EVS 2017 Data Review

Chris Rice

2024-03-09

According to the [website]: *The European Values Study is a large-scale, cross-national and longitudinal survey research program on how Europeans think about family, work, religion, politics, and society. Repeated every nine years in an increasing number of countries, the survey provides insights into the ideas, beliefs, preferences, attitudes, values, and opinions of citizens all over Europe.*

Beginning steps were to create an account with the aforementioned website in order to extract either the SPSS or STATA file. In this case, we downloaded the STATA file in order to create an EVS 2017 analysis. The original data set had hundreds of columns to choose from however, variables of interest were as follows:

1. Respondent's age (continuous)
2. Respondent's country (categorical)
3. Respondent's education (categorical)
4. Respondent's sex (categorical)
5. Respondent's opinion on if jobs are scarce should national citizens have priority over immigrants (categorical)
6. Respondent's opinion on if a child suffers when the mother is working (categorical)

Table 1 - Categorical Variable Descriptive Statistics

```
knitr::opts_chunk$set(warning = FALSE)

load("C:/Users/ricecakes/Desktop/Git1/HW-2/Rcode/df1c.RData")

calculate_frequencies <- function(data, var_name) {
  data %>%
    count(!!sym(var_name), name = "Frequency") %>%
    mutate(Percentage = Frequency / sum(Frequency) * 100,
           Category = as.character(!!sym(var_name)),
           Variable = var_name) %>%
    select(Variable, Category, Frequency, Percentage)
}

load("C:/Users/ricecakes/Desktop/Git1/HW-2/Rcode/df1c.RData")
# List of categorical variables
categorical_vars <- c("country", "education", "child_suffers", "jobs_are_scarce")

# Calculate frequencies and percentages for each variable and bind rows together
load("C:/Users/ricecakes/Desktop/Git1/HW-2/Rcode/df1c.RData")
categorical_summary <- map_dfr(categorical_vars, ~calculate_frequencies(df1c, .x))
```

```r
# Display the table
load("C:/Users/ricecakes/Desktop/Git1/HW-2/Rcode/df1c.RData")
knitr::kable(categorical_summary, booktabs = TRUE, caption = "Descriptive Summary of Categorical Variabl
  kable_styling(latex_options = c("striped", "scale_down")) %>%
  column_spec(1, width = "4cm") %>%
  column_spec(2, width = "4cm")
```

Table 1: Descriptive Summary of Categorical Variables

Variable	Category	Frequency	Percentage
country	Albania	1435	2.4142804
country	Armenia	1500	2.5236381
country	Austria	1644	2.7659073
country	Azerbaijan	1800	3.0283657
country	Belarus	1548	2.6043945
country	Bosnia and Herzegovina	1724	2.9005014
country	Bulgaria	1558	2.6212187
country	Croatia	1487	2.5017665
country	Czechia	1811	3.0468724
country	Denmark	3362	5.6563141
country	Estonia	1304	2.1938827
country	Finland	1199	2.0172280
country	France	1870	3.1461355
country	Georgia	2194	3.6912413
country	Germany	2170	3.6508631
country	Great Britain	1788	3.0081766
country	Hungary	1514	2.5471920
country	Iceland	1624	2.7322588
country	Italy	2277	3.8308826
country	Latvia	1335	2.2460379
country	Lithuania	1448	2.4361520
country	Montenegro	1003	1.6874727
country	Netherlands	2404	4.0445506
country	North Macedonia	1117	1.8792692
country	Norway	1122	1.8876813
country	Poland	1352	2.2746391
country	Portugal	1215	2.0441468
country	Romania	1613	2.7137521
country	Russia	1825	3.0704263
country	Serbia	1499	2.5219557
country	Spain	1209	2.0340523
country	Sweden	1194	2.0088159
country	Switzerland	3174	5.3400182
country	Ukraine	1612	2.7120697
country	NA	2507	4.2178404
education	Bachelor or equivalent	6508	10.9492244
education	Doctoral or equivalent	555	0.9337461
education	Less than primary	510	0.8580369
education	Lower secondary	8588	14.4486692

education	Master or equivalent	8397	14.1273260
education	Post-secondary non tertiary	2705	4.5509607
education	Primary	3028	5.0943841
education	Short-cycle tertiary	4553	7.6600828
education	Upper secondary	24121	40.5817827
education	dont know	81	0.1362765
education	no answer	308	0.5181870
education	other	84	0.1413237
child_suffers	agree	16484	27.7331000
child_suffers	agree strongly	5556	9.3475554
child_suffers	disagree	25122	42.2658905
child_suffers	disagree strongly	10918	18.3687203
child_suffers	dont know	1158	1.9482486
child_suffers	no answer	200	0.3364851
jobs_are_scarce	agree	16691	28.0813621
jobs_are_scarce	agree strongly	20308	34.1666947
jobs_are_scarce	disagree	8263	13.9018810
jobs_are_scarce	disagree strongly	4309	7.2495710
jobs_are_scarce	dont know	622	1.0464686
jobs_are_scarce	multiple answers Mail	1	0.0016824
jobs_are_scarce	neither agree nor disagree	9062	15.2461388
jobs_are_scarce	no answer	182	0.3062014

This table shows categorical variables with their respective variable, category, frequency and percentage. We can observe count totals within each category as well as percentages of total response by type. Technically, "jobs are scarce" and "child suffers" could have been converted from categorical (since they are ordinal) to continuous (which we will do before running the regression model). However, in order to have more meaningful and easy to interpret values for our readers, we decided best to keep them true to form (categorical).

Table 2 - Continuous Variable Descriptive Statistics

```
df1c <- df1c %>%
  mutate(age = ifelse(age >= 82, 82, age))


df1c <- df1c %>%
  mutate(age = as.numeric(age),  # Attempt to convert age to numeric
         age = ifelse(is.na(age), NA, age))  # Non-numeric become NA


df1c %>%
  summarise(
    Mean = mean(age, na.rm = TRUE),
    SD = sd(age, na.rm = TRUE),
    Min = min(age, na.rm = TRUE),
    `25% Quantile` = quantile(age, 0.25, na.rm = TRUE),
    Median = median(age, na.rm = TRUE),
```

```
    `75% Quantile` = quantile(age, 0.75, na.rm = TRUE),
    Max = max(age, na.rm = TRUE)
  ) %>%
  knitr::kable(caption = "Descriptive Statistics for Age")
```

Table 2: Descriptive Statistics for Age

Mean	SD	Min	25% Quantile	Median	75% Quantile	Max
49.78912	17.79272	18	35	50	64	82

Age is our only continuous variable. However, within the age column was "82 and older". This response was changed to equal 82. Thus, when we see 82 the respondent's age can be considered greater than or equal to 82. Hence, the maximum age of 82. Our results show a mean age of roughly 50, an s.d. of nearly 18, with a minimum age of 18.

Below we highlight the confirmation our age variable holds no NA values. We can say with confidence our descriptive statistics surrounding age are pure.

No NAs in "age" confirmed

```
unique(df1c$age)
```

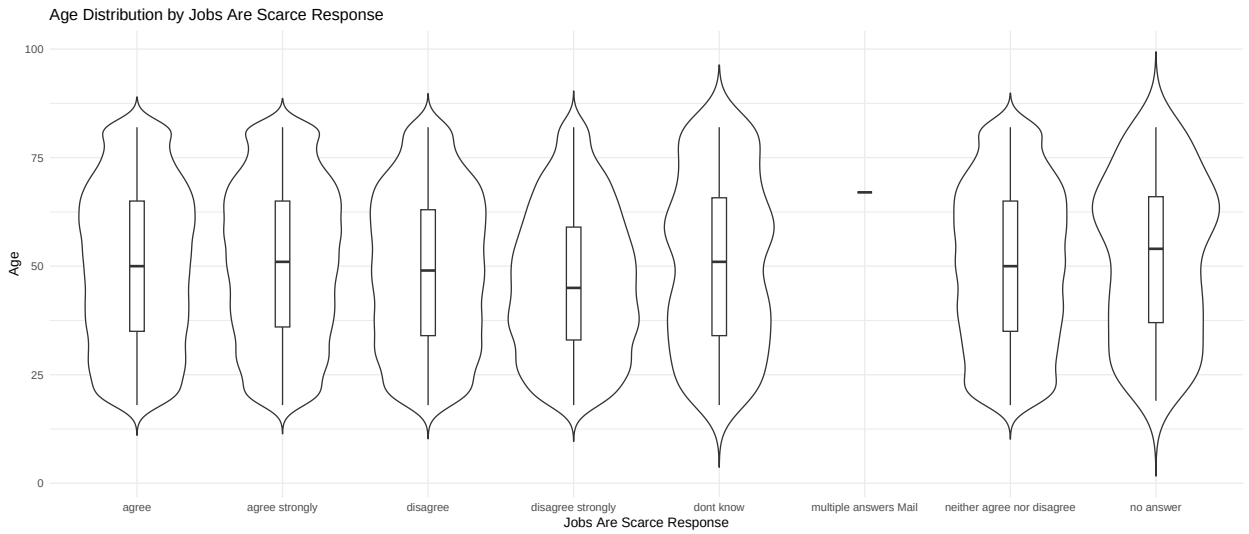
```
##  [1] 69 55 70 42 23 22 21 20 37 31 65 63 18 68 38 27 62 64 49 28 59 26 34 58 78
## [26] 54 46 19 39 60 36 71 50 57 25 82 47 77 44 24 53 30 29 41 56 66 43 40 51 72
## [51] 32 75 33 35 45 48 73 76 67 74 61 80 81 52 79
```

Age Impact on Dependent Variables

Plot 1 Age/Jobs are Scarce

```
library(ggplot2)
df1c <- df1c %>%
  mutate(jobs_are_scarce = as.factor(jobs_are_scarce))


ggplot(df1c, aes(x = jobs_are_scarce, y = age)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white") +  # narr bxplt for median/quartiles
  labs(title = "Age Distribution by Jobs Are Scarce Response",
       x = "Jobs Are Scarce Response",
       y = "Age") +
  theme_minimal()
```



We can see median and quartiles for each response category surrounding jobs_are_scarce. The specific question was: "When jobs are scarce, employers should give priority to [NATIONALITY] people over immigrants?" It appears from the box plots we can infer younger respondents tended to disagree strongly while observing median age rise when agreeing strongly.

Plot 2 Age/Child Suffers

```
library(ggplot2)

# Ensure 'child_suffers' is treated as a factor
df1c <- df1c %>%
  mutate(child_suffers = as.factor(child_suffers))

# Plotting the relationship between 'child_suffers' responses and 'age'
ggplot(df1c, aes(x = child_suffers, y = age)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white") +  # Add narr bxplt for med/qrtiles
  labs(title = "Age Distribution by Child Suffers Response",
       x = "Child Suffers Response",
       y = "Age") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  #
```

We can see median and quartiles for each response category surrounding child_suffers. The specific question was: "When a mother works for pay, the children suffer?" It appears from the box plots we can infer younger respondents tended to disagree strongly while observing median age rise when agreeing strongly.

Regression Models

```
library(dplyr)
library(stargazer)

load("C:/Users/ricecakes/Desktop/Git1/HW-2/Rcode/df2c.RData")

df2c <- df2c %>%
  mutate(age_squared = age^2,
         respondent_sex = as.factor(respondent_sex),
         education = as.factor(education))

model1 <- lm(jobs_are_scarce_numeric ~ age + age_squared + respondent_sex + education, data = df2c)
model2 <- lm(child_suffers_numeric ~ age + age_squared + respondent_sex + education, data = df2c)

stargazer(model1, model2, type = "text",
          title = "Regression Models Results",
          align = TRUE,
          header = FALSE,
          model.numbers = FALSE,
          digits = 3,  # Adjust number of decimal places, if desired
          p.auto = TRUE,  # Ensure automatic p-value significance levels are used
          signif.codes = TRUE)  # Ensure significance codes are shown
```

```
##
## Regression Models Results
## ================================================================================
##                                                  Dependent variable:
##                                          ---------------------------------------
##                                          jobs_are_scarce_numeric child_suffers_numeric
```

```
## ------------------------------------------------------------------------------
## age                                    0.001                0.001
##                                       (0.002)              (0.001)
##
## age_squared                          -0.00004**           -0.0001***
##                                       (0.00002)            (0.00001)
##
## respondent_sexfemale                  -0.102               0.074
##                                       (0.536)              (0.418)
##
## respondent_sexmale                    -0.059               -0.008
##                                       (0.536)              (0.418)
##
## respondent_sexno answer               0.027                0.756
##                                       (0.611)              (0.476)
##
## educationDoctoral or equivalent       0.361***             0.147***
##                                       (0.058)              (0.045)
##
## educationdont know                    -0.572***            -0.497***
##                                       (0.147)              (0.114)
##
## educationLess than primary            -0.723***            -0.716***
##                                       (0.061)              (0.047)
##
## educationLower secondary              -0.570***            -0.468***
##                                       (0.022)              (0.017)
##
## educationMaster or equivalent         -0.162***            -0.058***
##                                       (0.022)              (0.017)
##
## educationno answer                    -0.557***            -0.668***
##                                       (0.078)              (0.060)
##
## educationother                        -0.054               -0.139
##                                       (0.144)              (0.112)
##
## educationPost-secondary non tertiary  -0.462***            -0.236***
##                                       (0.030)              (0.023)
##
## educationPrimary                      -0.496***            -0.474***
##                                       (0.030)              (0.023)
##
## educationShort-cycle tertiary         -0.292***            -0.153***
##                                       (0.025)              (0.020)
##
## educationUpper secondary              -0.595***            -0.346***
##                                       (0.018)              (0.014)
##
## Constant                              2.842***             2.965***
##                                       (0.537)              (0.419)
##
## ------------------------------------------------------------------------------
## Observations                          59,438               59,438
```

```
## R2                                          0.034                 0.041
## Adjusted R2                                 0.033                 0.041
## Residual Std. Error (df = 59421)            1.312                 1.023
## F Statistic (df = 16; 59421)              128.964***            158.107***
## ============================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
##
## Regression Models Results
## ====
## TRUE
## ----
```

Here we took our categorical (and ordinal) dependent variables and made them continuous (numeric) so we could produce regression model results. These models show some statistical significance in predicting the dependent variables, but the low R^2 values indicate that they explain only a small portion of the variance. This suggests that while some predictors may have a statistically significant relationship with the outcome, they don't capture most of the complexity of the dependent variables. Further exploration of additional variables, interaction effects, or different modeling techniques might be needed.