Waseda University School of Political Science and Economices

# Econometrics I Research Paper

LIEN YU HSIANG

1A202G34

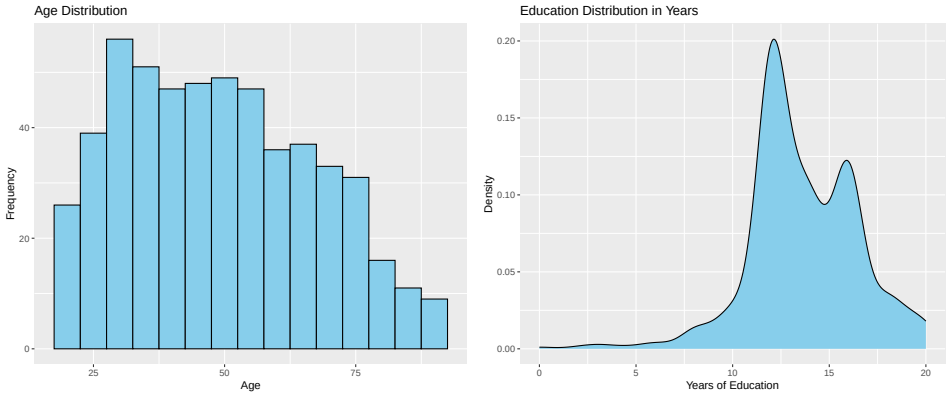
January 29, 2024

# 1 Introduction

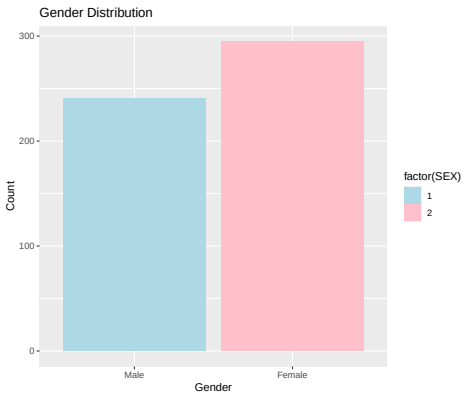
This paper endeavors to conduct a thorough examination of the dataset 'GSSdata2018.csv' through the application of statistical methods such as regression tests and hypothesis testing. The primary objective is to derive meaningful insights from the data, uncovering patterns and relationships that may exist. The use of robust statistical techniques ensures the reliability of our findings. The 'GSSdata2018.csv' dataset serves as the focal point for our analysis, with regression tests enabling us to explore connections between different variables. Additionally, hypothesis testing is employed to scrutinize specific assumptions, adding a layer of statistical rigor to our conclusions. The intention is to delve beyond superficial observations, utilizing meticulous statistical analysis to reveal nuanced trends within the data. By employing reliable statistical testing methods, the paper aims to establish credible and insightful conclusions, contributing to a more comprehensive understanding of the dataset and facilitating the proposal of informed research outcomes.

# 2 Data Description

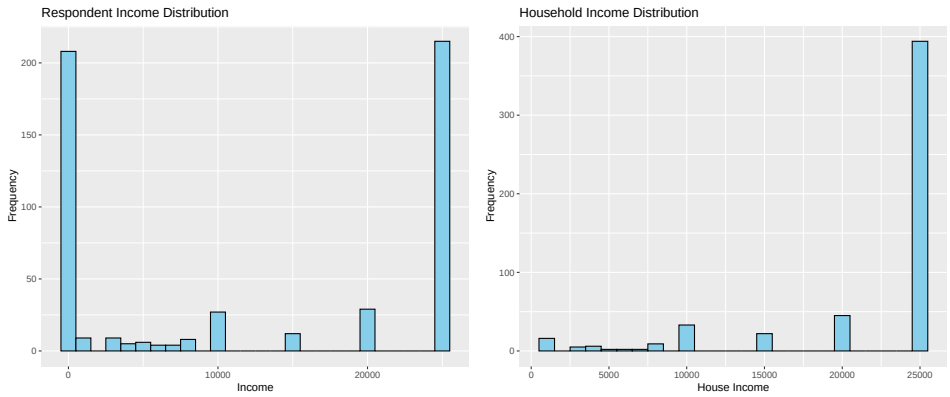
The dataset been utilized is 'GSSdata2018', which is a dataset that made by nationally representative survey of adults in the United States conducted since 1972. And the dataset is publically avalible so anyone who wish to discover further insight can obtain the data. Before accessing the data through statistical means, it is important to fist look and summerized some basic but vital aspect of the data.



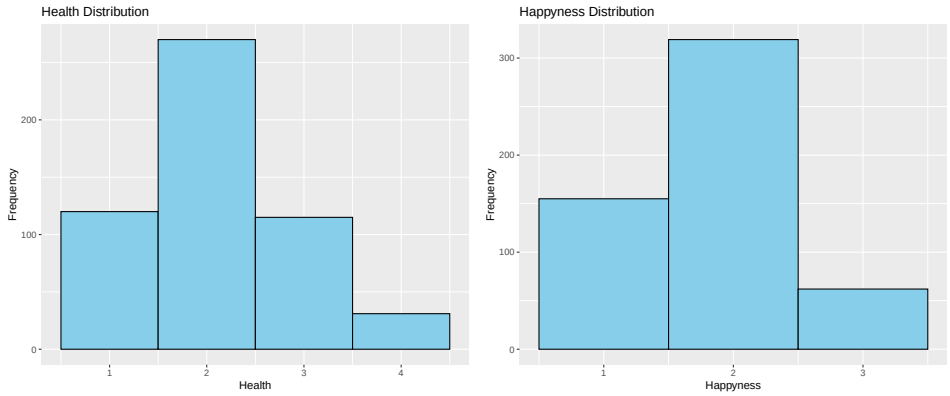

The upper two graphs depicting the distribution of recipient age and education, notable patterns emerge. The predominant age group among recipients falls within the range of 25 to 50 years. Simultaneously, a significant proportion of recipients possess either 12 or 16 years of education, pointing towards a prevalence of high school or university graduates within the dataset. This observation highlights the demographic composition of the majority of recipients, indicating a concentration of individuals who have completed their education at either the high school or university level.




The above graph simply showcase the distribution of gender within the dataset. Which we can see that the number of different is not very big, indicating a farily eqaul distribution among gender.




Examining the two upper graphs representing personal income and household income for each recipient, a noteworthy pattern emerges. It appears that the number of individuals reporting $0 in income is approximately equal to those reporting incomes exceeding $250,000. This observation prompts the hypothesis that individuals with high incomes may have a partner who stays at home and generate low or none income, although further statistical analysis is required to draw definitive conclusions.




Before making any accessment, we first have to notice that for both health and happiness, the lower the number the better they are. So after examining the upper two graphs, it is evident that the health condition of recipients tends to be 2 out of 4 on the scale, which indicates a good health condition. Additionally, the level of happiness observed is consistently moderate, indicating a mid-level of overall well-being, which is pretty happy.

# 3 Econometrics Models

The econometrics that is utilized for the statistical analysis is listed below:

Model 1. $health = \beta_0 + income\beta_1 + \epsilon$

Model 2. $happy = \beta_0 + \beta_1 health + \beta 2income + \epsilon$

Model 3. $education = \beta_0 + \beta_1 father - education + \beta_2 mother - education + \beta_3 spouse - education + \epsilon$

Model 4. $logincome = \beta_0 + \beta_1 logspouse - income + \epsilon$

Model 5. $income = \beta_0 + \beta_1 age + \beta_2 education + \beta_3 babies + \epsilon$

Model 6. $logincome = \beta_0 + \beta_1 age + \beta_2 education + \beta_3 babies + \epsilon$

(The seperation of male and female dataset is utilized for Model 5 and 6.)

# 4 Estimation Results

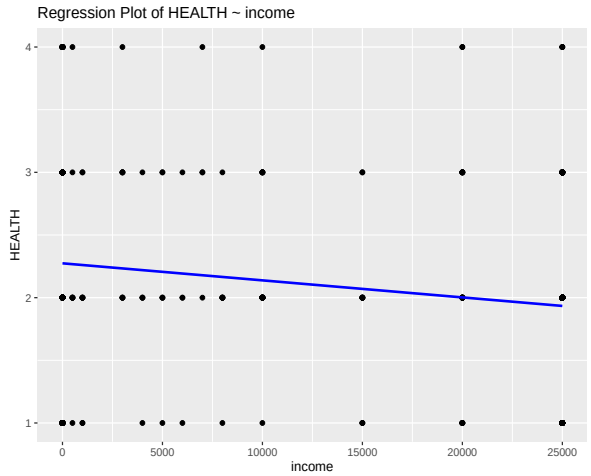
**Model 1** $health = \beta_0 + income\beta_1 + \epsilon$

This model aim to discover the relationships between health and income. The usual assumption will be that for a richer person, his health condition should be better since he have access to better medication resources.

health ~ income

| | (1) |
|---|---|
| (Intercept) | 2.274*** |
| | (0.051) |
| income | 0.000*** |
| | (0.000) |
| Num.Obs. | 536 |
| R2 | 0.037 |
| R2 Adj. | 0.035 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Standard errors in parentheses.




By inspecting the data, we can see that a person's health condition do not depend on once's income, which is different from usual assumption.

**Model 2** $happy = \beta_0 + \beta_1 health + \beta 2 income + \epsilon$

This model seeks to investigate the links between happiness, health, and income, operating on the assumption that individuals tend to experience greater happiness when they enjoy both financial prosperity and good health.

happy ~ health + income

| | (1) |
|---|---|
| (Intercept) | 1.293*** |
| | (0.081) |
| HEALTH | 0.245*** |
| | (0.032) |
| income | 0.000 |
| | (0.000) |
| Num.Obs. | 536 |
| R2 | 0.102 |
| R2 Adj. | 0.099 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Standard errors in parentheses.


By inspecting the table and graph, we can find out that happiness is affect by health but not income. More specificly, a 1 increase in scale of healthiness increase 0.245 in scale of happiness. So we can conclude that the usual assumption is partially support by the empirical data and statistical analysis.

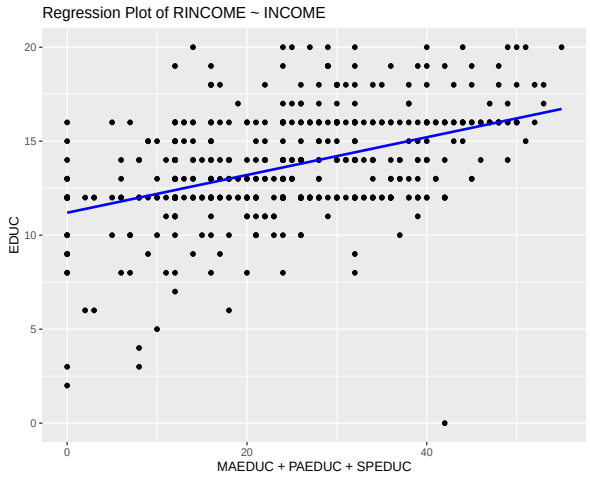
**Model 3** $education = \beta_0 + \beta_1 father - education + \beta_2 mother - education + \beta_3 spouse - education + \epsilon$

This model aims to explore the interconnections between an individual's educational attainment and the educational backgrounds of their parents and spouse. Conventionally, it's often presumed that there exists a strong correlation between an individual's level of education and that of their parents and spouse.

education ~ fother/mother/spouse education

| | (1) |
|---|---|
| (Intercept) | 10.619*** |
| | (0.251) |
| MAEDUC | 0.185*** |
| | (0.021) |
| PAEDUC | 0.106*** |
| | (0.018) |
| SPEDUC | 0.035* |
| | (0.015) |
| Num.Obs. | 536 |
| R2 | 0.251 |
| R2 Adj. | 0.247 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Standard errors in parentheses.




Upon examining the table, it becomes evident that there exists a strong correlation between an individual's level of education and that of their parents. To be precise, for each additional year of education attained by the mother and father, the individual's education increases by approximately 0.185 and 0.106 years, respectively. However, the correlation between an individual's education and that of their spouse appears to be relatively weak and lacks statistical significance.

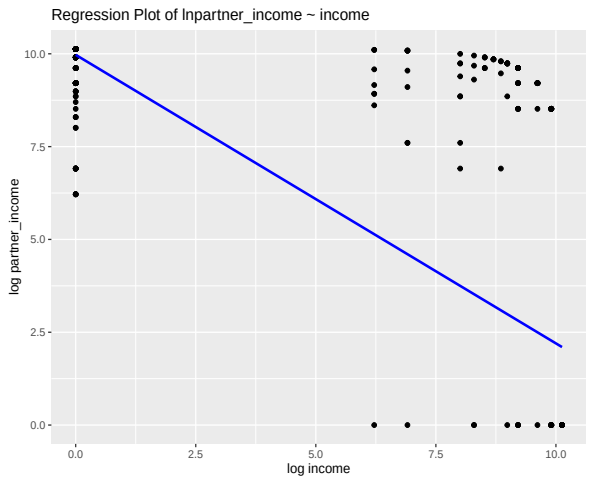
**Model 4** $logincome = \beta_0 + \beta_1 logspouse - income + \epsilon$

This model seeks to explore the association between the logarithm of an individual's personal income and the logarithm of their spouse's income. Conventionally, it is assumed that higher personal income for one individual might correlate with their spouse having the option to be a stay-at-home parent or not work, thereby implying a negative correlation between the two variables.

Log partner_income ~ Log income

| | (1) |
|---|---|
| (Intercept) | 9.972*** |
| | (0.213) |
| lnincome | −0.777*** |
| | (0.028) |
| Num.Obs. | 533 |
| R2 | 0.596 |
| R2 Adj. | 0.595 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Standard errors in parentheses.




Upon inspecting the table, it becomes evident that there exists a negative correlation between an individual's income and that of their spouse. To be precise, for every 1% increase in an individual's income, there is a corresponding decrease of 0.777% in their spouse's income. This finding aligns with

the conventional assumption. However, it is imperative to acknowledge that the dataset originates from the US government. Consequently, when examining countries where dual-income households constitute the predominant family structure, the relationship between personal income and spousal income may diverge.

**Model 5** $income = \beta_0 + \beta_1 age + \beta_2 education + \beta_3 babies + \epsilon$

**Model 6** $logincome = \beta_0 + \beta_1 age + \beta_2 education + \beta_3 babies + \epsilon$

Models 5 and 6 aim to explore the relationship between income or log-income and key determinants of income: age, education, and the number of dependents in the subject's family. According to labor economics theory, income typically exhibits a positive correlation with age, implying that income tends to rise as individuals age. Similarly, human capital theory suggests a positive relationship between income and education; individuals with higher levels of education generally possess greater human capital, resulting in higher earning potential. Additionally, the number of dependents in a family is expected to impact income, as greater financial resources are typically required to support a larger family size. Statistical analyses using these models were conducted on three distinct datasets: male, female, and combined datasets. By examining datasets categorized by gender, we aim to uncover gender-specific insights into how various factors influence income levels.

Income table

| | male_level | male_log | female_level | female_log | both | both_log |
|---|---|---|---|---|---|---|
| (Intercept) | 12446.056** | 8.050*** | 2630.627 | 6.418*** | 7.212*** | 7.212*** |
| | (3897.552) | (1.546) | (3760.614) | (1.577) | (1.102) | (1.102) |
| AGE | −180.286*** | −0.093*** | −135.079*** | −0.085*** | −0.088*** | −0.088*** |
| | (41.774) | (0.017) | (35.564) | (0.015) | (0.011) | (0.011) |
| EDUC | 744.565** | 0.204* | 1116.500*** | 0.269** | 0.233*** | 0.233*** |
| | (239.936) | (0.095) | (231.050) | (0.097) | (0.068) | (0.068) |
| BABIES | 779.652 | 0.429 | −669.234 | −0.553 | −0.242 | −0.242 |
| | (1807.740) | (0.717) | (1364.615) | (0.572) | (0.444) | (0.444) |
| Num.Obs. | 241 | 241 | 295 | 295 | 536 | 536 |
| R2 | 0.109 | 0.138 | 0.121 | 0.127 | 0.128 | 0.128 |
| R2 Adj. | 0.098 | 0.128 | 0.112 | 0.118 | 0.123 | 0.123 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Upon inspecting the data, several conclusions become apparent. Firstly, there is a negative correlation between age and income for both males and females. Additionally, education positively correlates with income for both genders, with each additional year of education leading to substantial increases in income. Specifically for males, a one-year increase in age corresponds to a 0.9% decrease in income, while each additional year of education results in a significant 20.4% increase. Furthermore, each additional baby in the family appears to positively impact male income by 42.9%, although this effect lacks statistical significance. In contrast, females experience similar negative correlations between age and income, albeit with more pronounced effects. A one-year increase in age corresponds to an 8.5% reduction in income

for females. Intriguingly, each additional baby in the family significantly decreases female income by 55.3%. Moreover, for females, each additional year of education results in a remarkable 26.9% increase in income, surpassing the growth rate observed in males. The disparity in the impact of children on income between genders is striking. While male income benefits from an increased number of babies, female income suffers. This observation suggests the presence of entrenched gender roles within society, influencing how family dynamics intersect with income generation. In summary, while age and education consistently influence income levels for both genders, the divergent effects of family size underscore the complex interplay between societal norms and economic outcomes for men and women.

# 5 Conclusion

In conclusion, this paper undertook a comprehensive analysis of the 'GSSdata2018' dataset, employing a variety of econometric models to uncover meaningful relationships among key variables. Through robust statistical testing, we aimed to elucidate nuanced patterns within the data, ultimately contributing to a deeper understanding of socio-economic dynamics in the United States. Our findings revealed intriguing insights into the determinants of health, happiness, educational attainment, and income levels.

Notably, while conventional assumptions regarding the positive relationship between income and health were not supported by our analysis, we discovered significant associations between education and income levels. For instance, each additional year of education correlated with a substantial increase in income for both males and females, with males experiencing a 20.4% boost and females a remarkable 26.9% surge. Additionally, our examination of family dynamics unveiled gender disparities in income outcomes, with each additional baby in the family positively impacting male income by 42.9% while significantly reducing female income by 55.3%. These findings underscore the complex interplay between individual characteristics, societal norms, and economic outcomes, providing valuable insights for future research and policy formulation.

# 6 Appendix

The appendix include the original R code of my analysis.

```
library(tidyverse)
library(modelsummary)
library(readr)
library(ggplot2)

setwd("/Users/y.h.lien/Desktop/Github/Econometrics-I")
data <- read_csv("GSSdata2018.csv")
View(data)

head(data)
dim(data)

summary(data)

data <- data %>% mutate(income =
                          500*(RINCOME == 1) +
                          1000*(RINCOME == 2) +
                          3000*(RINCOME == 3) +
                          4000*(RINCOME == 4) +
                          5000*(RINCOME == 5) +
                          6000*(RINCOME == 6) +
                          7000*(RINCOME == 7) +
                          8000*(RINCOME == 8) +
                          10000*(RINCOME == 9) +
                          15000*(RINCOME == 10) +
                          20000*(RINCOME == 11) +
                          25000*(RINCOME == 12) ) %>% mutate(lnincome = log(income + 1))

data <- data %>% mutate(house_income =
                          500*(INCOME == 1) +
                          1000*(INCOME == 2) +
                          3000*(INCOME == 3) +
                          4000*(INCOME == 4) +
                          5000*(INCOME == 5) +
                          6000*(INCOME == 6) +
                          7000*(INCOME == 7) +
                          8000*(INCOME == 8) +
                          10000*(INCOME == 9) +
                          15000*(INCOME == 10) +
                          20000*(INCOME == 11) +
                          25000*(INCOME == 12) ) %>% mutate(lnhouse_income = log(house_income + 1))

data <- data %>% mutate(partner_income = house_income - income) %>% mutate(lnpartner_income = log(par

data <- data %>% mutate(belief = 1*(HELL == 1)*(HEAVEN == 1))

mdata <- data %>% filter(SEX == 1)
fdata <- data %>% filter(SEX == 2)

# Data description
```

```
# Data visulization

# Age distribution
ggplot(data, aes(x = AGE)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")

# Education distribution
ggplot(data, aes(x = EDUC)) +
  geom_density(fill = "skyblue", color = "black") +
  labs(title = "Education Distribution in Years", x = "Years of Education", y = "Density")

# Gender distribution
ggplot(data, aes(x = factor(SEX), fill = factor(SEX))) +
  geom_bar() +
  scale_fill_manual(values = c("1" = "lightblue", "2" = "pink")) +
  labs(title = "Gender Distribution", x = "Gender", y = "Count") +
  scale_x_discrete(labels = c("1" = "Male", "2" = "Female"))

# Respondent Income distribution
ggplot(data, aes(x = RINCOME)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Respondent Income Distribution", x = "RIncome", y = "Frequency")

ggplot(data, aes(x = income)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black") +
  labs(title = "Respondent Income Distribution", x = "Income", y = "Frequency")


# Family Income distribution
ggplot(data, aes(x = INCOME)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Household Income Distribution", x = "Income", y = "Frequency")

ggplot(data, aes(x = house_income)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black") +
  labs(title = "Household Income Distribution", x = "House Income", y = "Frequency")


# Partner Income distribution
ggplot(data, aes(x = partner_income)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black") +
  labs(title = "Partner Income Distribution", x = "Partner Income", y = "Frequency")

# Health distribution
ggplot(data, aes(x = HEALTH)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Health Distribution", x = "Health", y = "Frequency")

# Happyness distribution
ggplot(data, aes(x = HAPPY)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
```

```
  labs(title = "Happyness Distribution", x = "Happyness", y = "Frequency")



# Econometric models (Regression models)

# Health with personal income
model_health_income <- lm(HEALTH ~ income, data)

summary(model_health_income)
modelsummary(model_health_income,
             gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
             stars = TRUE,
             notes = "Standard errors in parentheses.",
             title = "health ~ income")

ggplot(data, aes(x = income, y = HEALTH)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of HEALTH ~ income",
       x = "income",
       y = "HEALTH")

# Happy with Health and income
model_happy_health_income <- lm(HAPPY ~ HEALTH + income + AGE + EDUC, data)

summary(model_happy_health_income)
modelsummary(model_happy_health_income,
             gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
             stars = TRUE,
             notes = "Standard errors in parentheses.",
             title = "happy ~ health + income + AGE + EDUC")

ggplot(data, aes(x = HEALTH + income + AGE + EDUC , y = HAPPY)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of HAPPY ~ HEALTH + income",
       x = "HEALTH + income + AGE + EDUC",
       y = "HAPPY")

ggplot(data, aes(x = HEALTH, y = HAPPY)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of HAPPY ~ RINCOME",
       x = "HEALTH",
       y = "HAPPY")

# Personal education with father/mother educatione
model_edu <- lm(EDUC ~ MAEDUC + PAEDUC + SPEDUC, data)

summary(model_edu)
modelsummary(model_edu,
             gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
             stars = TRUE,
```

```
             notes = "Standard errors in parentheses.",
             title = "education ~ fother/mother/spouse education")

ggplot(data, aes(x = MAEDUC + PAEDUC + SPEDUC, y = EDUC)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of RINCOME ~ INCOME",
       x = "MAEDUC + PAEDUC + SPEDUC",
       y = "EDUC")

# Personal income and partner income
model_income_partner <- lm(partner_income ~ income, data)

summary(model_income_partner)
modelsummary(model_income_partner,
             gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
             stars = TRUE,
             notes = "Standard errors in parentheses.",
             title = "partner_income ~ income")

ggplot(data, aes(x = income, y = partner_income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of partner_income ~ income",
       x = "income",
       y = "partner_income")

model_income_partner_log <- lm(lnpartner_income ~ lnincome, data)

summary(model_income_partner_log)
modelsummary(model_income_partner_log,
             gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
             stars = TRUE,
             notes = "Standard errors in parentheses.",
             title = "Log partner_income ~ Log income")

ggplot(data, aes(x = lnincome, y = lnpartner_income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of lnpartner_income ~ income",
       x = "log income",
       y = "log partner_income")

# income and AGE, EDUC, BABIES male and female data
models <- list()
models[["male_level"]] <- lm(income ~ AGE + EDUC + BABIES, mdata)
models[["male_log"]]   <- lm(lnincome ~ AGE + EDUC + BABIES, mdata)
models[["female_level"]] <- lm(income ~ AGE + EDUC + BABIES, fdata)
models[["female_log"]]   <- lm(lnincome ~ AGE + EDUC + BABIES, fdata)
models[["both"]]   <- lm(lnincome ~ AGE + EDUC + BABIES, data)
models[["both_log"]]   <- lm(lnincome ~ AGE + EDUC + BABIES, data)

modelsummary(models,
             gof_omit = "Log.Lik.|AIC|BIC|F|RMSE",
```

```
             stars = TRUE,
             notes = "Standard errors in parentheses.",
             title = "Income table")

ggplot(data, aes(x = AGE, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of income ~ AGE",
       x = "AGE",
       y = "income")

ggplot(mdata, aes(x = AGE, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Male Regression Plot of income ~ AGE",
       x = "AGE",
       y = "income")

ggplot(fdata, aes(x = AGE, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Female Regression Plot of income ~ AGE",
       x = "AGE",
       y = "income")

ggplot(data, aes(x = EDUC, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of income ~ EDUC",
       x = "EDUC",
       y = "income")

ggplot(mdata, aes(x = EDUC, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Male Regression Plot of income ~ EDUC",
       x = "EDUC",
       y = "income")

ggplot(fdata, aes(x = EDUC, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Female Regression Plot of income ~ EDUC",
       x = "EDUC",
       y = "income")

ggplot(data, aes(x = BABIES, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Regression Plot of income ~ BABIES",
       x = "BABIES",
       y = "income")

ggplot(mdata, aes(x = BABIES, y = income)) +
```

```
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Male Regression Plot of income ~ BABIES",
       x = "BAABIES",
       y = "income")

ggplot(fdata, aes(x = BABIES, y = income)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Female Regression Plot of income ~ BABIES",
       x = "BABIES",
       y = "income")
```