Whether married with older partner has a better life feeling than younger ones?

Yangming Hu

2020-10-19

Abstract

This study investigated the question of interest that whether married with older partner has a better life feeling. The study mainly applied statistical approaches based on linear regression models to study the question of interest. It was found that there are strong evidence show the facts that people married with no younger partner are indeed have a better average life feeling score than those married with younger partner significantly. This finding suggests people should find no younder partner than themselves if they want to find a higher level of life feeling in their future life.

Introduction

It is widely know that for marriages especailly in morden days, people is not constrained to find partners with the same ages as themselves. Married with people much younger or older than themselves become common things in recent years. One of potential reasons is that it is claimed married with older poeple would be more happier than married with younger ones. This study investigated this question of interest applied linear regression models. It was found that controlled lots of important covariates such as age, gender, education level, income, hours of work and etc, the study shows that married with no younger partner are indeed have a better average life feeling score than those married with younger partner. This is important as it suggests that people should find no younder partner than themselves if they want to find a higher level of life feeling in their future life. The study was organized as following: the introduction section introduced the goal of the study and main findings, methods used. The data section introduced the data and discussions. The model section introduced the linear regression model used in the study and gave discussions. Results section illustrated all of the results obtained in the study and finally, the discussion section discussed the findings, the weakness of the study and the possible future work. The study is originally hold in the link: https://github.com/Jaydenhu123/STA304_PS2/blob/main/Jarden_304_PS2.pdf.

Data

The source of the data studied in the study is the 2017 Canadian General Social Survey Data. The GSS data contains many instances with lots of variables. This study did the work based on the Rohan Alexander and Sam Caetano (2020). The response is the feeling score of life which is an ordinal scaled variable from 0 to 10 and the categorical factor with 3 levels - older, same and younger aged partners as the main interested variable. The models also include various of important covariates such as age, gender, education, income and etc.

The 2017 GSS survey is well-tested and designed that the answers covered lots of information of peersons which are important. However, there are also lots of non-responses to some questions. The target population of the survey is all of the people living in the 10 provinces of Canada with ages no younger than 15. The frame is a list of landline and cellular telephone numbers in Canada. The samples are the units collected by the survey.

The survey used a stratified sampling method that it divides Canada into several areas and draw samples in the strata of these areas. There are non-responses problems that not all of people might be studied, if people do not have telephone numbers. Also, people who are not want to answer the questions are not investigated, and the non-answers to questions also cause non-response bias. The survey deal with these non-response bias by using methods like estimation instead of actual values, it is a good trade-off as some responses could be estimated from other sources such as tax files and etc. Table 1 shows a summary of the features of the data used in this study.

Table 1: Summary of features

name	type	mean	disp	median	min	max	nlevs
age	numeric	48.81	12.29	48.8	20.3	80	0
age_diff	factor	NA	0.61	NA	1477.0	2494	3
feelings_life	numeric	8.39	1.39	8.0	0.0	10	0
number_marriages	numeric	1.12	0.37	1.0	1.0	4	0
children_in_household	factor	NA	0.40	NA	390.0	3824	4
lives_with_partner	factor	NA	0.01	NA	37.0	6355	2
sex	factor	NA	0.48	NA	3066.0	3326	2
own_rent	factor	NA	0.14	NA	882.0	5510	2
average_hours_worked	factor	NA	0.37	NA	11.0	4037	5
education	factor	NA	0.37	NA	2348.0	4044	2
income_family	factor	NA	0.59	NA	166.0	2638	6
self_rated_health	factor	NA	0.62	NA	79.0	2439	5
self_rated_mental_health	factor	NA	0.64	NA	46.0	2294	5

Model

The linear regression model in the study uses the feeling score of life as the response which is an ordinal scaled variable from 0 to 10 and the categorical factor with 3 levels - older, same and younger aged partners as the main interested variable. The models also include various of important covariates. The inear regression model is described as below:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_p + \epsilon$$

where y is feeling score of life, x are various covarites and the main interested facor age difference. The ϵ is i.i.d. ~ N(0,σ^2). And for model checks and diagnostics, main the 5 aspects are investigated:

1) independent assumption: the observations should be independent
2) linearity assumption: the relationship between response and factors should be linearity
3) constant variance assumption: the residuals should have the constant variance
4) normality assumption: the residuals should follow normal distribution
5) unsual points: there are should no outliers, influence points

And in this study, the above assumptions are mainly checked using model diagnostic plots. The whole procedure of model building, model diagnostics are performed in the R software which is designed originally for statistical analysis. At last, linear regression model is choosen among different types of model because it is suitable for the topic and very easy to interpret, other models like bayes models are too complicated and time costing in fitting big data set, also there are no special prior information in this study.

Results

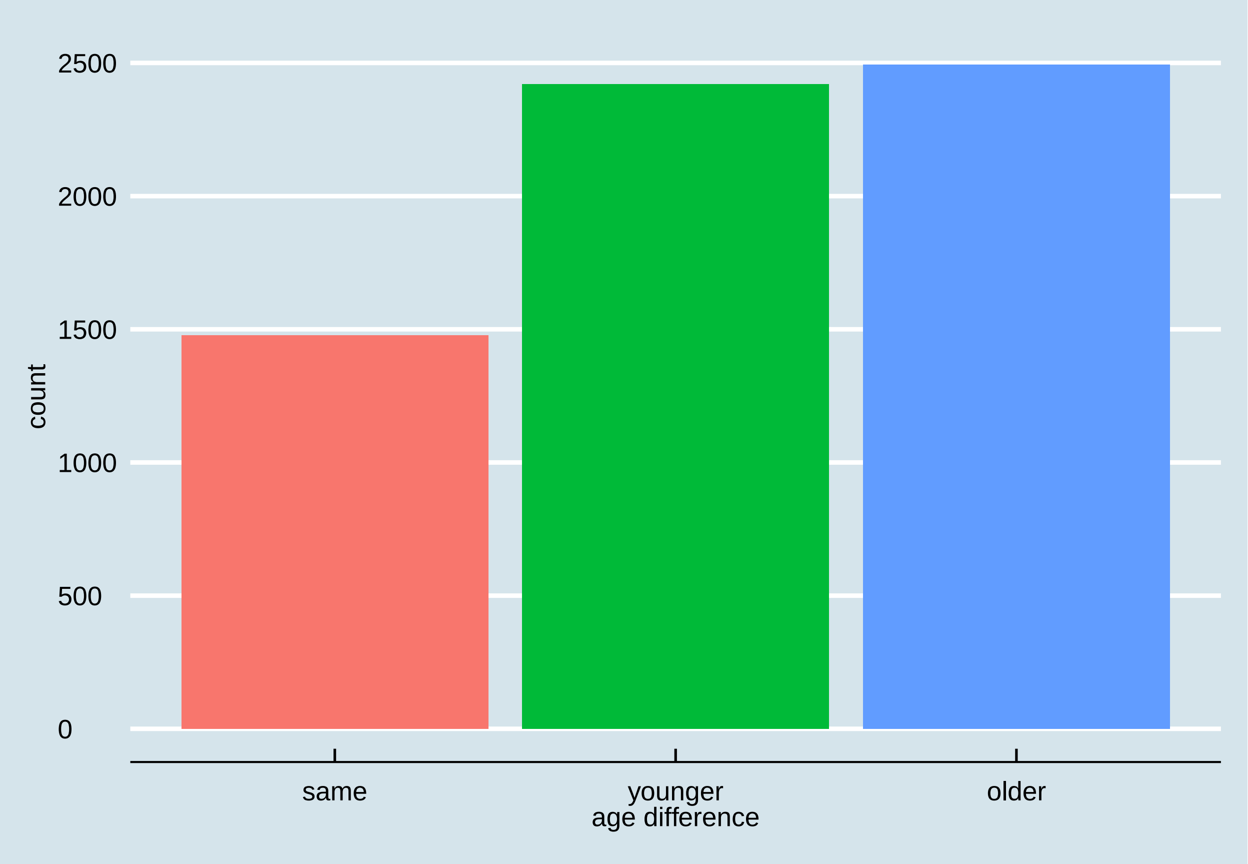

Figure 1: Distribution of the interested factor age difference. There are three levels - older, same and younger indicating the invidiuals have partner with a older, same and younger age respectively

Figure 1 shows the distribution of the interested factor age difference. There are three levels - older, same and younger indicating the invidiuals have partner with a older, same and younger age respectively. It shows the data has more instances married with people having older or younger age than the instances themselves.

Figure 2 shows the distribution of the life feeling score grouped by the interested factor age difference. The distributions are across the three levels of older, same and younger age difference respectively. Just from the boxes, it appears that the three levels have similar distributions of life feeling score, however, if look carefully, it can be found the average median levels of life feeling scores are different across the three levels of age difference.

Table 2: Simple linear model with only interested factor age difference

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.455	0.036	233.944	0.000
age_diffyounger	-0.078	0.046	-1.707	0.088

	Estimate	Std. Error	t value	Pr(>|t|)
age_diffolder	-0.088	0.046	-1.941	0.052

Table 2 shows the estimates of simple linear model with only interested factor age difference. The two dummy variables older and younger levels are both significant at a level of 10% but not signiicant at a level of 5%.

Table 3: Full linear model with interested factor age difference along with all of covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.620	0.389	22.169	0.000
age	0.005	0.002	3.123	0.002
age_diffyounger	-0.035	0.042	-0.829	0.407
age_diffolder	-0.072	0.042	-1.725	0.085
number_marriages	0.085	0.043	1.993	0.046
children_in_householdOne child	-0.062	0.048	-1.296	0.195
children_in_householdThree or more children	0.081	0.070	1.158	0.247
children_in_householdTwo children	0.050	0.047	1.059	0.289
lives_with_partnerYes	-0.099	0.202	-0.487	0.626
sexMale	-0.127	0.037	-3.412	0.001
own_rentRented	-0.216	0.048	-4.505	0.000
average_hours_worked0.1 to 29.9 hours	0.219	0.372	0.588	0.557
average_hours_worked30.0 to 40.0 hours	0.190	0.371	0.513	0.608
average_hours_worked40.1 to 50.0 hours	0.268	0.373	0.718	0.473
average_hours_worked50.1 hours and more	0.247	0.374	0.660	0.510
educationBelow Bachelor	0.212	0.033	6.351	0.000
income_family$125,000 and more	0.064	0.046	1.404	0.160
income_family$25,000 to $49,999	-0.209	0.065	-3.188	0.001
income_family$50,000 to $74,999	-0.007	0.056	-0.125	0.901
income_family$75,000 to $99,999	-0.143	0.054	-2.663	0.008
income_familyLess than $25,000	-0.428	0.105	-4.084	0.000
self_rated_healthFair	-0.696	0.075	-9.261	0.000
self_rated_healthGood	-0.353	0.048	-7.414	0.000
self_rated_healthPoor	-0.872	0.146	-5.954	0.000
self_rated_healthVery good	-0.224	0.043	-5.183	0.000
self_rated_mental_healthFair	-1.871	0.086	-21.717	0.000
self_rated_mental_healthGood	-0.975	0.046	-21.341	0.000
self_rated_mental_healthPoor	-3.190	0.187	-17.101	0.000
self_rated_mental_healthVery good	-0.452	0.040	-11.177	0.000

Table 3 shows the estimates of full linear model with interested factor age difference along with all of covariates but no interaction effects. Now, only the dummy variable older is significant at a level of 10%. Also, it can be found lots of covariates are significant indicating they are important covariates which should be inlcuded in the model.

Table 4: Best subset linear model with interested factor age difference along with selected covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.819	0.108	81.303	0.000
age	0.005	0.002	3.248	0.001

	Estimate	Std. Error	t value	Pr(>|t|)
age_diffyounger	-0.036	0.042	-0.865	0.387
age_diffolder	-0.074	0.042	-1.758	0.079
number_marriages	0.084	0.043	1.965	0.050
children_in_householdOne child	-0.063	0.048	-1.315	0.189
children_in_householdThree or more children	0.083	0.069	1.200	0.230
children_in_householdTwo children	0.049	0.047	1.035	0.301
sexMale	-0.117	0.036	-3.266	0.001
own_rentRented	-0.222	0.048	-4.641	0.000
educationBelow Bachelor	0.213	0.033	6.381	0.000
income_family$125,000 and more	0.065	0.046	1.434	0.152
income_family$25,000 to $49,999	-0.202	0.065	-3.092	0.002
income_family$50,000 to $74,999	-0.004	0.056	-0.065	0.948
income_family$75,000 to $99,999	-0.141	0.054	-2.634	0.008
income_familyLess than $25,000	-0.419	0.104	-4.026	0.000
self_rated_healthFair	-0.694	0.075	-9.242	0.000
self_rated_healthGood	-0.352	0.048	-7.386	0.000
self_rated_healthPoor	-0.875	0.146	-5.989	0.000
self_rated_healthVery good	-0.223	0.043	-5.176	0.000
self_rated_mental_healthFair	-1.872	0.086	-21.739	0.000
self_rated_mental_healthGood	-0.976	0.046	-21.368	0.000
self_rated_mental_healthPoor	-3.192	0.186	-17.145	0.000
self_rated_mental_healthVery good	-0.452	0.040	-11.189	0.000

Table 4 shows the estimates of best subset linear model selected by AIC backward approach with interested factor age difference along with selected covariates. Still, the results are similar with the previous full model that only the dummy variable older is significant at a level of 10%.

Table 5: Best subset linear model with interested factor age difference along with selected covariates as well as interactions among them

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.191	0.192	47.857	0.000
age	-0.001	0.003	-0.365	0.715
age_diffyounger	-0.822	0.241	-3.411	0.001
age_diffolder	-0.246	0.238	-1.033	0.302
number_marriages	0.079	0.043	1.857	0.063
children_in_householdOne child	-0.205	0.097	-2.114	0.035
children_in_householdThree or more children	-0.252	0.137	-1.839	0.066
children_in_householdTwo children	-0.027	0.096	-0.279	0.780
lives_with_partnerYes	-0.696	0.436	-1.596	0.111
sexMale	-0.112	0.036	-3.129	0.002
own_rentRented	-0.409	0.105	-3.891	0.000
educationBelow Bachelor	0.210	0.033	6.293	0.000
income_family$125,000 and more	0.065	0.045	1.426	0.154
income_family$25,000 to $49,999	-0.214	0.065	-3.279	0.001
income_family$50,000 to $74,999	0.012	0.056	0.207	0.836
income_family$75,000 to $99,999	-0.146	0.054	-2.726	0.006
income_familyLess than $25,000	-0.403	0.104	-3.877	0.000
self_rated_healthFair	-0.670	0.167	-4.014	0.000
self_rated_healthGood	-0.284	0.098	-2.888	0.004

	Estimate	Std. Error	t value	Pr(>|t|)
self_rated_healthPoor	0.234	0.405	0.579	0.563
self_rated_healthVery good	-0.231	0.086	-2.677	0.007
self_rated_mental_healthFair	-2.015	0.185	-10.869	0.000
self_rated_mental_healthGood	-1.005	0.098	-10.290	0.000
self_rated_mental_healthPoor	-1.515	0.439	-3.451	0.001
self_rated_mental_healthVery good	-0.460	0.082	-5.592	0.000
age:age_diffyounger	0.013	0.004	3.233	0.001
age:age_diffolder	0.003	0.004	0.743	0.457
age_diffyounger:children_in_householdOne child	0.286	0.124	2.311	0.021
age_diffolder:children_in_householdOne child	0.091	0.124	0.739	0.460
age_diffyounger:children_in_householdThree or more children	0.621	0.176	3.532	0.000
age_diffolder:children_in_householdThree or more children	0.258	0.180	1.437	0.151
age_diffyounger:children_in_householdTwo children	0.167	0.123	1.359	0.174
age_diffolder:children_in_householdTwo children	0.033	0.123	0.272	0.786
age_diffyounger:lives_with_partnerYes	1.211	0.547	2.215	0.027
age_diffolder:lives_with_partnerYes	0.325	0.539	0.604	0.546
age_diffyounger:own_rentRented	0.183	0.126	1.451	0.147
age_diffolder:own_rentRented	0.283	0.127	2.217	0.027
age_diffyounger:self_rated_healthFair	-0.217	0.205	-1.061	0.289
age_diffolder:self_rated_healthFair	0.165	0.204	0.808	0.419
age_diffyounger:self_rated_healthGood	-0.069	0.124	-0.553	0.581
age_diffolder:self_rated_healthGood	-0.101	0.125	-0.813	0.416
age_diffyounger:self_rated_healthPoor	-1.166	0.458	-2.545	0.011
age_diffolder:self_rated_healthPoor	-1.420	0.466	-3.047	0.002
age_diffyounger:self_rated_healthVery good	-0.046	0.112	-0.413	0.679
age_diffolder:self_rated_healthVery good	0.058	0.110	0.530	0.596
age_diffyounger:self_rated_mental_healthFair	0.456	0.238	1.917	0.055
age_diffolder:self_rated_mental_healthFair	-0.059	0.225	-0.261	0.794
age_diffyounger:self_rated_mental_healthGood	0.126	0.121	1.037	0.300
age_diffolder:self_rated_mental_healthGood	-0.049	0.122	-0.403	0.687
age_diffyounger:self_rated_mental_healthPoor	-1.291	0.542	-2.381	0.017
age_diffolder:self_rated_mental_healthPoor	-2.599	0.515	-5.042	0.000
age_diffyounger:self_rated_mental_healthVery good	0.033	0.105	0.314	0.754
age_diffolder:self_rated_mental_healthVery good	-0.015	0.105	-0.140	0.889

Table 5 also consider various different interactions between the interested factor age difference with the covariates. It shows the best subset linear model by AIC backward approach with interested factor age difference along with covariates as well as interactions between the interested factor age difference with the covariates. The result shows that there are some significant interaction effects in the model, and the age difference dummy variable younger level is very significant with p value less than 0.05. The model result shows that fixed other factors, the main effect of married with younger partner is significantly negative that the average level of life feeling score is about 0.822 units lower than that of married with same aged partner. For married older partner, there is also a negative effect but it is not significant with a p value larger than 0.05.

Table 6: Best subset linear model with interested factor age difference along with selected covariates as well as interactions among them after dropping outliers with absolute standardized residuals larger than 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.227	0.160	57.604	0.000

	Estimate	Std. Error	t value	Pr(>|t|)
age	-0.001	0.003	-0.484	0.629
age_diffyounger	-0.814	0.201	-4.045	0.000
age_diffolder	-0.205	0.199	-1.030	0.303
number_marriages	0.099	0.036	2.736	0.006
children_in_householdOne child	-0.221	0.081	-2.738	0.006
children_in_householdThree or more children	-0.263	0.114	-2.302	0.021
children_in_householdTwo children	-0.063	0.081	-0.774	0.439
lives_with_partnerYes	-0.819	0.357	-2.293	0.022
sexMale	-0.085	0.030	-2.837	0.005
own_rentRented	-0.442	0.089	-4.992	0.000
educationBelow Bachelor	0.220	0.028	7.893	0.000
income_family$125,000 and more	0.067	0.038	1.759	0.079
income_family$25,000 to $49,999	-0.147	0.055	-2.660	0.008
income_family$50,000 to $74,999	0.060	0.047	1.276	0.202
income_family$75,000 to $99,999	-0.093	0.045	-2.075	0.038
income_familyLess than $25,000	-0.315	0.090	-3.500	0.000
self_rated_healthFair	-0.860	0.141	-6.112	0.000
self_rated_healthGood	-0.269	0.083	-3.258	0.001
self_rated_healthPoor	0.060	0.334	0.180	0.857
self_rated_healthVery good	-0.241	0.072	-3.352	0.001
self_rated_mental_healthFair	-1.995	0.163	-12.214	0.000
self_rated_mental_healthGood	-0.944	0.082	-11.526	0.000
self_rated_mental_healthPoor	-0.770	0.455	-1.693	0.091
self_rated_mental_healthVery good	-0.444	0.068	-6.488	0.000
age:age_diffyounger	0.013	0.003	3.724	0.000
age:age_diffolder	0.003	0.003	0.780	0.435
age_diffyounger:children_in_householdOne child	0.247	0.103	2.388	0.017
age_diffolder:children_in_householdOne child	0.080	0.104	0.767	0.443
age_diffyounger:children_in_householdThree or more children	0.595	0.147	4.038	0.000
age_diffolder:children_in_householdThree or more children	0.229	0.150	1.525	0.127
age_diffyounger:children_in_householdTwo children	0.129	0.103	1.254	0.210
age_diffolder:children_in_householdTwo children	-0.008	0.103	-0.075	0.940
age_diffyounger:lives_with_partnerYes	1.057	0.461	2.293	0.022
age_diffolder:lives_with_partnerYes	0.146	0.447	0.327	0.744
age_diffyounger:own_rentRented	0.289	0.107	2.707	0.007
age_diffolder:own_rentRented	0.401	0.108	3.705	0.000
age_diffyounger:self_rated_healthFair	0.141	0.175	0.806	0.420
age_diffolder:self_rated_healthFair	0.423	0.173	2.446	0.014
age_diffyounger:self_rated_healthGood	-0.090	0.104	-0.867	0.386
age_diffolder:self_rated_healthGood	-0.099	0.105	-0.941	0.347
age_diffyounger:self_rated_healthPoor	-1.254	0.385	-3.255	0.001
age_diffolder:self_rated_healthPoor	-1.253	0.390	-3.209	0.001
age_diffyounger:self_rated_healthVery good	0.000	0.094	0.001	0.999
age_diffolder:self_rated_healthVery good	-0.009	0.092	-0.101	0.920
age_diffyounger:self_rated_mental_healthFair	0.213	0.210	1.014	0.310
age_diffolder:self_rated_mental_healthFair	-0.227	0.198	-1.148	0.251
age_diffyounger:self_rated_mental_healthGood	0.147	0.102	1.444	0.149
age_diffolder:self_rated_mental_healthGood	-0.024	0.103	-0.233	0.816
age_diffyounger:self_rated_mental_healthPoor	-2.111	0.559	-3.774	0.000
age_diffolder:self_rated_mental_healthPoor	-3.856	0.523	-7.371	0.000
age_diffyounger:self_rated_mental_healthVery good	0.015	0.087	0.175	0.861
age_diffolder:self_rated_mental_healthVery good	-0.001	0.087	-0.007	0.995

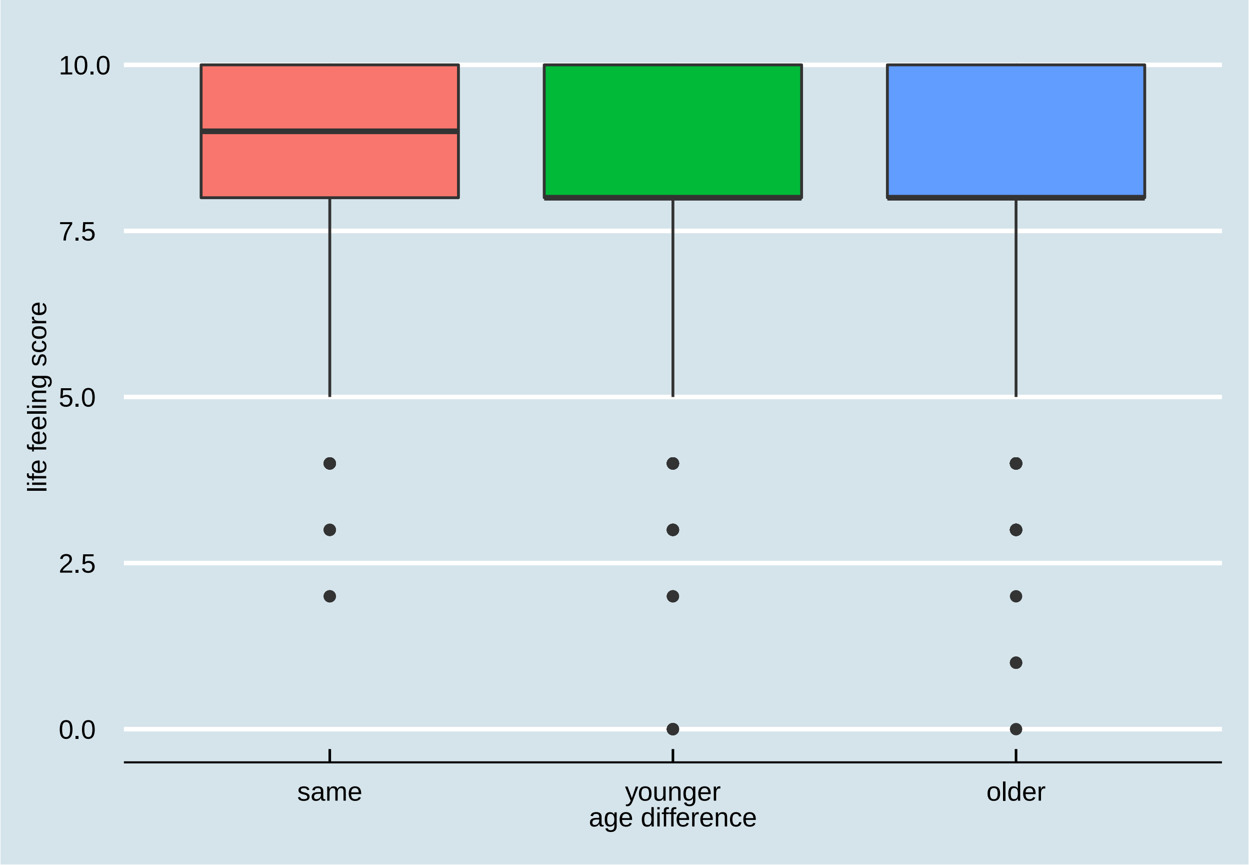

Figure 2: Distribution of the life feeling score grouped by the interested factor age difference. The distributions are across the three levels of older, same and younger age difference respectively

However, the findings in table 5 might be affected by outlers, table 6 shows the best subset linear model by AIC backward approach with interested factor age difference along with covariates as well as interactions between the interested factor age difference with the covariates after removing outliers with absolute standardized residuals larger than 2. And it can be found that only the estimates changed a little between the models in table 5 and table 6, the directions, significances of the interested factor are the same. So this means that after removing the outliers, the inferences are still consistent.

Discussion

After all of the above work. First, we dicuss the validation of the model obtained which is very important as model fail to pass the model checks are meaningless. And there are mainly 4 assumptions checked in this study.

1) independent assumption: the top left residuals plot in the figure 3 shows that the residuals points are randomly distributed arouned the zero mean line, it means there is no denepdent pattern.
2) linearity assumption: the top left residuals plot in the figure 3 shows there is no special curve pattern, the linearity asumption is true.
3) constant variance assumption: the top left residuals plot in the figure 3 shows that the spread of residuals does not change across x-axis obviously, the constant variance assumption is true.
4) normality assumption: the top right normal Q-Q plot in the figure 3 shows that the residuals fit the straight line well overall except some points far from the two ends, however, in practice, the results are robust in these cases, the normality assumption is acceptable.

Besides the 4 main assumptions, unusual data points could also be checked, the residuals plot and normal Q-Q plots show no residuals have obviously large absolute residuals. However, the bottom right leverage plot shows there are some possible high leverage points, the bottom left cook's distance plot shows there are some possible strong influence points. Future work of deal with these possible unusual data points could be investigated to improve the fitness of the model.

Second, we dicuss about the procedures of the models. A simple linear model without covariates was not enough to investigate the goal of the study as there are many important covariate not included in the model which could cause omitted variables bias, it is the same for only considering first order linear regression model without adding interation effects. In our study, there are significant different results obtained between a model with interaction effects and a model without considering interaction effects. However, high order terms are not considered in this study anymore. Future work could study this problem to improve the model.

At last, we dicussed some weaknesses in the study. The study is performed on a subset of 2017 GSS data, however, this data is known to have non-response bias and adjusted by some estimations, if the adjusted results are not close to the actual ones, then it would lead seriously biasness in our final model, thus, the inferences based on the final model are not reliable anymore. There are some procedures of recoding variables in this study, different designs of recoding might result in different results and inferences. Also the response in this study is an ordinal response variable scaled from 0 to 10, the range is not wide enough for linear regression models.

However, the findings in this study might be still useful that it shows married with same aged partner has significantly higher level of life feeling score than married with younger partner, although there is no significant difference between same aged partner and older partner, the finding already indicates the facts that people married with no younger partner are indeed have a better average life feeling score than those married with younger partner significantly.

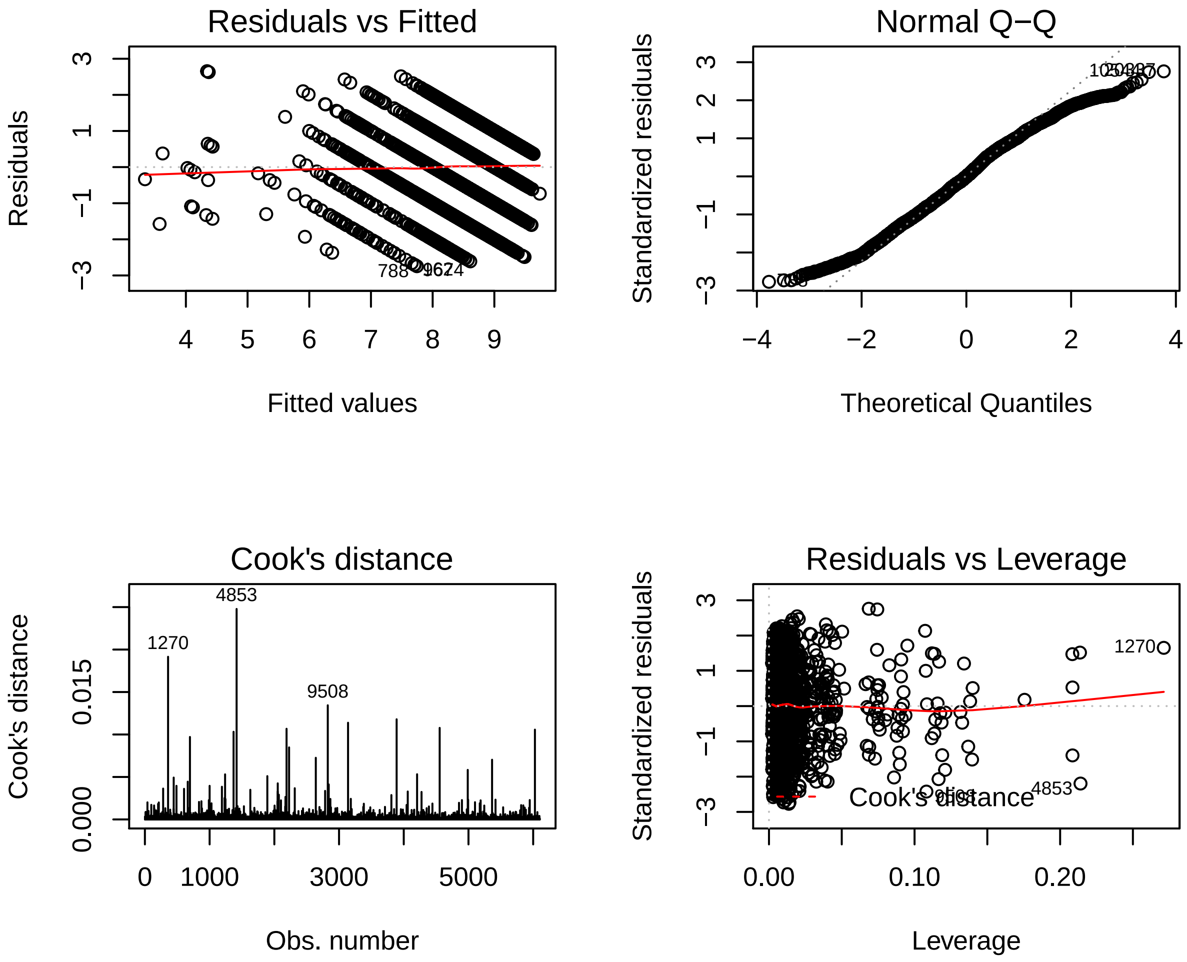

Figure 3: Model diagnostics for the final linear regression model

References

1. Alboukadel Kassambara (2019). ggpubr: ‘ggplot2’ Based Publication Ready Plots. R package version 0.2.4. https://CRAN.R-project.org/package=ggpubr
2. Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
3. Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
4. Hadley Wickham, Romain Franois, Lionel Henry and Kirill Muller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.Rproject.org/package=dplyr
5. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
6. Jeffrey B. Arnold (2019). ggthemes: Extra Themes, Scales and Geoms for ‘ggplot2’. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes
7. Probst P, Au Q, Casalicchio G, Stachl C, Bischl B (2017). “Multilabel Classification with R Package mlr.”
8. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
9. Rohan Alexander and Sam Caetano (2020). Source R code for cleaning 2017 GSS Survey Data.
10. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.

Appendix

The Github repo link for the source files of the study cound be found in the website: https://github.com/Jaydenhu123/STA304_PS2/.