressful	10 (19.6%)	13 (11.8%)	2 (5.0%)	1 (3.4%)	1 (9.1%)	27 (11.2%)
A bit stressful	24 (47.1%)	54 (49.1%)	15 (37.5%)	16 (55.2%)	1 (9.1%)	110 (45.6%)
Quite a bit stressful	9 (17.6%)	30 (27.3%)	15 (37.5%)	9 (31.0%)	6 (54.5%)	69 (28.6%)
Extremely stressful	5 (9.8%)	10 (9.1%)	5 (12.5%)	3 (10.3%)	2 (18.2%)	25 (10.4%)
sex						
Male	17 (33.3%)	50 (45.5%)	29 (72.5%)	15 (51.7%)	6 (54.5%)	117 (48.5%)
Female	34 (66.7%)	60 (54.5%)	11 (27.5%)	14 (48.3%)	5 (45.5%)	124 (51.5%)
income						
Less than $25,000	8 (15.7%)	21 (19.1%)	6 (15.0%)	4 (13.8%)	3 (27.3%)	42 (17.4%)
$25,000 to $49,999	21 (41.2%)	39 (35.5%)	11 (27.5%)	10 (34.5%)	3 (27.3%)	84 (34.9%)
$50,000 to $74,999	14 (27.5%)	29 (26.4%)	14 (35.0%)	8 (27.6%)	2 (18.2%)	67 (27.8%)
$75,000 to $99,999	3 (5.9%)	11 (10.0%)	1 (2.5%)	6 (20.7%)	0 (0%)	21 (8.7%)
$100,000 to $124,999	3 (5.9%)	2 (1.8%)	5 (12.5%)	0 (0%)	2 (18.2%)	12 (5.0%)
$125,000 or more	2 (3.9%)	8 (7.3%)	3 (7.5%)	1 (3.4%)	1 (9.1%)	15 (6.2%)
number_works_perweek						
>0 to 15 hours	2 (3.9%)	5 (4.5%)	0 (0%)	2 (6.9%)	0 (0%)	9 (3.7%)
16 to 29 hours	6 (11.8%)	14 (12.7%)	3 (7.5%)	6 (20.7%)	1 (9.1%)	30 (12.4%)
30 to 40 hours	36 (70.6%)	73 (66.4%)	24 (60.0%)	17 (58.6%)	9 (81.8%)	159 (66.0%)
41 hours and above	7 (13.7%)	18 (16.4%)	13 (32.5%)	4 (13.8%)	1 (9.1%)	43 (17.8%)
disability_status						
Yes	36 (70.6%)	92 (83.6%)	32 (80.0%)	23 (79.3%)	11 (100%)	194 (80.5%)
No	15 (29.4%)	18 (16.4%)	8 (20.0%)	6 (20.7%)	0 (0%)	47 (19.5%)
feel_isolated						
Yes	8 (15.7%)	23 (20.9%)	18 (45.0%)	13 (44.8%)	5 (45.5%)	67 (27.8%)
No	43 (84.3%)	87 (79.1%)	22 (55.0%)	16 (55.2%)	6 (54.5%)	174 (72.2%)

Here are the description of each variable:

stress_lvl: Level of stress in life. It is self-rated and has scale from 1 to 5. (1 is not at all stressful and 5 is extremely stressful)

sex: sex of respondent (male and female only)

income: personal income group before tax, gap of $25000 and has 6 interval, starting from less than $25000, up to $125000 or more

number_works_perweek: number of hours worked per week at job. The groups are >0 to 15 hours, 16 to 29 hours, 30 to 40 hours and 41 hours and above.

disability_status: disability status.

feel_isolated: Whether feeling isolated at work.

job_satisfaction: how satisfied with the job. The scale is from 1 to 5. (1 is very satisfied and 5 is very dissatisfied)

```
## Warning in matrix(c(1, -0.1544, -0.1164, -0.1632, -0.0667, 0.1385, -0.1544, :
## Êý¾Ý³¤¶È[35]²»ÊÇ¾ØÕóÐÐÊý[6]µÄÕû±¶
```

Figure 1: Pairwise Correlation Between Explanatory variables

	Sex	Income	Number of hours worked per week	Job satisfaction	Disability status	Feel isolated
Sex	1.0000	-0.1544	-0.1164	-0.1632	-0.0667	0.1385
Income	-0.1544	1.0000	0.2973	0.0602	0.0643	-0.2264
Number of hours worked per week	-0.1164	0.2973	1.0000	0.0152	-0.0783	-0.1018
Job satisfaction	-0.1632	0.0602	0.0152	1.0000	-0.1031	-0.2474
Disability status	-0.0667	0.0643	-0.0783	-0.1031	1.0000	0.0717

	Sex	Income	Number of hours worked per week	Job satisfaction	Disability status	Feel isolated
Feel isolated	-0.0879	-0.1018	-0.2474	0.0717	1.0000	1.0000

We use pairwise correlation to check the multicollinearity of the MLR model. Multicollinearity is a situation that two or more explanatory variables are highly correlated. If that happens, our model will be unstable and the estimated regression coefficients vary widely. In this model, all of the absolute correlation coefficients are small enough to accept all of the explanatory variables for now.

Model

A multiple linear regression model with propensity score matching is used to predict the relationship between the stress level with other explanatory variables (sex, income, numbers of hours worked per week, feel isolated, job satisfaction). We apply propensity score matching to estimate the effect of the treatment group (job satisfaction) we select. We use job satisfaction as the treatment group in the propensity score. To have two groups in our model, observations with job satisfaction equals “very satisfied” or “satisfied” considered to be Group 1. And observations with job satisfaction equals “very dissatisfied” or “dissatisfied” considered to be Group 2. We omit the observations with job satisfaction equal “neither satisfied nor dissatisfied”.

The estimated model is:

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$$

where $ y$ refers to the estimated stress level; b_0 refers to the overall intercept; b_1 refers to the slope of sex X_1 refers to the value of sex (1 represents male and 2 represents female) b_2 refers to the slope of income X_2 refers to the value of income before tax b_3 refers to the slope of numbers of hours worked per week X_3 refers to the value of numbers of hours worked per week b_4 refers to the slope of whether feel isolated X_4 refers to the value of whether feel isolated (1 represents Yes and 2 represents No) b_5 refers to the slope of the disability status X_5 refers to the value of the disability status (1 represents Yes and 2 represents No)

Figure 2: Residuals vs Fitted plot

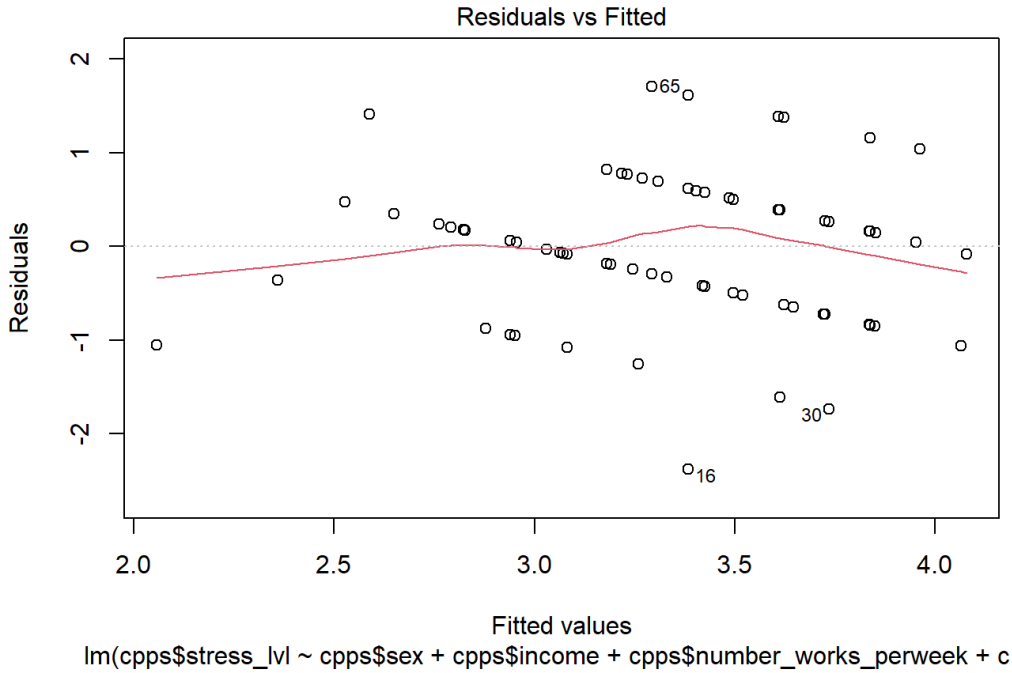

Figure 2 is the residual versus fitted plot. It shows the linear relationship between explanatory variables and the response variable. In this Figure 2, though the horizontal line can be treated as flat, the plot has pattern. It may cause because there exist too many category variables. So it does not meetthe equal variance assumption

Figure 3: Normal Q-Q plot

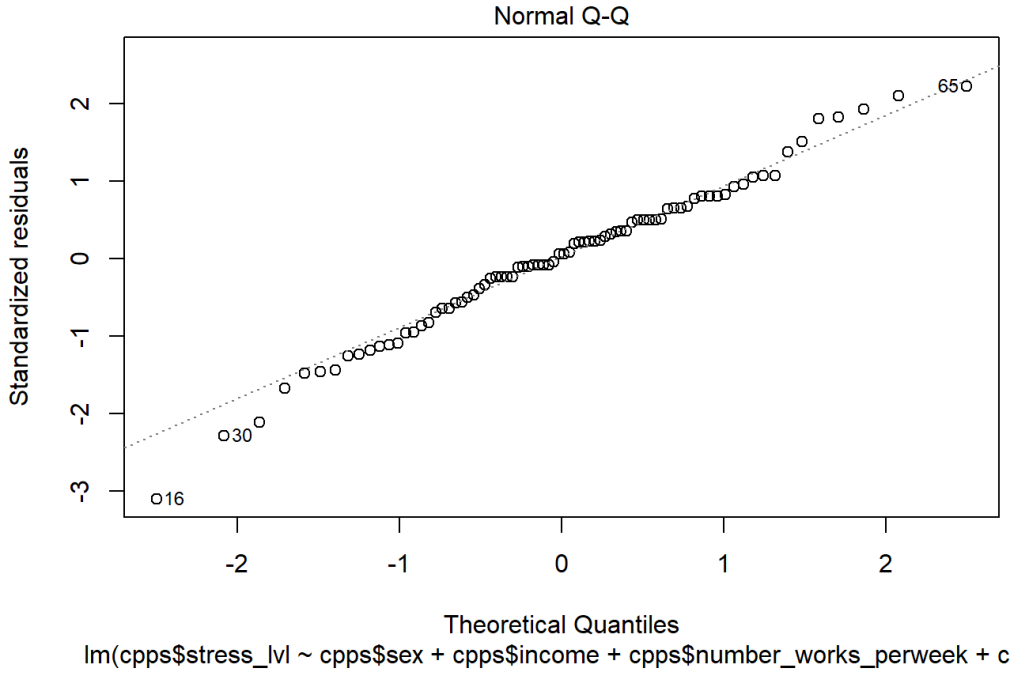

Figure 3 is the normal Q-Q plot of standardize residuals. It checks whether the residuals are normally distributed. This plot is little heavy tail. Except those extreme cases, we can still conclude that our model follows normal distribution.

Figure 4: standardized residuals versus fitted value plot

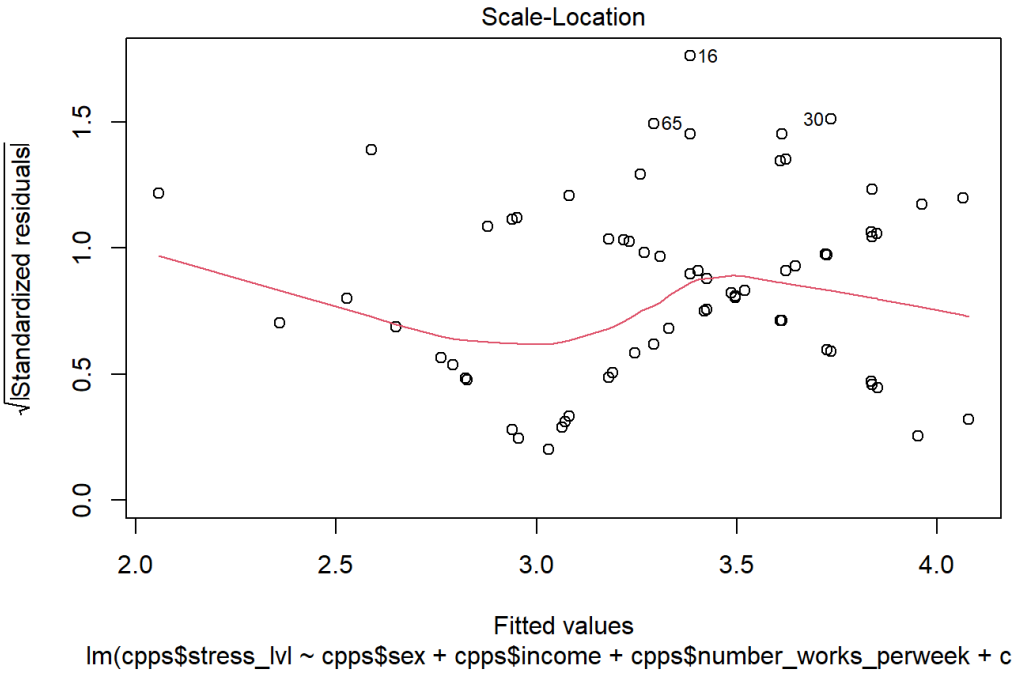

Figure 4 is standardized residuals versus fitted value plot. The line is not flat and the points are not randomly distributed along the line. So the assumption of equal variance does not achieved.

Figure 5: residuals vs leverage plot

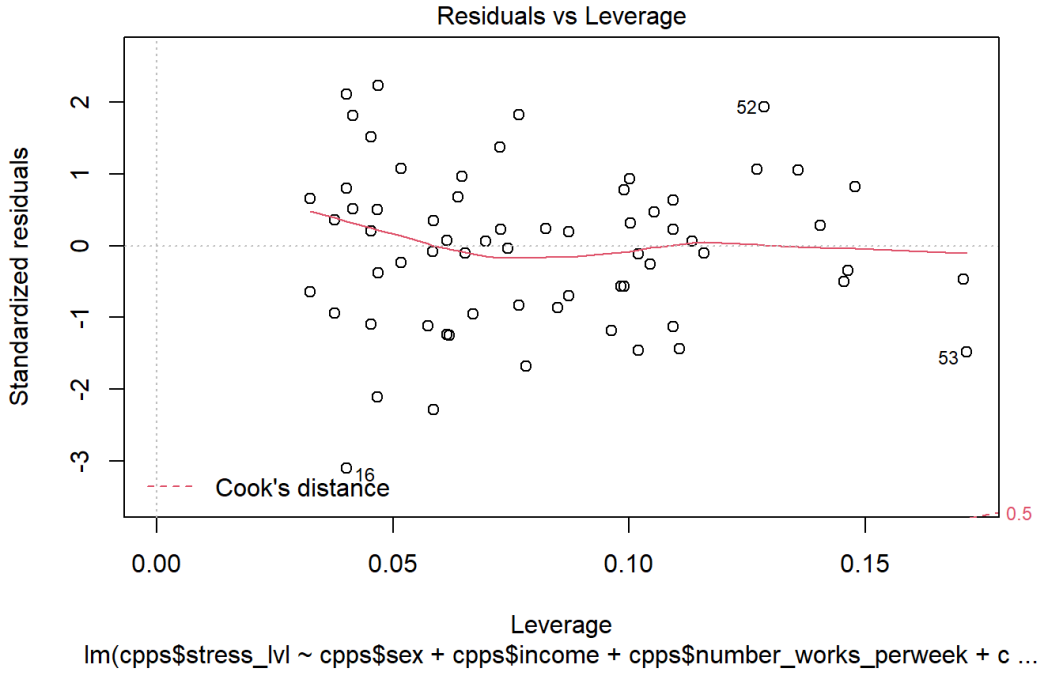

Figure 5 is the residuals vs leverage plot. It checks the influential cases. We can see that the numbers of the influential points are not big enough and there exists a horizontal cook's distance line. So we can conclude that it has a linear relationship between the stress level and the other explanatory variables.

Results

We first estimate the model by discussing the p-value:

Figure 3: Estimated Coefficients Summary

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.9424	0.6051	4.863	0.0000063
Sex	0.2294	0.1777	1.291	0.2008000
Income	0.1130	0.0715	1.580	0.1184000
Numbers of hours worked per week	0.2396	0.1315	1.823	0.0724000
feel isolated	-0.4324	0.1861	-2.323	0.0229000
disability status	-0.3012	0.2278	-1.322	0.1902000

Only p-value of feel isolated is less than 0.05, we reject the hypothesis which feel isolated is not 0. The p-value of the other variables are greater than 0.05, so we fail to reject the Null hypothesis Then, the fitted equation of multiple linear regression model is:

$$\hat{s}tresslevel = 2.9424 - 0.4324 * feelisolated$$

We also use AIC to build the model.

Final model using backward elimination of AIC is:

$$\hat{s}tresslevel = 2.9016 + 0.1131 * income + 0.2476 * numbersofhoursworkedperweek - 0.4278 * feelisolated$$

Since AIC is likelihood-based criterion and balancing goodness of fit and penalizing complexity of the model, in the discussion part, we discuss with the model using backward elimination of AIC.

From Figure 6, it shows the respondents' stress level has a positive relationship with whether they feel isolated. If they feel isolated at work, then their stress level in the life also become large.

Discussion

Summary

Our goal is to use the multiple linear regression model with the propensity score to predict the impact of the stress level on working status. We analyze the observations from the aspects of sex, income, the number of hours worked, feel isolated while working and disability status. We first obtain data from General Social Survey. Then we set the propensity score using “job satisfaction” variable as the treatment group. Next, we construct a multiple linear regression model and analyze it from the aspect of p-value and AIC method. We use the model applying AIC to do the conclusion and point out the weakness later.

Results Analysis

Our model is

$$\hat{s}tresslevel = 2.9016 + 0.1131 * income + 0.2476 * numbersofhoursworkedperweek - 0.4278 * feelisolated$$

. Then, holding other factors, if the income increases by $1, the stress level increases by 0.1131. Holding other factors, if the number or hours worked per week rises by 1 hour, then the stress level increases by 0.2476 Holding other factors, if people feel isolated at work, then stress level decreases by 0.4278 (0.4278). If people does not feel isolated at work, then the stress level decreases by 0.8556 (0.4278*2).

Significance

This model is able to predict workers’ stress levels based on working status and conditions is crucial. For employers, it can make employees more productive and earn more profits. For the governments, it helps boost the economy and increase the utility level of citizens.

Conclusions

The final model with propensity score and AIC is

$$\hat{s}tresslevel = 2.9016 + 0.1131 * income + 0.2476 * numbersofhoursworkedperweek - 0.4278 * feelisolated$$

.

It shows that the stress level is related to persons’ income, numbers of hours worked and whether they feel isolated at work.The stress level is positively related to all three variables.

Weakness & Next Steps

Firstly, our data sample is not big enough. The observations with any missing answers are removed. In future, we could make more attempts to get a completed interview.

Also, our data selection is not good as well. The dataset set almost all variables as the category variables. Variables such as income and age are divided into groups. With too many category variables, it is hard to do the pairwise correlation to determine the relationship between each explanatory variable. Next time, numerical variables are more straightforward to be used.

References

Ggplot2 colors : How to change colors automatically and manually? (n.d.). Retrieved December 21, 2020, from http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually

Hao Zhu (2020). kableExtra: Construct Complex Table with ‘kable’ and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

Kemeny, A. (2018, June). General Social Survey Cycle 30: Canadians at Work and Home. Retrieved December 21, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss30/gss30/more_doc/GSSC30ENgid.pdf

Mental health in the workplace. (n.d.). Retrieved December 23, 2020, from https://www.who.int/teams/mental-health-and-substance-use/mental-health-in-the-workplace

(n.d.). Retrieved December 21, 2020, from https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.html

Propensity score matching. (2020, December 02). Retrieved December 21, 2020, from https://en.wikipedia.org/wiki/Propensity_score_matching

Thavaneswaran, A., & Lix, L. (2008, April 22). Propensity Score Matching in Observational Studies. Retrieved December 21, 2020, from https://www.umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/protocol/media/propensity_score_matching.pdf

The Methodology Center (n.d.). AIC vs. BIC. Retrieved December 21, 2020, from https://www.methodology.psu.edu/resources/AIC-vs-BIC/