Replicating research on marriage, happiness, and income inequality

Data 598 A, Winter 2020

Lauren Heintz, Tara Wilson, Ben Mathieu and Will Wright

Repository:
https://github.com/UW-MSDS-DATA-598-Reproducibility-WI20/brodeur-heintz-wilson-wright-replication-project

Paper:
Grunberg, Rebecca L. (Sloan School Of Management, MIT); Kim, Hyejun (Sloan School Of Management, MIT); Kim, Minjae (Sloan School Of Management, MIT). (2014). Marriage and happiness Grunberg Kim Kim.pdf [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/25655/MB980L

Marriage and happiness: Providing evidence against a relationship between inequality and happiness

WHAT

- Previous study by Oishi, Kesebir, and Diener (2011) shows that Americans' average happiness is negatively correlated with income inequality
  - This paper (chosen) questions the validity of this data and this papers conclusions, and seeks to disprove the relationship with a linear regression that accounts for several other factors including marriage, race, and gender.
  - Ultimately, this study finds no significant relationship to income inequality but instead to marriage

WHY

- Interest in the social sciences, something applicable that we can all relate to
- Data available
- Primary language used was R

Scientific claims of interest

Paper by Grunberg, Kim, & Kim (2014) creates a multi-level linear regression model, accounting for interactions between other variables to sort out the ultimate relationship to the happiness of Americans from years 1972-2012

Dependent Variable:

Happiness

Independent Variables:

- Gini coefficient → **Statistical measure between** 0 and 1 **for the distribution of wealth in a nation. Higher number indicates greater income inequality**
- Income
- Age
- Sex
- Race
- Married

Original Study Findings

Fig. 1. Income inequality in the United States from 1947 to 2009, as indexed by the Gini coefficient.

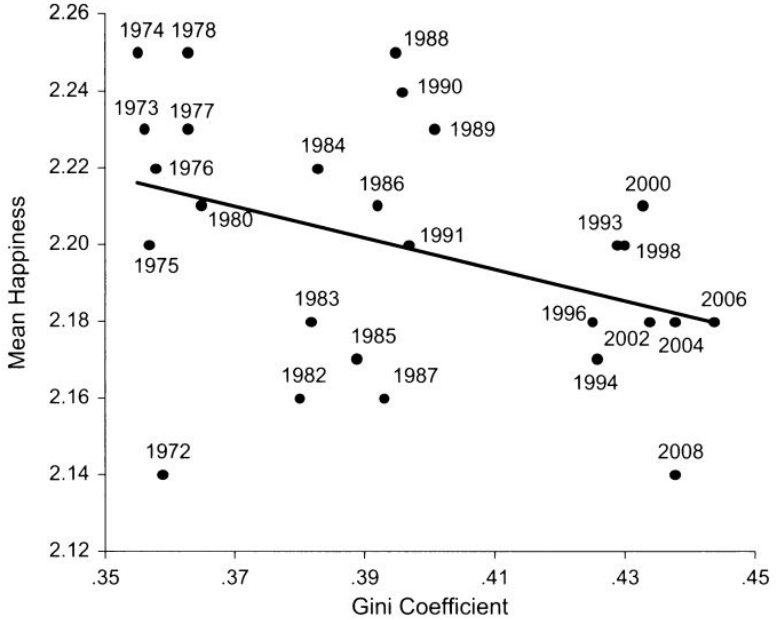


Fig. 2. Scatter plot (with best-fitting regression line) showing mean American happiness scores as a function of income inequality, as indexed by the Gini coefficient, from 1972 to 2008.

Replication results for figure 3

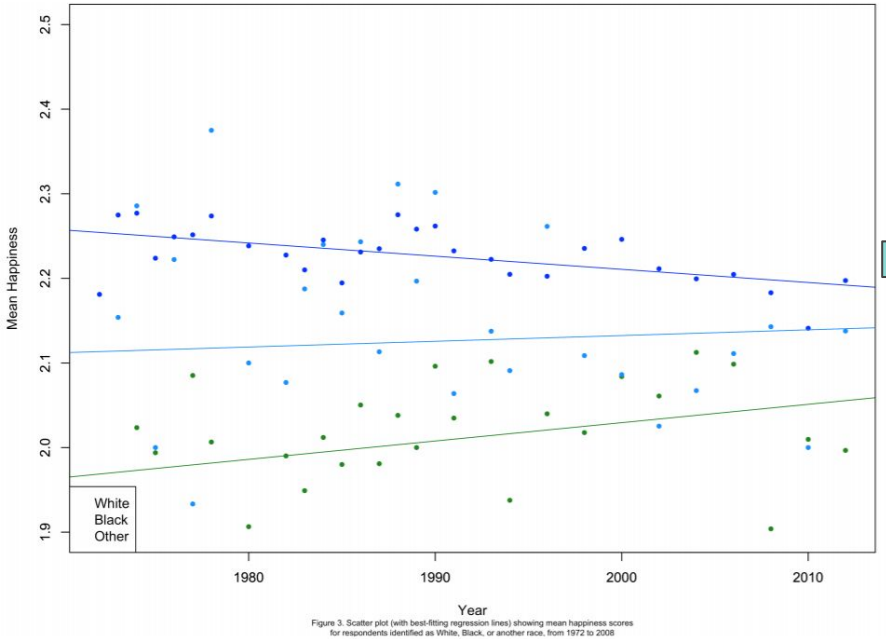


Figure 3. Scatter plot (with best-fitting regression lines) showing mean happiness scores for respondents identified as White, Black, or another race, from 1972 to 2008

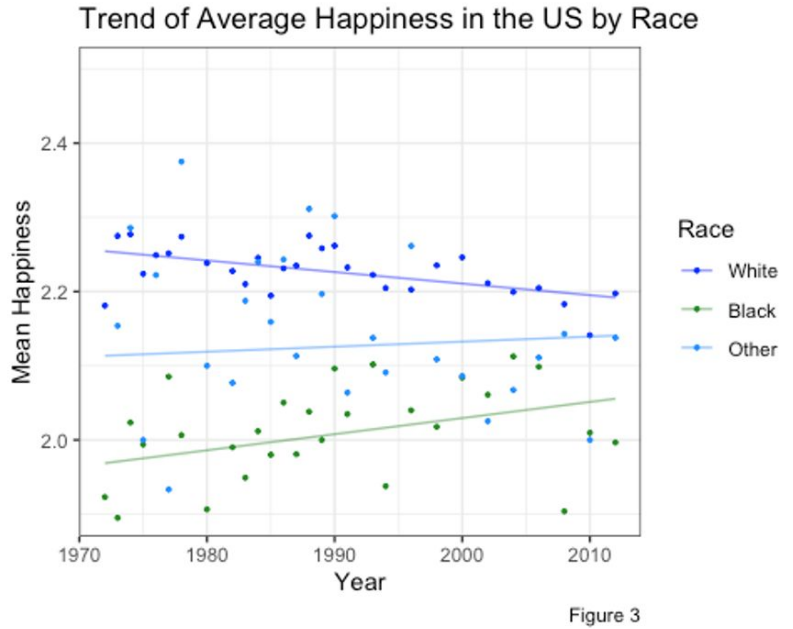
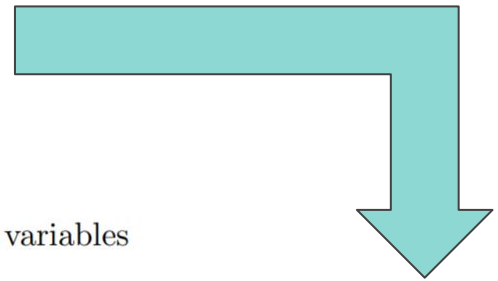


Figure 3

Replication results for table 1

	Estimate	p value
Logged Income	0.19	0.00
Age	0.03	0.00
Sex	0.01	0.23
White	0.12	0.00
Married	0.23	0.00

Table 1: Correlations between happiness and control variables



	Estimate	p value
Logged Income	0.19	0.00
Age	0.03	0.00
Sex	0.00	0.23
White	0.12	0.00
Married	0.23	0.00

Table 1: Correlations between happiness and control variables

Replication results for table 2

Table 2: Multilevel Linear Models

	Dependent variable:					
	Happiness					
	(1)	(2)	(3)	(4)	(5)	(6)
Gini	−0.385**	−0.461**	−0.413**	−0.386**	−0.177	0.332
	(0.194)	(0.189)	(0.195)	(0.194)	(0.175)	(0.210)
REALINClog		0.125***				
		(0.003)				
Age			0.001***			
			(0.0002)			
factor(SEX)2				0.007		
				(0.006)		
White					0.188***	
					(0.008)	
Married						0.298***
						(0.006)
Constant	2.351***	1.140***	2.306***	2.348***	2.112***	1.899***
	(0.077)	(0.081)	(0.078)	(0.077)	(0.070)	(0.084)
Observations	48,318	43,564	48,318	48,318	48,318	48,318
Log Likelihood	−46,611.120	−41,048.910	−46,590.320	−46,614.720	−46,308.240	−45,295.860
Akaike Inf. Crit.	93,230.250	82,107.830	93,190.630	93,239.430	92,626.470	90,601.730
Bayesian Inf. Crit.	93,265.390	82,151.240	93,234.560	93,283.360	92,670.400	90,645.650
Note:						*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:					
	Happiness					
	(1)	(2)	(3)	(4)	(5)	(6)
Gini	−0.385**	−0.461**	−0.424**	−0.386**	−0.177	0.331
	(0.194)	(0.189)	(0.196)	(0.194)	(0.175)	(0.210)
REALINClog		0.125***				
		(0.003)				
Age			0.001***			
			(0.0002)			
factor(sex)2				0.007		
				(0.006)		
White					0.188***	
					(0.008)	
Married						0.298***
						(0.006)
Constant	2.351***	1.140***	2.309***	2.348***	2.112***	1.900***
	(0.077)	(0.081)	(0.078)	(0.077)	(0.070)	(0.084)
Observations	48,318	43,564	48,158	48,318	48,318	48,306
Log Likelihood	−46,611.120	−41,048.920	−46,419.990	−46,614.720	−46,308.240	−45,282.940
Akaike Inf. Crit.	93,230.250	82,107.830	92,849.970	93,239.430	92,626.470	90,575.880
Bayesian Inf. Crit.	93,265.390	82,151.240	92,893.890	93,283.360	92,670.400	90,619.810
Note:						*p<0.1; **p<0.05; ***p<0.01

Highlights of replication

Primary Divergence:

- Finding new data set (GSS, Global Social Survey)
- Merging to the other given data sets, as well as some encoding updates needed to match variable use in original paper
- Narrowing date range to match that of the original paper

Secondary Divergences:

“Figure 3”: Utilized ggplot2 package instead of standard R visualization libraries

“Table 1”: Additionally used kableExtra to produce output table instead of xtable alone

“Table 2”: Utilized new data which led to slight variability of some cells, utilized kable package in addition to stargazer

Contribution of members

Benjamin Brodeur Mathieu:

Debugged original code, searched for new dataset, wrote the methods and conclusion sections of the report, reviewed the paper and the presentation for typos and bugs

Tara Wilson:

Debugged original code, replicated table 1 & 2, setup git repository, updated readme with guidances, reviewed paper

Will Wright:

Debugged original code, searched for new dataset, replicated figure 3, removed rendering bugs, added dynamic plotting

Lauren Heintz:

Wrote introduction and abstract, added references to original paper, setup RMD and other paper requirements, reviewed paper, wrote presentation

Challenges & reflections

1. R code did not run because data files were missing columns
2. Missing columns were critical data
3. Finding the data online was difficult, but once we found the data modifying it and merging it with the original dataset went relatively smoothly.

Thanks!