Waseda University

School of Political Science and Economics

Empirical report: Does education effect happiness in America?

Daniel Fabio Groth, student ID: 1A249134-9

Econometrics, Fall 2024

Table of contents

List of Tables

List of Figures

Introduction

In this report, I will use data from the 2018 GSS to investigate the relationship between education and happiness. Specifically, I will examine whether individuals with higher levels of education are more likely to report being happy. I will also explore whether this relationship varies between different demographic groups.

Research background

These questions are important because they can help us understand the factors that contribute to well-being and happiness. Previous research has found that education is positively associated with well-being, with individuals with higher levels of education reporting higher levels of happiness and life satisfaction. However, the relationship between education and happiness is complex and may vary between different groups of people. For example, some studies have found that the relationship between education and happiness is stronger for individuals because of social networks and involvement with the wider world, and that having higher educations means being employed and having higher income levels which are positively related with happiness.

Literature Review

Previous research has found a positive relationship between education and political ideology. For example, a study by Chen (2012) says that "In short, individuals who receive more education have more extensive social networks as well as greater involvement with the wider world; these life conditions are positively related with happiness". Another study by Cuñado & De Gracia (2012) says that "we find that people with a higher education level have higher income levels and a higher probability of being employed, and thus, report higher levels of happiness". In my research, I will build on this previous work by examining the relationship between education and happiness on americans but also include different demographic groups with data from the 2018 GSS exploring.

Hypothesis

I hypothesize that individuals in america with higher levels of education will be more likely to report being happy. This hypothesis is based on previous research that has found a positive relationship between education and well-being. I also expect that this relationship will be stronger for certain demographic groups, such as younger individuals.

Preview of findings

Using the 2018 GSS data, and a logistic regression model, I will examine the relationship between education and happiness. I will also control for demographic variables such as age, sex, marital status and respondents income to see if the relationship between education and happiness holds after controlling for these variables. I expect to find that individuals with higher levels of education are more likely to report being happy, and that this relationship is stronger for certain demographic groups. What I did find was that the relationship between education and happiness was negative, which was contrary to what I expected. Interestingly, it seems that being divorced or never married is a significant predictor of happiness. This could be due to the fact that individuals who are divorced or never married have more freedom and autonomy, which could contribute to their happiness.

Roadmap

The report is structured as follows. In the next section, I will describe the data and variables used in the analysis. In the following section, I will present the results of the analysis, including descriptive statistics and regression analysis. Finally, I will discuss the implications of the findings and suggest directions for future research.

Data

The General Social Survey (GSS) is a survey conducted in the United States to monitor social change and study the growing complexity of American society. The survey is conducted every two years and is designed to provide a snapshot of the opinions and behaviors of the American people. The GSS collects data on a wide range of topics, including attitudes towards social issues, political beliefs, and demographic characteristics. The data is publicly available for download from the GSS website here.

Variables

The variables going to be used in this analysis can be seen in the following Table 1.

Variable	Definition
Sex	Sex of respondent (1 = Male, 2 = Female)
Age	Age in years at time of interview
Race	Race (1 = White, 2 = Black, 3 = Other)
Education	Highest year of school completed
Happy	Self-rated happiness (0 = Not happy, 1 = Happy)
Marital	Marital status (From 0-5)
Respondents Income	(Income variable from 0-12)

Table 1: Variables and Definitions (GSS 2018)

Education is measured as the highest year of school completed, ranging from 0 to 20 years. Happiness is measured as a self-rated happiness score, ranging from 1 to 3, with 1 being "Very Happy", 2 being "Pretty Happy", and 3 being "Not Too Happy". In this analysis, I will use education as the independent variable and happiness as the dependent variable. I will also transform the happiness variable into a binary variable for the regression analysis, which will consist of "Not Happy" and "Happy", as is seen above in Table 1.

Marital status are measured from 0 to 5 where 1 is "Married", 2 is "Widowed", 3 is "Divorced", 4 is "Separated" and 5 is "Never married". Respondents income is measured from 0 to 12 where 0 is "Under 1000", 1 is "1000-2999", 2 is "3000-3999", 3 is "4000-4999", 4 is "5000-5999", 5 is "6000-6999", 6 is "7000-7999", 7 is "8000-9999", 8 is "10000-12499", 9 is "12500-14999", 10 is "15000-17499", 11 is "17500-19999" and 12 is "20000 and over".

Descriptive Statistics

In this section, I will present descriptive statistics for the variables used in the analysis. This will provide an overview of the sample and help to identify any patterns or trends in the data.

	AGE	EDUC	HAPPY	MARITAL	RACE	SEX	RINCOME
Mean	48.93	13.63	1.71	2.72	1.42	1.55	6.49
Std.Dev	18.11	2.83	0.45	1.67	0.69	0.50	5.55
Min	18.00	0.00	1.00	1.00	1.00	1.00	0.00
Q1	34.00	12.00	1.00	1.00	1.00	1.00	0.00
Median	48.00	13.00	2.00	3.00	1.00	2.00	9.00
Q3	63.00	16.00	2.00	5.00	2.00	2.00	12.00
Max	89.00	20.00	2.00	5.00	3.00	2.00	12.00
MAD	22.24	1.48	0.00	2.97	0.00	0.00	4.45
IQR	29.00	4.00	1.00	4.00	1.00	1.00	12.00
CV	0.37	0.21	0.27	0.61	0.49	0.32	0.85
Skewness	0.29	-0.42	-0.93	0.30	1.37	-0.20	-0.20
SE.Skewness	0.11	0.11	0.11	0.11	0.11	0.11	0.11
Kurtosis	-0.87	1.87	-1.14	-1.53	0.41	-1.96	-1.85
N.Valid	536.00	536.00	536.00	536.00	536.00	536.00	536.00
Pct.Valid	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 2: Descriptive Statistics for GSS 2018 (N = 536)

The mean age of the sample is 48.93 years, with a standard deviation of 18.11. The mean education level is 13.63 years, with a standard deviation of 2.83. The mean happiness score is 1.7, with a standard deviation of 0.45. The sample is predominantly white, which can make the results biased, if the sample is not representative of the population. The Skewness is negative for the happiness variable, which indicates that the distribution is left-skewed, this is also true for education. The Kurtosis is negative for the happiness variable, which indicates that the distribution is platykurtic. The sample size is 536, with no missing values for any of the variables used in the analysis.

Data Visualization

In this section, I will present visualizations of the data to help illustrate the relationships between the variables. This will provide a more intuitive understanding of the data and help to identify any patterns or trends.

Figure 1: Distribution of Happiness Scores

Figure 1 shows the two categories of the happiness variable, with the majority of respondents reporting being happy. This is consistent with previous research that has found that the majority of people report being happy. The distribution of the happiness variable is left-skewed, and this could be shown better before merging the "Happy" variable with "Pretty Happy" and "Very Happy". I acknowledge that this could be a limitation of the analysis, and that the results may be biased because of merging the categories.

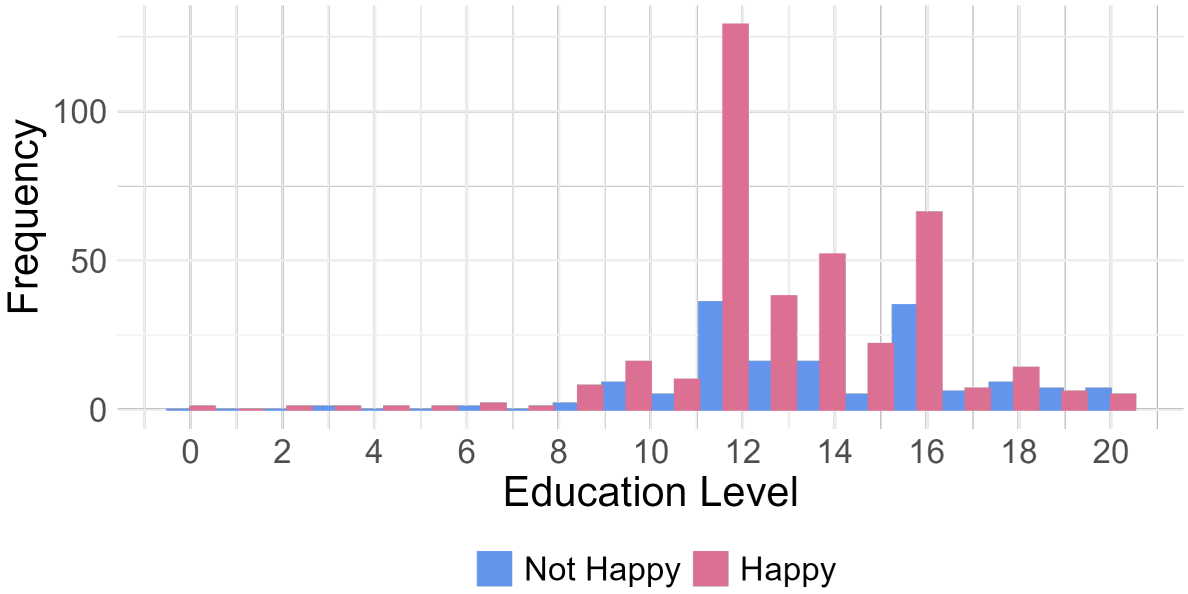

Figure 2: Distribution of Education Levels

Figure 2 shows that the distribution of education levels is roughly normal, with a peak around 12 years of education. This is consistent with the fact that most people in the sample have completed high school, which is typically 12 years of education. As the distribution of education levels is right-modal, the normality assumption for the regression analysis is not violated.

Lets look at the relationship between education and happiness.

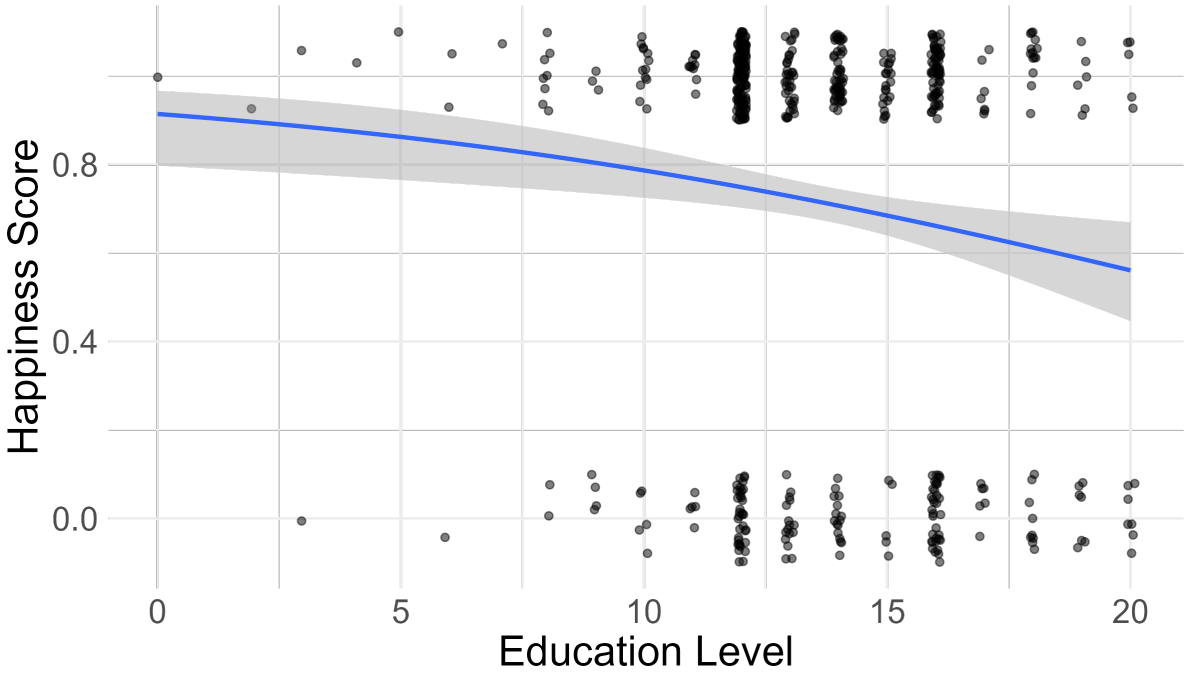

Figure 3: Relationship between Education and Happiness

Figure 3 shows the relationship between education and happiness. The scatterplot shows that there is a negative relationship between education and happiness, with individuals with higher levels of education being less likely to report being happy. This is contrary to what I expected, and could be due to the fact that the sample is not representative of the population, or that there are other factors influencing the relationship between education and happiness. More analysis is needed to understand this relationship better.

Regression Analysis

In this section, I will present the results of the regression analysis examining the relationship between education and happiness. I will estimate a logistic regression model to determine whether education is a significant predictor of happiness, controlling for demographic variables.

Model Specification

The model I am going to use is called an Logistic Regression model, which is a type of regression analysis used to predict the probability of a binary outcome. In this case, the dependent variable is happiness, which is a binary variable indicating whether an individual is happy or not happy. The independent variable of interest is education, which is a continuous variable representing the highest year of school completed.

The logistic regression model I am going to use is specified as follows:

$$log(\frac{P(\text{Happy} = 1)}{1 - P(\text{Happy=1})}) = \beta_0 + \beta_1 \times \text{Education}$$

where β_0 is the intercept, β_1 is the coefficient for education. The coefficient β_1 represents the change in the log-odds of being happy for a one-unit increase in education. A positive coefficient indicates that higher levels of education are associated with a higher probability of being happy, while a negative coefficient indicates the opposite. The left side of the equation is the log-odds of being happy, which is transformed to a probability using the logistic function. This model has no error term as it is captured by the binoimal distribution, not by an error term.

```
================================================
                        Dependent variable:
                    ---------------------------
                              happy_binary
------------------------------------------------
EDUC                           -0.106***
                                (0.036)

Constant                       2.362***
                                (0.507)

------------------------------------------------
Observations                      536
Log Likelihood                 -317.731
Akaike Inf. Crit.               639.461
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

The results of the logistic regression model are presented in the table above. The coefficient for education is negative, which is statistically significant at the 0.01 level. This indicates that for a one-unit increase in education, the log-odds of being happy decrease by 0.106. This is contrary to what I expected, and suggests that higher levels of education are associated with a lower probability of being happy. However, it is important to note that this relationship may be influenced by other factors, and further analysis is needed to understand the relationship between education and happiness better.

Adding Demographic Variables

In this section, I will control for demographic variables such as age, sex, marital status and respondents income to see if the relationship between education and happiness holds after controlling for these variables.

The new model is specified as follows:

$$log(\frac{P(\text{Happy} = 1)}{1 - P(\text{Happy=1})}) = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Age} + \beta_3 \times \text{Marital status} + \beta_4 \times \text{Respondents income}$$

where β_0 is the intercept, β_1 is the coefficient for education, β_2 is the coefficient for age, β_3 is the coefficient for marital status, and β_4 is the coefficient for respondents income. The coefficients β_1 represents the change in the log-odds of being happy for a one-unit increase in education, holding all other variables constant. A positive coefficient indicates that higher levels of education are associated with a higher probability of being happy, while a negative coefficient indicates the opposite. The same goes for the other coefficients. The left side of the equation represents the log-odds of being happy, and the model has no error term as it is modeled through the binomial distribution.

	Dependent variable:	
	happy_binary	
	(1)	(2)
EDUC	-0.106***	-0.091*
	(0.036)	(0.052)
SEX2		0.497
		(1.040)
AGE		0.002
		(0.007)
MARITALWidowed		0.683*
		(0.396)
MARITALDivorced		1.461***
		(0.305)
MARITALSeparated		0.699
		(0.608)
MARITALNever married		1.413***
		(0.281)
RINCOME		0.013
		(0.020)
EDUC:SEX2		-0.029
		(0.073)

Constant	2.362***	1.237
	(0.507)	(0.835)
Observations	536	536
Log Likelihood	-317.731	-295.010
Akaike Inf. Crit.	639.461	610.020
Note:	*p<0.1; **p<0.05; ***p<0.01	

The results of the logistic regression model with demographic variables are presented in the table above. The coefficient for education is still negative, but now only statistically significant at the 0.1 level. Which means that the relationship between education and happiness is not as strong as before.

By using an interaction term between education and sex we can see if the relationship between education and happiness is different for the two genders, while both are not statistically significant, the coefficient for the interaction term is negative for male and positive for females which could indicate that the relationship between education and happiness is different for the two genders, but not by much which can be seen in Figure 4.

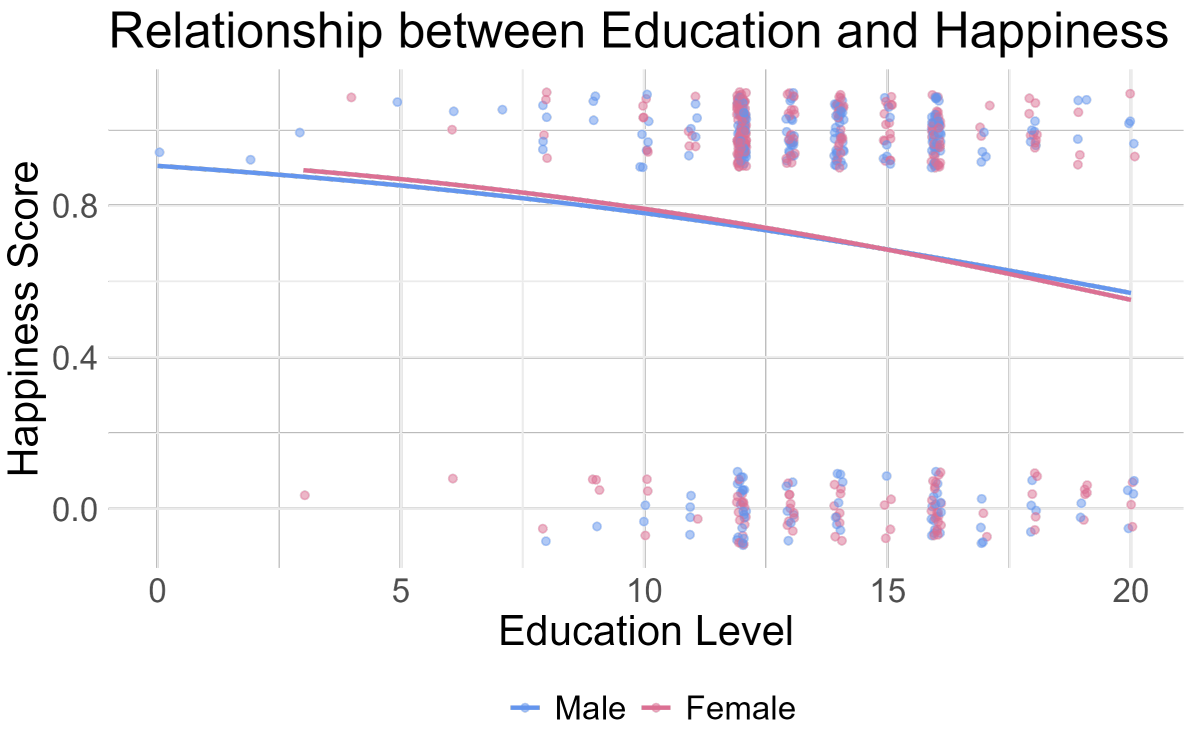

Figure 4: Relationship between Education and genders

Age does not seem to be a significant predictor of happiness, with a coefficient of 0.002, and the same can be said for being separated and the respondents income. But looking at marital status, being divorced and never married seems to be significant predictors of

happiness, with a positive coefficient of 1.461 and 1.413 respectively. This indicates that individuals who are divorced or never married are more likely to report being happy, compared to individuals who are married. Albeit not what I was looking for, it is interesting to see that being divorced or never married is a significant predictor of happiness.

Model Comparison

I will do an ANOVA test which is a statistical test used to compare the fit of two or more models to determine which model is the best fit for the data. In this case, I will compare the first model, which only includes education, with the second model, which includes demographic variables. The null hypothesis is that the two models are equally good fits for the data, while the alternative hypothesis is that the second model is a better fit for the data.

```
Analysis of Deviance Table

Model 1: happy_binary ~ EDUC
Model 2: happy_binary ~ EDUC * SEX + AGE + MARITAL + RINCOME
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       534     635.46
2       526     590.02  8   45.442 3.034e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Doing an ANOVA test between the two models, we can see that the second model is a better fit for the data, with a p-value of 0.001. This indicates that the second model, which includes demographic variables, is a better predictor of happiness than the first model, which only includes education. But the model might be overfitted, because of adding too many variables and splitting the sample size into smaller groups.

Conclusion

In this report, I have examined the relationship between education and happiness using data from the 2018 GSS. I found that individuals with higher levels of education are less likely to report being happy, which is contrary to what I expected. This relationship holds even after controlling for demographic variables. Some weaknesses of the analysis is that the sample size is predominantly white, which can make the results biased, if the sample is not representative of the population. That is also why I did not include race in the model, because the samples would be too small for each group. Another weakness is that I transformed the happiness variable into a binary variable, which could have biased the results when they were merged.

Overall, the results of the analysis suggest that the relationship between education and happiness is complex and may be influenced by other factors not included in this analysis.

Interestingly, I found that being divorced or never married is a significant predictor of happiness, with individuals who are divorced or never married being more likely to report being happy compared to individuals who are married. This is an unexpected finding, and further research is needed to understand the relationship between marital status and happiness better.

Usage of AI

Copilot

As I am writing this document in Rstudio, there is an integration of copilot which sometimes automatically suggests code snippets. Sometimes it works great and my latex math gets written perfectly, and other times it just gives me a bunch of random unrelevant latex math or code.

Here is a Link to copilot.

ChatGPT

Here is the link to the conversation where I asked questions:

Link to ChatGPT

Appendix

```
rm(list = ls())
library(tidyverse)
library(stargazer)
# Load data
url <- "https://raw.githubusercontent.com/DanielFabioG/data/refs/heads/main/GSSdata2018_

data <- read.csv(url)

# Clean the dataset for only the variables I need for the analysis
data <- data %>%
  select(SEX,AGE,RACE,EDUC,HAPPY, MARITAL, RINCOME)

  # Making the happy variable binary
data$happy_binary <- ifelse(data$HAPPY == 1, 0, 1)

# Change the happy_binary variable to a categorical "YES, NO" variable
data$happy_categorical <- factor(data$happy_binary,
                                 levels = c(0, 1), labels = c("Not Happy", "Happy"))

# Make marital status a factor
data$MARITAL <- factor(data$MARITAL,
                       levels = c(1, 2, 3, 4, 5),
                       labels = c("Married", "Widowed", "Divorced", "Separated", "Never

# Make race factor
```

```r
data$RACE <- factor(data$RACE,
                    levels = c(1, 2, 3),
                    labels = c("White", "Black", "Other"))

# Descriptive statistics
#descr(data, style = "rmarkdown")

# Plot the distribution of the happiness variable
fig1 <-  ggplot(data, aes(x = happy_categorical,
                          fill = happy_categorical)) +
  geom_bar() +
  labs(title = "",
       x = "Happiness Score",
       y = "Frequency",
       fill = "")+
  theme_minimal()+
  # need to fix the sizes of all text for ggsave so it gets bigger
  theme(text = element_text(size = 20))+
  scale_fill_manual(values = c("cornflowerblue", "palevioletred"))+
  theme(legend.position = "none")

ggsave("documentobjects/figures/happiness_distribution.png",
       plot = fig1, width = 8, height = 4.5, dpi = 300)

# Create a histogram of the education variable
fig2 <- data %>%
ggplot(aes(x = EDUC, fill = happy_categorical, color = happy_categorical)) +
  geom_histogram(bins = 20, position = "dodge") +
  labs(title = "",
       x = "Education Level",
       y = "Frequency",
       color = "",
       fill = "")+
  theme_minimal()+
  scale_fill_manual(values = c("cornflowerblue", "palevioletred"))+
  scale_color_manual(values = c("cornflowerblue", "palevioletred"))+
  theme(legend.position = "bottom")+
  scale_x_continuous(breaks = seq(0, 20, by = 2))+
  theme(text = element_text(size = 20))

ggsave("documentobjects/figures/education_distribution.png",
       plot = fig2, width = 8, height = 4.5, dpi = 300)

# Create a scatterplot of education and happiness
fig3 <- ggplot(data, aes(x = EDUC, y = happy_binary)) +
  geom_jitter(width = 0.1, height = 0.1, alpha = 0.5) +
  labs(title = "",
```

```
       x = "Education Level",
       y = "Happiness Score")+
  stat_smooth(method="glm", method.args=list(family="binomial"))+
  theme_minimal()+
  theme(text = element_text(size = 20))

ggsave("documentobjects/figures/education_happiness.png",
       plot = fig3, width = 8, height = 5, dpi = 300)

# fix SEX variable to be a factor
data <- data %>%
  mutate(SEX = factor(SEX, levels = c(1, 2), labels = 1:2))

# Fit the ordered logistic regression model
model.fit1 <- glm(happy_binary ~ EDUC, data = data, family = "binomial")

# Display the results
stargazer(model.fit1, type = "text")

# Fit the ordered logistic regression model with demographic variables

model.fit2 <- glm(happy_binary ~ EDUC * SEX + AGE + MARITAL + RINCOME,
                  data = data, family = "binomial")

# Display the results
stargazer(model.fit1, model.fit2, type = "text")

fig3 <- ggplot(data = data, mapping = aes(x = EDUC, y = happy_binary, color = SEX)) +
  geom_jitter(width = 0.1, height = 0.1, alpha = 0.5) +
  stat_smooth(method="glm", method.args=list(family="binomial"), se = FALSE)+
  theme_minimal()+
  theme(text = element_text(size = 20))+
  labs(title = "Relationship between Education and Happiness",
       x = "Education Level",
       y = "Happiness Score",
       color = "",
       )+
  scale_color_manual(values = c("cornflowerblue", "palevioletred"),
                     labels = c("Male", "Female"))+
  theme(legend.position = "bottom")

ggsave("documentobjects/figures/education_sex.png",
       plot = fig3, width = 8, height = 5, dpi = 300)

# Anova
anova(model.fit1, model.fit2, test = "Chisq")
```

References

Chen, W. (2012). How education enhances happiness: Comparison of mediating factors in four east asian countries. *Social Indicators Research*, *106*, 117–131.

Cuñado, J., & De Gracia, F. P. (2012). Does education affect happiness? Evidence for spain. *Social Indicators Research*, *108*, 185–196.