Can Test Scores Predict Future Contribution of an Individual in Society: An Empirical Analysis in R

Soham Seal Jadavpur University Department of Economics Roll no. - 002300301079

2025-07-25

Contents

Introduction

This report presents an empirical investigation into the factors that predict "Responsible Social Contribution." The study leverages the rich, longitudinal data from the India Human Development Survey (IHDS) to explore the relative importance of traditional academic achievement versus a broader set of "alternative" factors, such as an individual's relationship with their school and their home environment.

Project Context & Motivation

The contemporary education system places overwhelming emphasis on quantifiable assessments - marks, grades, ranks, and scores. This project fundamentally challenges this paradigm by testing whether school marks performance is

actually a good predictor of responsible social contribution or alternative school-life factors are superior predictors of an individual's future contribution to society compared to traditional academic marks.

The research addresses a critical gap in educational policy: Are we measuring the right things? Does our marks-obsessed culture identify and nurture the individuals who ultimately become responsible contributors to society, or are we missing the actual predictors of meaningful social contribution? Do students that score more contribute more to the society? In other words, does the society need to produce more high scorers for its betterment?

Research Problem Statement

Primary Research Question: Do traditional school marks predict an individual's capacity to become a responsible contributor to society?

Secondary Research Question: Do alternative school-life factors do a better job at predicting an individual's capacity to become a responsible contributor to society?

Underlying Challenge: The current education system may be fundamentally flawed in its assessment methods, potentially overlooking individuals with high social contribution potential while over-emphasizing those who excel in traditional academic metrics.

Hypothesis Framework

Core Hypothesis

Alternative school-life factors (study fondness, curiosity, resilience, social integration, adaptability, etc.) are stronger predictors of responsible social contribution than quantifiable academic performance (marks, grades, ranks).

Specific Hypotheses to Test

- **H0**: School marks are significant predictors of responsible social contribution.
- **H1**: School marks are significant predictors of responsible social contribution.
- **H2**: Individual alternative school-life factors (like relation with school, child relation with school, school type, house environment) are stronger predictors than marks.
- **H3**: Combined alternative models show alternative factors over marks in predictive power.
- **H4**: The relationship holds across different demographic factors.

Objective

My objective in doing this project is to dig deeper into how the society's education system works and benefits itself and is a change in policy and/or mindset necessary to foster us towards the right direction. This project should also inspire others as it is not as robust as I would like it to be due to lack of data and time constraint.

Data and Methodology

R Code

```
required_packages <- c(
  "dplyr",
```

```
  "tidyr",
  "readr",
  "ggplot2",
  "psych",
  "skimr",
  "ggcorrplot",
  "here",
  "tools",
  "rmarkdown",
  "viridis",
  "stargazer",
  "ggthemes",
  "pandoc"
)

# Function to check, install, and load packages
install_and_load_packages <- function(packages) {
  for (package_name in packages) {
    if (!require(package_name, character.only = TRUE)) {
      message(paste("Installing package:", package_name))
      install.packages(package_name, dependencies = TRUE)
      # Try loading again after installation
      if (!require(package_name, character.only = TRUE)) {
        stop(paste("Failed to install and load package:", package_name, ". Please install manually."))
      }
    }
  }
}

# Set CRAN mirror
options(repos = c(CRAN = "https://cran.r-project.org"))

# Call the function to install and load all required packages
install_and_load_packages(required_packages)

# Attaching required packages
library(here)
library(rmarkdown)

# sourcing
setwd(here("R"))


# ============================================================================
# STEP 1: LOAD AND PREPARE DATA
# ============================================================================

# Define the path to the raw IHDS data
raw_data_dir <- here("data", "ihds-data", "data")
data_files <- list.files(path = raw_data_dir, full.names = TRUE)

data <- list()

for (file_path in data_files) {
```

```
  data_name <- tools::file_path_sans_ext(basename(file_path))
  message(paste("  - Importing:", basename(file_path)))
  df <- read.delim(file_path, sep = "\t", header = TRUE, stringsAsFactors = FALSE)
  data[[data_name]] <- df
}

# ================================================================================
# CREATE PANEL DATASET
# ================================================================================

# Create panel dataset using tracking information
panel_data <- data[["panel_individual"]]

# ================================================================================
# STEP 3: SAVE PANEL DATASET
# ================================================================================

saveRDS(panel_data, file = here("processed-data", "raw-panel-data.rds"))
print("Raw panel dataset saved as 'raw-panel-data.rds'")

panel_data = readRDS(here("processed-data", "raw-panel-data.rds"))

# Extracting necessary columns
necessary_cols = c(

    "HHBASE", # unique id for household
    "HHSPLITID", # unique id for split household
    "IDPERSON", # Person id, unique 12[IHDS2] or 11[IHDS1] byte string
    "PBASE", # multisurvey Person id for each household
    "XRO5", # age in 2005
    "RO5", # age in 2012
    "URBAN", # rural-urban Census 2001 for IHDS-I; 2011 for IHDS-II
    "XRO3", # Sex - revised

    "WKANYPLUS", # work participation (farm business w|s animal) 2012
    "XED4", # Educ: Attended school 2005

    "WKBUSINESS", # work business 2012
    "WKSALARY", # work wage salary 2012
    "WKFARM", # work farm 2012
    "WKANIMAL", # work animal 2012

    # **Dependent Variable: Responsible Social Contribution**

    # **Happiness & Well-being**

    "TO2Y", # smoke, alcohol, yesno
    "TO3", # Smoke cigarettes [IHDS2 only]
    "TO4", # Smoke bidis or hukkah [IHDS2 only]

    # **Health**

    "BAZ", # BMI for age zscore from zanthro(US) months>=24
```

```
"MB5", # High BP (lifestyle diseases)
"MB6", # Heart disease (lifestyle diseases)
"MB7", # Diabetes (lifestyle diseases)
"MB14", # Mental illness (lifestyle diseases)
"MB15", # STD or AIDS (diseases for lack of hygiene)
"MB4", # Tuberculosis (diseases for lack of hygiene)
"SM8", # Diarrhoea with blood (diseases for lack of hygiene)

"MB24", # Days hosp
"SM17", # Days hosp

"MB2Y", # yesno disease
"MB19", # yesno Treatment
"MB25", # Cost Dr/hosp
"MB27", # Cost Medicine
"MB29", # Med Insurance Rs [IHDS2 only]

"SM2Y", # yesno sick [IHDS2 only]
"SM12", # yesno Treatment
"SM18", # Dr/hosp Rs
"SM20", # Medicine Rs
"SM22", # Med Insurance Rs [IHDS2 only]

# **Social Responsibility**

"AN6", # Animal care: Frequency
"AN5Y", # Animal care work (reverse coded)

# **Productive Contribution**

"RO7", # Primary Activity Status [IHDS2 only]
"WKEARNPLUS", # Earnings est
"XWKEARN", # sum ag nonag salary farm animal business 2005
"INCOME", # Annual individual income 2012
"XINCOME", # HH Annual income Rs2012
"COTOTAL", # Annual consumption expenditure
"INCBENEFITS", # all govt benefits Rs
"POOR2", # Poverty 2012 Tendulkar cutoffs yesno (reverse coding required)
"WKDAYS", # work days /year (farm, business, wage|salary) >= 200

# **Independent Variables**

"ED6", # Educ: Completed Years, never,<1=0
"ED8", # Educ: secondary class
"XTA8B", # Test reading score
"XTA9B", # Test math score
"XTA10B", # Test Writing level [IHDS1 values]
"XCS24Y", # Scholarship
"ED13", # Educ: Degree class
# CONTROL:  (hh income 2005)
"XWKEARNPLUS", # Earnings est.: sum w|s farm business animal
"XED6", # Educ: Completed Years

# **Alternative Predictors**
```

```
    # 1# **Relation with school**

    "XTA5", # Test enjoy school
    "XTA6", # Test teacher not nice
    "XCS10", # School Hrs/week
    "XCS11", # Homework Hrs/week
    "XED7", # Educ: Ever repeated
    # CONSTRAINT:
    "XTA3", # test attended school

    "XCH2", # Child: School enrollment
    "XCH15", # Child: Average student
    "XCH16", # Child: School enjoyment
    "XCH18", # Child: Ever praised
    "XCH17", # Child: # Repeats
    "XCH19", # Child: Ever beaten
    "XCH9", # Child: Fair(unfair) Teacher
    "XCH10", # Child: Good Teacher
    "XCH11", # Child: Biased Teacher
    "ED6", # Educ: Completed Years, never,<1=0
    # CONSTRAINT:
    "XCH1Y", # Child: EdHe questions

    # 2# **School type**

    "XCS4", # School type
    "XCS25", # School fees
    "XCS12", # Pvt Tuition Hrs/week

    # 3# **House environment**

    "HHEDUCF", # Highest female adult educ [max=15]
    "HHEDUCM", # Highest male adult educ [max=15]
    "HHLITERATE" # Any adult (or head) in hh literate
  )

# ============================================================================
# CONSTRUCT COMPOSITE MEASURES
# ============================================================================

# 1.1 DEPENDENT VARIABLE(y): Responsible Social Contribution
construct_dependent_variable <- function(data) {
  data %>%
    mutate(
      # Happiness & Well-being (z scale)
      happiness_wellbeing = (
        as.vector(scale(1 - TO2Y)) +  # smoke, alcohol, yesno (reverse coded)
        as.vector(scale(5 - TO3)) +  # Smoke cigarettes (reverse coded)
        as.vector(scale(5 - TO4)) # Smoke bidis or hukkah (reverse coded)
      ) / 3,

      # Health (z scale)
```

```
baz_indicator = coalesce(BAZ, 0),  # BMI normalized

# --- Lifestyle Diseases (Z-Scaled) ---
lifestyle_diseases = coalesce(as.vector(scale({
  # 1. Convert relevant columns to binary (0/1) using pmin(.x, 1). NA remains NA.
  binary_lifestyle_diseases <-
    select(., c("MB5", "MB6", "MB7", "MB14")) %>%
    mutate(across(everything(), ~ pmin(.x, 1, na.rm = FALSE)))

  # 2. Count the number of diseases by summing binaries (na.rm=TRUE handles NAs in sum)
  num_lifestyle_diseases <-
    rowSums(binary_lifestyle_diseases, na.rm = TRUE)

  all_original_lifestyle_na <-
    rowSums(is.na(
      select(., c("MB5", "MB6", "MB7", "MB14")))) == length(c("MB5", "MB6", "MB7", "MB14"))
  num_lifestyle_diseases[all_original_lifestyle_na] <- 0

  # 4. Reverse code the count (Max possible lifestyle diseases = 4)
  length(c("MB5", "MB6", "MB7", "MB14")) - num_lifestyle_diseases
})), 0),

# --- Hygiene-related Diseases (Z-Scaled) ---
hygiene_diseases = coalesce(as.vector(scale({
  # 1. Convert relevant columns to binary using pmin(.x, 1). NA remains NA.
  binary_hygiene_diseases <- select(., c("MB15", "MB4", "SM8")) %>%
    mutate(across(everything(), ~ pmin(.x, 1, na.rm = FALSE)))

  # 2. Count the number of diseases
  num_hygiene_diseases <- rowSums(binary_hygiene_diseases, na.rm = TRUE)

  # 3. Refine NA handling for the count
  all_original_hygiene_na <-
    rowSums(is.na(
      select(., c("MB15", "MB4", "SM8")))) == length(c("MB15", "MB4", "SM8"))
  num_hygiene_diseases[all_original_hygiene_na] <- NA_real_

  # 4. Reverse code the count (Max possible hygiene diseases = 3)
  length(c("MB15", "MB4", "SM8")) - num_hygiene_diseases
})), 0),

# healthy days
healthy_days = coalesce(as.vector(scale({
  # 1. Sum the days in hospital. If either MB24 or SM17 is NA, their sum will be NA.
  days_in_hospital <- MB24 + SM17

  # 2. Get the maximum observed total days in hospital across the entire column.
  # This serves as our 'Max_Old_Value' for reverse coding.
  max_observed_days = max(days_in_hospital, na.rm = TRUE)
  min_observed_days = min(days_in_hospital, na.rm = TRUE)

  # 3. Apply reverse coding: (Max observed days - Current total days)
  # Less days in hospital should mean a higher social contribution score.
  (max_observed_days + min_observed_days) - days_in_hospital
```

```
  })), 0),

    # --- New: Healthcare Utilization (Z-Scaled) ---
  healthcare_utilization = ({
    mb_utilization_component <- case_when(
      MB2Y == 1 ~ (coalesce(as.vector(scale(MB19)), 0) +
                   coalesce(as.vector(scale(MB25)), 0) +
                   coalesce(as.vector(scale(MB27)), 0))/3,
      MB2Y == 0 ~ 0,
      TRUE ~ 0
    )

    sm_utilization_component <- case_when(
      SM2Y == 1 ~ (coalesce(as.vector(scale(SM12)), 0) +
                   coalesce(as.vector(scale(SM18)), 0) +
                   coalesce(as.vector(scale(SM20)), 0))/3,
      SM2Y == 0 ~ 0,
      TRUE ~ 0
    )

    # Combine both components. Sum will be NA if either component is NA.
    (mb_utilization_component + sm_utilization_component)/2
  }),

health = (

  baz_indicator * 0.2 +
  lifestyle_diseases * 0.1 +
  hygiene_diseases * 0.1 +
  healthy_days * 0.2 +
  healthcare_utilization * 0.2

),

# Animal Care (z scale)
animal_care = {

  an6_scaled <- as.vector(scale(.data$AN6))
  an5y_scaled <- as.vector(scale(.data$AN5Y))

  base_weight <- 0.15
  additional_weight <- 0.35

  effective_weight <- case_when(
    .data$WKANIMAL <= 2 ~ base_weight,
    .data$WKANIMAL > 2 ~ additional_weight,
    TRUE ~ 0
  )

  weighted_sum <- (an6_scaled * effective_weight) + (an5y_scaled * effective_weight)

  coalesce(weighted_sum, 0)
},
```

```
      # Productive Contribution (z scale)

      productive_contribution = ({
        productive_contr <- case_when(
            (RO7 <= 10 & WKDAYS >= 100 & XINCOME != 0) ~ (
              coalesce(as.vector(scale(WKEARNPLUS)), 0) + # Work Earnings est 2012
              coalesce(as.vector(scale(INCOME)), 0) + # Annual individual income 2012
              coalesce(as.vector(scale(XINCOME)), 0) + # HH Annual income Rs2012
              coalesce(as.vector(scale({COTOTAL * INCOME/XINCOME})), 0) # Annual individual consumption
              ) / 4,
            TRUE ~ 0
          )
        productive_contr
        }),

      # Overall Responsible Social Contribution (weighted average)
      responsible_social_contribution = (
        happiness_wellbeing * 0.20 +
        health * 0.30 +
        animal_care * 0.20 +
        productive_contribution * 0.30
      )
    )
}

# 4.2 INDEPENDENT VARIABLES(X_0): Traditional Predictors
construct_traditional_predictors <- function(data) {
  data %>%
    mutate(

      # Individual components (z scale)
      completed_years = coalesce(as.vector(scale(ED6)), 0),
      secondary_class = coalesce(as.vector(scale({
        secondary_cl = case_when(
            ED8 == "I" ~ 3,
            ED8 == "II" ~ 2,
            ED8 == "III" ~ 1,
            TRUE ~ 0
          )
        secondary_cl
        })), 0),
      reading_score = coalesce(as.vector(scale(XTA8B)), 0),
      math_score = coalesce(as.vector(scale(XTA9B)), 0),
      writing_level = coalesce(as.vector(scale(XTA10B)), 0),
      scholarship = coalesce(as.vector(scale(XCS24Y)), 0),
      degree_class = coalesce(as.vector(scale({
        degree_cl = case_when(
            ED13 == "I" ~ 3,
            ED13 == "II" ~ 2,
            ED13 == "III" ~ 1,
            TRUE ~ 0
          )
        degree_cl
        })), 0),
```

```
      control_predictor_trad =
        (coalesce(as.vector(scale(XWKEARNPLUS)), 0) + # hh income 2005
         coalesce(as.vector(scale(XED6)), 0) # completed years 2005
         ) / 2,

      # Academic Performance Composite (weighted average)
      academic_performance = (

        secondary_class * 0.2 +
        degree_class * 0.2 +
        scholarship * 0.2 +
        reading_score * 0.1 +
        writing_level * 0.1 +
        math_score * 0.1 +
        completed_years * 0.1

      )
    )
}

# 4.3 ALTERNATIVE PREDICTORS(X_1): Alternative School-life Factors
construct_alternative_predictors <- function(data) {
  data %>%
    mutate(
      # 1# Relation with school (z scale)

      enjoy_school = coalesce(as.vector(scale(XTA5)), 0), # Test enjoy school
      teacher_nice = coalesce(as.vector(scale(4 - XTA6)), 0), # Test teacher not nice (reverse coded)
      school_hrs = coalesce(as.vector(scale(XCS10)), 0), # School Hrs/week
      homework_hrs = coalesce(as.vector(scale(XCS11)), 0), # Homework Hrs/week

      relation_with_school = (

        enjoy_school * 0.35 + # Test enjoy school
        teacher_nice * 0.35 + # Test teacher not nice (reverse coded)
        school_hrs * 0.20 + # School Hrs/week
        homework_hrs * 0.10 # Homework Hrs/week
        ),

      # 2# Child's relation with school (z scale)

      ch_teacher = (
        coalesce(as.vector(scale(XCH9)), 0) + # Child: Fair(unfair) Teacher
        coalesce(as.vector(scale(5 - XCH10)), 0) + # Child: Good Teacher (reverse coded)
        coalesce(as.vector(scale(4 - XCH11)), 0) # Child: Biased Teacher (reverse coded)
      ) / 3,
      ch_average_student = coalesce(as.vector(scale(XCH15)), 0),  # Child: Average student
      ch_school_enjoyment = coalesce(as.vector(scale(XCH16)), 0), # Child: School enjoyment
      ch_resilience = (
        coalesce(as.vector(scale(ED6)), 0) + # Educ: Completed Years
        coalesce(as.vector(scale(XCH17)), 0) + # Child: # Repeats
        coalesce(as.vector(scale(XCH19)), 0) # Child: Ever beaten
      ) / 3,
      ch_ever_praised = coalesce(as.vector(scale(XCH18)), 0), # Child: Ever praised
```

```
      child_relation_with_school = (

        ch_teacher +
        ch_average_student +
        ch_school_enjoyment +
        ch_resilience +
        ch_ever_praised

      ) / 5,

      # 3# School type (z scale)

      school_type = coalesce(as.vector(scale(XCS4)), 0), # School type
      school_fees = coalesce(as.vector(scale(XCS25)), 0), # School fees

      school_type = (
        school_type + # School type
        school_fees # School fees
      ) / 2,

      # 4# House environment (z scale)

      highest_education_female = coalesce(as.vector(scale(HHEDUCF)), 0), # Highest female adult educ [ma
      highest_education_male = coalesce(as.vector(scale(HHEDUCM)), 0), # Highest male adult educ [max=15
      any_adult_literate = coalesce(as.vector(scale(HHLITERATE)), 0), # Any adult (or head) in hh litera

      house_environment = (

        highest_education_female + # Highest female adult educ [max=15]
        highest_education_male + # Highest male adult educ [max=15]
        any_adult_literate # Any adult (or head) in hh literate

      ) / 3,

      alternative_factors = (

        relation_with_school +
        child_relation_with_school +
        school_type +
        house_environment

      ) / 4

    )
}

extracted_panel_data <- panel_data %>%
  select(
    all_of(necessary_cols)
  )

# Construct all composite measures
extracted_panel_data <- extracted_panel_data %>%
```

```
  construct_dependent_variable() %>%
  construct_traditional_predictors() %>%
  construct_alternative_predictors()

extracted_panel_data <- extracted_panel_data %>%
  # participation (farm business w|s animal) 2012
  filter(WKANYPLUS != 0) %>%
  # Educ: Attended school 2005
  filter(XED4 == 1)

# save
saveRDS(extracted_panel_data, file = here("processed-data", "extracted-panel-data.rds"))
print("Extracted and cleaned panel dataset saved as 'extracted-panel-data.rds'")

extracted_panel_data = readRDS(here("processed-data", "extracted-panel-data.rds"))

# ============================================================================
# PREPARE DATA FOR ANALYSIS
# ============================================================================

# Create analysis-ready dataset with standardized variables
analysis_dataset <- extracted_panel_data %>%
  # STEP 1: Create all new factor variables using mutate()
  mutate(
    gender_f = factor(XRO3, levels = c(1, 2), labels = c("Male", "Female")),
    urban_rural_f = factor(URBAN, levels = c(1, 0), labels = c("Urban", "Rural")),
    business_f = factor(WKBUSINESS, levels = c(0, 1, 2, 3, 4), labels = c("No Business", "No Business",
    salary_f = factor(WKSALARY, levels = c(0, 1, 2, 3, 4), labels = c("No Salary", "No Salary", "No Sala
    farm_f = factor(WKFARM, levels = c(0, 1, 2, 3, 4), labels = c("No farm", "No farm", "No farm", "farm
    animal_f = factor(WKANIMAL, levels = c(0, 1, 2, 3, 4), labels = c("No animal", "No animal", "No anim
    work_type_f = case_when(
      WKSALARY >= 3 ~ "Salary",
      WKBUSINESS >= 3 ~ "Business",
      WKFARM >= 3 ~ "Farm",
      WKANIMAL >= 3 ~ "Animal"
    ) %>% factor(levels = c("Animal", "Farm", "Business", "Salary"))
  ) %>%
  # STEP 2: Select the final set of desired columns, including the new factors
  select(
    responsible_social_contribution,

    academic_performance,
      completed_years,
      secondary_class,
      reading_score,
      math_score,
      writing_level,
      scholarship,
      degree_class,

    alternative_factors,
      relation_with_school,
        enjoy_school, # Test enjoy school
        teacher_nice, # Test teacher not nice (reverse coded)
```

```
        school_hrs, # School Hrs/week
        homework_hrs, # Homework Hrs/week
      child_relation_with_school,
        ch_teacher,
        ch_average_student,
        ch_school_enjoyment,
        ch_resilience,
        ch_ever_praised,
      school_type,
      house_environment,

    control_predictor_trad,

    # Now include the new factor variables we just created in mutate()
    gender_f,
    urban_rural_f,
    work_type_f
  ) %>%
  # STEP 3: Filter out rows where ANY of the selected columns have an NA
  filter(!if_any(everything(), is.na))

# Save analysis-ready dataset
saveRDS(analysis_dataset, file = here("processed-data", "structured-dataset.rds"))
print("Analysis-ready dataset saved as 'structured_dataset.rds'")

print("=== DATA PREPARATION COMPLETE ===")
print("Ready for econometric analysis!")

# Function to ensure output directories exist
ensure_output_dirs <- function() {
  output_dirs <- c(
    here("output", "figures"),
    here("output", "tables"),
    here("output", "reports")
  )

  for (dir in output_dirs) {
    if (!dir.exists(dir)) {
      dir.create(dir, recursive = TRUE, showWarnings = FALSE)
      message(paste("Created directory:", dir))
    }
  }
}

# Helper function to save plots with standardized naming and formats
save_plot <- function(obj, name, width = 10, height = 8, dpi = 300) {
  # Ensure output directory exists
  ensure_output_dirs()

  # Clean the name (remove special characters, replace spaces with underscores)
  clean_name <- gsub("[^A-Za-z0-9_-]", "_", name)
  clean_name <- gsub("_{2,}", "_", clean_name)  # Replace multiple underscores with single
  clean_name <- gsub("^_|_$", "", clean_name)   # Remove leading/trailing underscores
```

```
  # Define file paths
  png_path <- here("output", "figures", paste0(clean_name, ".png"))

  # Save as PNG
  ggsave(filename = png_path, plot = obj, width = width, height = height,
         dpi = dpi, units = "in", device = "png")

  message(paste("Plot saved as:"))
  message(paste("  PNG:", png_path))

  return(list(png = png_path))
}

# Helper function to save tables with standardized naming and formats
save_table <- function(df, name, include_rownames = FALSE) {
  # Ensure output directory exists
  ensure_output_dirs()

  # Clean the name (remove special characters, replace spaces with underscores)
  clean_name <- gsub("[^A-Za-z0-9_-]", "_", name)
  clean_name <- gsub("_{2,}", "_", clean_name)  # Replace multiple underscores with single
  clean_name <- gsub("^_|_$", "", clean_name)   # Remove leading/trailing underscores

  # Define file paths
  csv_path <- here("output", "tables", paste0(clean_name, ".csv"))

  # Save as CSV
  write_csv(df, csv_path)

  message(paste("Table saved as:"))
  message(paste("  CSV:", csv_path))

  return(list(csv = csv_path))
}

# function to trim outliers
trim_outliers_iqr <- function(data, variable_name) {
  Q1 <- quantile(data[[variable_name]], 0.25, na.rm = TRUE)
  Q3 <- quantile(data[[variable_name]], 0.75, na.rm = TRUE)
  IQR_val <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR_val
  upper_bound <- Q3 + 1.5 * IQR_val

  data %>%
    filter(!!sym(variable_name) >= lower_bound, !!sym(variable_name) <= upper_bound)
}

# Q-Q Plot Functions
qq_plot <- function(data, column_name){
  ggplot(data = data,
         aes_string(sample = column_name)) +
    stat_qq() +
    stat_qq_line() +
    labs(title = paste("Q-Q Plot:", tools::toTitleCase(gsub("_", " ", column_name))),
```

```
         x = "Theoretical Quantiles", y = "Sample Quantiles") +
    theme_minimal()
}

# Load analysis dataset
structured_dataset = readRDS(here("processed-data", "structured-dataset.rds"))

# Create an sample analysis table for the report
sample_analysis_dataset = (
  structured_dataset %>%
    select(
      responsible_social_contribution,
      academic_performance,
      relation_with_school,
      child_relation_with_school,
      school_type,
      house_environment,
      alternative_factors
    )
)

save_table(head(sample_analysis_dataset), "sample_analysis_table")

# Q-Q Plots
qq_plot(sample_analysis_dataset, "responsible_social_contribution")
qq_plot(sample_analysis_dataset, "academic_performance")
qq_plot(sample_analysis_dataset, "relation_with_school")
qq_plot(sample_analysis_dataset, "child_relation_with_school")
qq_plot(sample_analysis_dataset, "school_type")
qq_plot(sample_analysis_dataset, "house_environment")
qq_plot(sample_analysis_dataset, "alternative_factors")

# Taking only the interquartile range of the dataset as Q-Q plots show non normality in tails
analysis_dataset <- structured_dataset %>%
   trim_outliers_iqr("responsible_social_contribution")  %>%
  trim_outliers_iqr("alternative_factors")

# Q-Q Plots
qq_plot(analysis_dataset, "responsible_social_contribution")
qq_plot(analysis_dataset, "academic_performance")
qq_plot(analysis_dataset, "relation_with_school")
qq_plot(analysis_dataset, "child_relation_with_school")
qq_plot(analysis_dataset, "school_type")
qq_plot(analysis_dataset, "house_environment")
qq_plot(analysis_dataset, "alternative_factors")

# Attach dataset to make variables available directly
attach(analysis_dataset)

# ============================================================================
# DESCRIPTIVE STATISTICS
# ============================================================================

# Key composites and controls for analysis
```

```
key_variables <- c(
  "responsible_social_contribution", "academic_performance", "alternative_factors",
  "relation_with_school", "child_relation_with_school", "school_type",
  "house_environment", "control_predictor_trad"
)

# Descriptive statistics using psych::describe()
descriptive_stats <- analysis_dataset %>%
  select(all_of(key_variables)) %>%
  psych::describe() %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Variable") # Convert row names to a column

# Save descriptive statistics
save_table(descriptive_stats, "descriptive_statistics_psych")

# ============================================================================
# VISUALIZATION OF DESCRIPTIVE STATISTICS
# ============================================================================
# Objective: To provide a visual overview of the central tendency and variability of key variables.

# Calculate means and standard deviations for plotting
summary_data <- analysis_dataset %>%
  select(all_of(key_variables)) %>%
  summarise_all(list(mean = mean, sd = sd), na.rm = TRUE) %>%
  pivot_longer(everything(), names_to = "Statistic", values_to = "Value") %>%
  separate(Statistic, into = c("Variable", "Type"), sep = "_(?=[^_]*$)") %>%
  pivot_wider(names_from = Type, values_from = Value)

# Create the dot plot with error bars
descriptive_plot <- ggplot(summary_data, aes(x = mean, y = reorder(Variable, mean))) +
  geom_point(size = 4, color = "darkblue") +
  geom_errorbarh(aes(xmin = mean - sd, xmax = mean + sd), height = 0.2, color = "gray60") +
  labs(title = "Mean and Standard Deviation of Key Variables",
       x = "Mean (with +/- 1 SD error bars)",
       y = "Variable") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title = element_text(face = "bold"))

# Save the plot
save_plot(descriptive_plot, "descriptive_statistics_plot")

# ============================================================================
# CORRELATION MATRIX
# ============================================================================

# Select numeric predictors and outcome for correlation matrix
numeric_vars <- analysis_dataset %>%
  select(all_of(key_variables)) %>%
  select_if(is.numeric)

# Calculate correlation matrix
cor_matrix <- cor(numeric_vars, use = "complete.obs")
```

```
# Save correlation matrix as table
cor_df <- as.data.frame(cor_matrix)
cor_df$Variable <- rownames(cor_df)
cor_df <- cor_df %>% select(Variable, everything())
save_table(cor_df, "correlation_matrix")

# Create correlation heatmap
cor_plot <- ggcorrplot(cor_matrix,
                       hc.order = TRUE,
                       type = "lower",
                       lab = TRUE,
                       lab_size = 3,
                       title = "Correlation Matrix of Key Variables",
                       ggtheme = ggplot2::theme_minimal())

# Save correlation plot
save_plot(cor_plot, "correlation_heatmap")

# ============================================================================
# DISTRIBUTION PLOTS
# ============================================================================

# Get numeric variables for distribution plots
numeric_composites <- c("responsible_social_contribution", "academic_performance",
                        "alternative_factors", "relation_with_school",
                        "child_relation_with_school", "school_type",
                        "house_environment", "control_predictor_trad")

# Filter to only include variables that exist and are numeric
available_numeric <- numeric_composites[numeric_composites %in% names(analysis_dataset)]
available_numeric <- available_numeric[sapply(analysis_dataset[available_numeric], is.numeric)]

# 1. Histograms + density for each composite
for (var in available_numeric) {
  hist_plot <- ggplot(analysis_dataset, aes_string(x = var)) +
    geom_histogram(aes(y = ..density..), bins = 30, alpha = 0.7, fill = "skyblue", color = "black") +
    geom_density(color = "red", size = 1) +
    labs(title = paste("Distribution of", tools::toTitleCase(gsub("_", " ", var))),
         x = tools::toTitleCase(gsub("_", " ", var)),
         y = "Density") +
    theme_minimal()

  save_plot(hist_plot, paste0("histogram_", var))
}

# 2. Boxplots
for (var in available_numeric) {
  # Boxplot by gender
  box_gender <- ggplot(analysis_dataset, aes_string(x = "gender_f", y = var, fill = "gender_f")) +
    geom_boxplot(alpha = 0.7) +
    labs(title = paste("Distribution of", tools::toTitleCase(gsub("_", " ", var)), "by Gender"),
         x = "Gender",
         y = tools::toTitleCase(gsub("_", " ", var))) +
```

```
    theme_minimal() +
    theme(legend.position = "none")

  save_plot(box_gender, paste0("boxplot_", var, "_by_gender"))

  # Boxplot by urban/rural
  box_urban <- ggplot(analysis_dataset, aes_string(x = "urban_rural_f", y = var, fill = "urban_rural_f")
    geom_boxplot(alpha = 0.7) +
    labs(title = paste("Distribution of", tools::toTitleCase(gsub("_", " ", var)), "by Location"),
         x = "Location",
         y = tools::toTitleCase(gsub("_", " ", var))) +
    theme_minimal() +
    theme(legend.position = "none")

  save_plot(box_urban, paste0("boxplot_", var, "_by_location"))

  # Boxplot by work type
  box_urban <- ggplot(analysis_dataset, aes_string(x = "work_type_f", y = var, fill = "work_type_f")) +
    geom_boxplot(alpha = 0.7) +
    labs(title = paste("Distribution of", tools::toTitleCase(gsub("_", " ", var)), "by Work Type"),
         x = "Work Type",
         y = tools::toTitleCase(gsub("_", " ", var))) +
    theme_minimal() +
    theme(legend.position = "none")

  save_plot(box_urban, paste0("boxplot_", var, "_by_work_type"))
}

# 3. Scatter plots of each predictor vs. outcome with simple linear trend
outcome_var <- "responsible_social_contribution"
predictor_vars <- available_numeric[available_numeric != outcome_var]

for (var in predictor_vars) {
  scatter_plot <- ggplot(analysis_dataset, aes_string(x = var, y = outcome_var)) +
    geom_point(alpha = 0.6, color = "steelblue") +
    geom_smooth(method = "lm", se = TRUE, color = "red") +
    labs(title = paste("Relationship between", tools::toTitleCase(gsub("_", " ", var)),
                       "and", tools::toTitleCase(gsub("_", " ", outcome_var))),
         x = tools::toTitleCase(gsub("_", " ", var)),
         y = tools::toTitleCase(gsub("_", " ", outcome_var))) +
    theme_minimal()

  save_plot(scatter_plot, paste0("scatter_", var, "_vs_", outcome_var))
}

# ============================================================================
# EXTENDED BIVARIATE ANALYSIS: HISTOGRAMS BY CATEGORICAL FACTORS
# ============================================================================
# Objective: To visualize the distribution of key continuous variables across different categorical grou

# Define the continuous variables to plot
continuous_vars_for_hist <- c(
  "responsible_social_contribution",
  "academic_performance",
```

```
  "relation_with_school",
  "child_relation_with_school",
  "school_type",
  "house_environment"
)

# Define the categorical variables
categorical_vars_for_hist <- c(
  "gender_f",
  "urban_rural_f",
  "business_f",
  "work_type_f""
)

for (cont_var in continuous_vars_for_hist) {
  for (cat_var in categorical_vars_for_hist) {
    # Create histogram with density, filled by categorical variable
    hist_plot_faceted <- ggplot(analysis_dataset, aes_string(x = cont_var, fill = cat_var)) +
      geom_histogram(aes(y = ..density..), bins = 30, alpha = 0.7, position = "identity", color = "blac
      geom_density(alpha = 0.2) +
      labs(title = paste("Distribution of", tools::toTitleCase(gsub("_", " ", cont_var)), "by", tools::
           x = tools::toTitleCase(gsub("_", " ", cont_var)),
           y = "Density") +
      theme_minimal() +
      theme(legend.position = "bottom")

    save_plot(hist_plot_faceted, paste0("histogram_", cont_var, "_by_", cat_var))
  }
}

# ============================================================================
# GROUP SUMMARIES
# ============================================================================

# Group summaries by gender and urban_rural
group_summaries <- analysis_dataset %>%
  group_by(work_type_f) %>%
  summarise(
    across(all_of(available_numeric),
           list(mean = ~round(mean(.x, na.rm = TRUE), 3),
                sd = ~round(sd(.x, na.rm = TRUE), 3)),
           .names = "{.col}_{.fn}"),
    n = n(),
    .groups = "drop"
  )

# Save group summaries
save_table(group_summaries, "group_summaries_by_work_type")

# Additional summary by gender only
gender_summaries <- analysis_dataset %>%
  group_by(gender_f) %>%
  summarise(
    across(all_of(available_numeric),
```

```
            list(mean = ~round(mean(.x, na.rm = TRUE), 3),
                 sd = ~round(sd(.x, na.rm = TRUE), 3)),
            .names = "{.col}_{.fn}"),
    n = n(),
    .groups = "drop"
  )

save_table(gender_summaries, "group_summaries_by_gender")

# Additional summary by location only
location_summaries <- analysis_dataset %>%
  group_by(urban_rural_f) %>%
  summarise(
    across(all_of(available_numeric),
           list(mean = ~round(mean(.x, na.rm = TRUE), 3),
                sd = ~round(sd(.x, na.rm = TRUE), 3)),
           .names = "{.col}_{.fn}"),
    n = n(),
    .groups = "drop"
  )

save_table(location_summaries, "group_summaries_by_location")

# Detach dataset to avoid masking issues
detach(analysis_dataset)

print("=== EXPLORATORY ANALYSIS COMPLETE ===")
print("All tables and figures saved to /output/ directory")

# --- Utility Functions ---
# Source utility functions for saving tables and plots
if (!exists("save_table")) {
  original_wd <- getwd()
  setwd(here())

  ensure_output_dirs <- function() {
    output_dirs <- c(here("output", "figures"), here("output", "tables"), here("output", "reports"))
    for (dir in output_dirs) {
      if (!dir.exists(dir)) {
        dir.create(dir, recursive = TRUE, showWarnings = FALSE)
        message(paste("Created directory:", dir))
      }
    }
  }

  setwd(original_wd)
}

# --- Data Loading and Preprocessing ---
# Load the pre-processed and analysis-ready dataset
analysis_dataset <- readRDS(here("processed-data", "analysis-ready-dataset.rds"))

# Create dummy variables explicitly with fastDummies
# Assuming gender, urban_rural, work_type are categorical; adjust names if different
```

```
analysis_dataset <- analysis_dataset %>%
  dummy_cols(select_columns = c("gender_f", "urban_rural_f", "work_type_f"),
             remove_first_dummy = TRUE, # Avoid multicollinearity
             remove_selected_columns = TRUE) # Remove original categorical columns

# Print column names to verify dummy variables
print("=== Dataset Columns After Dummy Coding ===")
print(colnames(analysis_dataset))

print("=== STARTING REGRESSION ANALYSIS ===")
print(paste("Sample size:", nrow(analysis_dataset)))

# ================================================================================
# REGRESSION MODELING
# ================================================================================
# Objective: To model the relationship between the independent variables and the
# dependent variable, 'responsible_social_contribution'. We will build a series of models
# to test the relative importance of academic vs. alternative predictors.
# Note: Replace dummy variable names with actual names from your dataset after dummy_cols

attach(analysis_dataset)

# --- Model 1(a): Academic Performance Only ---

model1a <- lm(responsible_social_contribution ~ academic_performance + control_predictor_trad +
                gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)

# --- Model 1(b): Academic Performance Only --- (Across Demographic Factors)
model1b <- lm(responsible_social_contribution ~ academic_performance + control_predictor_trad, data = a

# --- Model 2(a): relation_with_school Only ---
model2a <- lm(responsible_social_contribution ~ relation_with_school +
                gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)

# --- Model 2(b): relation_with_school Only --- (Across Demographic Factors)
model2b <- lm(responsible_social_contribution ~ relation_with_school, data = analysis_dataset)

# --- Model 3(a): child_relation_with_school Only ---
model3a <- lm(responsible_social_contribution ~ child_relation_with_school +
                gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)

# --- Model 3(b): child_relation_with_school Only --- (Across Demographic Factors)
model3b <- lm(responsible_social_contribution ~ child_relation_with_school, data = analysis_dataset)

# --- Model 4(a): school_type Only ---
model4a <- lm(responsible_social_contribution ~ school_type +
                gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)

# --- Model 4(b): school_type Only --- (Across Demographic Factors)
model4b <- lm(responsible_social_contribution ~ school_type, data = analysis_dataset)

# --- Model 5(a): house_environment Only ---
model5a <- lm(responsible_social_contribution ~ house_environment +
                gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)
```

```
# --- Model 5(b): house_environment Only --- (Across Demographic Factors)
model5b <- lm(responsible_social_contribution ~ house_environment, data = analysis_dataset)

# --- Model 6(a): Combined Model ---
model6a <- lm(responsible_social_contribution ~ relation_with_school + child_relation_with_school + scho
               gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)

# --- Model 6(b): Combined Model --- (Across Demographic Factors)
model6b <- lm(responsible_social_contribution ~ relation_with_school + child_relation_with_school + scho

# --- Model 7(a): Combined Model ---
model7a <- lm(responsible_social_contribution ~ alternative_factors +
               gender_f_Female + urban_rural_f_Rural + work_type_f_Business, data = analysis_dataset)

# --- Model 7(b): Combined Model --- (Across Demographic Factors)
model7b <- lm(responsible_social_contribution ~ alternative_factors, data = analysis_dataset)

# --- Model 8(a): Combined Model ---
model8a <- lm(responsible_social_contribution ~ academic_performance +
               relation_with_school + child_relation_with_school + school_type + house_environment +
               control_predictor_trad + gender_f_Female + urban_rural_f_Rural + work_type_f_Business, d

# --- Model 8(b): Combined Model --- (Across Demographic Factors)
model8b <- lm(responsible_social_contribution ~ academic_performance +
               relation_with_school + child_relation_with_school + school_type + house_environment +
               control_predictor_trad, data = analysis_dataset)

# --- Model 9(a): Combined Model ---
model9a <- lm(responsible_social_contribution ~ academic_performance + alternative_factors +
               control_predictor_trad + gender_f_Female + urban_rural_f_Rural + work_type_f_Business, d

# --- Model 9(b): Combined Model --- (Across Demographic Factors)
model9b <- lm(responsible_social_contribution ~ academic_performance + alternative_factors +
               control_predictor_trad, data = analysis_dataset)

print("=== MODELS CREATED SUCCESSFULLY ===")

# --- Model 10: Final Model of Best Fit ---
# model10 <- lm(data = analysis_dataset) if time permits
# print("=== MODEL OF BEST FIT CREATED SUCCESSFULLY ===")

detach(analysis_dataset)

# ============================================================================
# MODEL DIAGNOSTICS
# ============================================================================

# Custom VIF function in base R
calculate_vif <- function(model) {
  predictors <- names(coef(model))[-1] # Exclude intercept
  if (length(predictors) <= 1) return(NULL) # VIF not meaningful for single predictor
  vif_values <- numeric(length(predictors))
  names(vif_values) <- predictors
```

```
  for (i in seq_along(predictors)) {
    formula <- as.formula(paste(predictors[i], "~", paste(predictors[-i], collapse = "+")))
    temp_model <- lm(formula, data = model$model)
    r_squared <- summary(temp_model)$r.squared
    vif_values[i] <- 1 / (1 - r_squared)
  }
  return(vif_values)
}

# Function to create and save Q-Q and residuals vs. fitted plots
create_diagnostic_plots <- function(model, model_name) {
  # Residuals vs. Fitted
  resid_plot <- ggplot(data = data.frame(fitted = fitted(model), residuals = residuals(model)),
                       aes(x = fitted, y = residuals)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    labs(title = paste("Residuals vs. Fitted:", model_name), x = "Fitted Values", y = "Residuals") +
    theme_minimal()

  # Q-Q Plot
  qq_plot <- ggplot(data = data.frame(residuals = residuals(model)),
                    aes(sample = residuals)) +
    stat_qq() +
    stat_qq_line() +
    labs(title = paste("Q-Q Plot:", model_name), x = "Theoretical Quantiles", y = "Sample Quantiles") +
    theme_minimal()

  # Save plots
  ggsave(here("output", "figures", paste0("resid_", model_name, ".png")), resid_plot, width = 5, height
  ggsave(here("output", "figures", paste0("qq_", model_name, ".png")), qq_plot, width = 5, height = 4)

  return(list(resid_plot = resid_plot, qq_plot = qq_plot))
}

perform_diagnostics <- function(model, model_name) {
  # Linearity: Ramsey RESET test
  reset_test <- resettest(model)
  linearity_check <- paste(case_when(
    reset_test$p.value >= 0.1 ~ "Linear",
    reset_test$p.value >= 0.05 ~ "Almost Linear",
    reset_test$p.value >= 0.01 ~ "Nearly Linear",
    TRUE ~ "Non-Linear"
  ), as.character(reset_test$p.value))

  # Independence: Durbin-Watson test
  dw_test <- dwtest(model)
  dw_result <- paste0("DW = ", round(dw_test$statistic, 2), ", p = ", round(dw_test$p.value, 3))

  # Homoscedasticity: Breusch-Pagan test
  bp_test <- bptest(model)
  bp_result <- paste0("BP = ", round(bp_test$statistic, 2), ", p = ", round(bp_test$p.value, 3))

  # Normality: Use Anderson-Darling for large datasets, Shapiro-Wilk for smaller ones
  n <- length(residuals(model))
```

```
if (n > 5000) {
  norm_test <- ad.test(residuals(model))
  norm_result <- paste0("AD = ", round(norm_test$statistic, 2), ", p = ", round(norm_test$p.value, 3)
  norm_note <- "Anderson-Darling used (n > 5000)"
} else if (n >= 3) {
  norm_test <- shapiro.test(residuals(model))
  norm_result <- paste0("W = ", round(norm_test$statistic, 2), ", p = ", round(norm_test$p.value, 3))
  norm_note <- "Shapiro-Wilk used"
} else {
  norm_result <- "NA"
  norm_note <- "Sample size < 3"
}

# Multicollinearity: VIF (using custom function)
vif_result <- "NA"
vif_values <- calculate_vif(model)
if (!is.null(vif_values)) {
  if (any(vif_values > 5)) {
    vif_result <- "High"
  } else {
    vif_result <- "Low"
  }
}

# Outliers: Cook's distance
cooks_d <- cooks.distance(model)
outliers_result <- paste0(sum(cooks_d > 1), " obs > 1")

# Exogeneity: Check correlation between residuals and predictors
exog_result <- "NA"
predictors <- names(coef(model))[-1]
if (length(predictors) > 0) {
  correlations <- sapply(predictors, function(pred) {
    cor(residuals(model), model$model[[pred]], use = "complete.obs")
  })
  exog_result <- if (any(abs(correlations) > 0.3)) "Potential Endogeneity" else "No Issues"
}

# Recommendations
recommendations <- c()
if (reset_test$p.value < 0.05) recommendations <- c(recommendations, "Consider non-linear terms or tra
if (dw_test$p.value < 0.05) recommendations <- c(recommendations, "Consider robust standard errors fo
if (bp_test$p.value < 0.05) recommendations <- c(recommendations, "Consider robust standard errors fo
if (grepl("p = [0.]+[0-4]", norm_result)) recommendations <- c(recommendations, "Residuals not normal
if (vif_result == "High") recommendations <- c(recommendations, "High multicollinearity detected. Che
if (sum(cooks_d > 1) > 0) recommendations <- c(recommendations, "Significant outliers detected. Inves
if (exog_result == "Potential Endogeneity") recommendations <- c(recommendations, "Potential endogene

if (length(recommendations) == 0) recommendations <- "None"
else recommendations <- paste(recommendations, collapse = " ")

# Create and save diagnostic plots
create_diagnostic_plots(model, model_name)
```

```
  # Create a summary data frame
  diagnostics_df <- data.frame(
    Model = model_name,
    Linearity = linearity_check,
    Independence = dw_result,
    Homoscedasticity = bp_result,
    Normality = norm_result,
    Normality_Note = norm_note,
    Multicollinearity = vif_result,
    Outliers = outliers_result,
    Exogeneity = exog_result,
    Recommendations = recommendations
  )

  return(diagnostics_df)
}

# List of all models
model_list <- list(
  model1a = model1a, model1b = model1b,
  model2a = model2a, model2b = model2b,
  model3a = model3a, model3b = model3b,
  model4a = model4a, model4b = model4b,
  model5a = model5a, model5b = model5b,
  model6a = model6a, model6b = model6b,
  model7a = model7a, model7b = model7b,
  model8a = model8a, model8b = model8b,
  model9a = model9a, model9b = model9b
)

# Perform diagnostics for all models
diagnostics_results <- lapply(names(model_list), function(name) {
  perform_diagnostics(model_list[[name]], name)
})

# Combine results into a single table
diagnostics_table <- do.call(rbind, diagnostics_results)

# Print and save the diagnostics table
print("=== MODEL DIAGNOSTICS ===")
print(diagnostics_table)
write_csv(diagnostics_table, here("output", "tables", "regression_diagnostics.csv"))

# ============================================================================
# MODEL SUMMARIES & SAVING
# ============================================================================

# Print summaries for all models
for (name in names(model_list)) {
  cat("\n")
  cat("========================================================================\n")
  cat("                        MODEL SUMMARY:", name, "\n")
  cat("========================================================================\n")
  print(summary(model_list[[name]]))
```

```
}

# Save the R model objects for later use
saveRDS(model_list, file = here("processed-data", "regression-models.rds"))

# Create a proper CSV-formatted table for all models
model_summaries <- bind_rows(
  lapply(names(model_list), function(name) {
    tidy(model_list[[name]], conf.int = TRUE) %>%
      mutate(Model = name)
  }),
  .id = "id"
) %>%
  select(Model, everything(), -id)

# Save as proper CSV
write_csv(model_summaries, here("output", "tables", "regression_models_summary.csv"))

print("=== REGRESSION ANALYSIS COMPLETE ===")
print("Regression models, diagnostics, and summary table have been saved.")

# Knitting the main rmarkdown file

rmd_file <- here("reports", "empirical-study-report.Rmd")
render(input = rmd_file)

#end
```

Data Source and Sample

This study utilizes data from the India Human Development Survey (IHDS), a comprehensive panel survey that provides a wealth of information on various aspects of life in India. Our analysis focuses on a subset of the IHDS panel data, specifically individuals for whom we have information on their educational experiences in the first wave of the survey and their social and economic outcomes in the second wave.

Link: https://www.icpsr.umich.edu/web/ICPSR/studies/37382/datadocumentation

After cleaning and pre-processing the data, our final sample consists of:

```
## [1] "34302 individuals."
```

Here the Final Analysis Table:

Table 1: Sample Analysis Table

responsible_social_contribution	academic_performance	relation_with_school	child_relation_with_school	school_type	house_environment	alternative_factors
0.7386076	1.2012892	0	0.1297400	0	1.0260412	0.2889453
0.7063235	1.5699592	0	0.1297400	0	1.0260412	0.2889453
0.5275149	0.4001753	0	0.0872240	0	1.0260412	0.2783163
0.7063235	1.2012892	0	0.1297400	0	1.0260412	0.2889453
0.4728467	-0.1291436	0	0.0163639	0	1.0260412	0.2606013
-0.4340736	0.3576592	0	0.0588799	0	0.6227395	0.1704048

Only first few rows are shown.

Variable Construction

A key feature of this study is the construction of composite variables to measure our key concepts of interest. These composite variables are created by combining several individual survey items into a single, more reliable measure.

Dependent Variable: Responsible Social Contribution

Our dependent variable, `responsible_social_contribution`, is a composite index designed to capture a broad range of positive social behaviors and outcomes. It is constructed as a weighted average of four sub-components: Happiness and Well-being, Health, Social Responsibility, and Productive Contribution.

Independent Variables

Our independent variables are divided into two main categories:

1. **Academic Performance:** A composite measure of traditional educational attainment, including years of schooling, test scores, and scholarships.
2. **Alternative Factors:** A composite measure capturing a broader range of school- and home-related factors, including relationship with school, school type, and home environment.

Analytical Approach

Our analysis proceeds in three main steps:

1. **Data Preparation:** We begin by importing the raw IHDS data, selecting the relevant variables, and constructing our composite measures. We then clean the data by removing observations with missing values.
2. **Exploratory Data Analysis:** We then conduct a thorough exploratory data analysis to understand the characteristics of our data. This includes calculating descriptive statistics, creating a correlation matrix, and visualizing the distributions of our key variables.
3. **Regression Analysis:** Finally, we use ordinary least squares (OLS) regression to model the relationship between our independent variables and our dependent variable. We estimate a series of four nested regression models to test our central hypothesis.

Exploratory Data Analysis

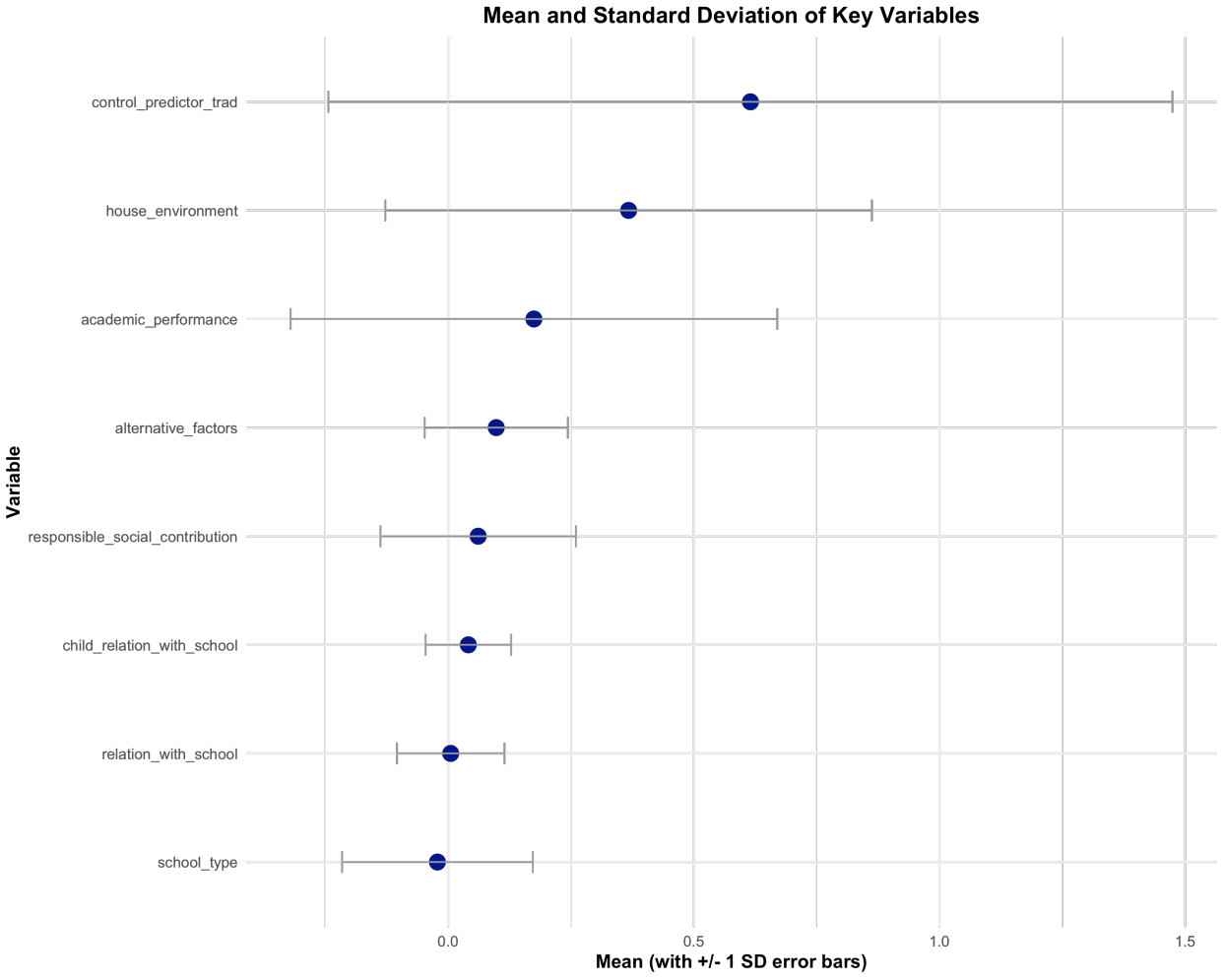
Descriptive Statistics

We begin our exploratory analysis by examining the descriptive statistics for our key variables. The following table provides a summary of the mean, standard deviation, and other key statistics for each of our composite variables.

Table 2: Descriptive Statistics of Key Variables

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
responsible_social_contribution	1	34302	0.06120001	0.19890653	0.07502388	0.07008833	0.1795705	-0.4739468	0.6003331	1.074280	-0.3491249	0.2049847	0.001074
academic_performance	2	34302	0.17461784	0.51934	-0.0866276	0.07922092	0.2178319	-0.8453338	2.328778	3.174111	1.955171	3.885947	0.0026737

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
alternative_factors	3	34302	0.09803	0.31457774	0.10300381	0.10119998	0.31538086	-0.3548577	0.5427059	0.8975636	-0.2140668	-0.08299311	0.0007871
relation_with_school	4	34302	0.00533	0.10937370	0.00000000	0.00000000	0.00000000	-2.2943380	1.3065030	3.6008411	-2.2847606	37.58529290	0.0005905
child_relation_with_school	5	34302	0.04119	0.30870644	0.04470790	0.04064744	0.30630343	-1.5131075	1.2014482	2.7145556	0.6616988	7.37108260	0.0004701
school_type	6	34302	-0.02169	0.19411920	0.00000000	-0.00107160	0.00000000	-0.7656853	2.0938026	2.8594879	2.0939501	21.37218301	0.0010481
house_environment	7	34302	0.36730	0.94950558	0.37483083	0.37758755	0.89611675	-1.6774275	1.2178625	2.8952950	-0.3249902	-0.07147670	0.0026730
control_predictor_trad	8	34302	0.61526	0.62588001	0.24526437	0.49768276	0.30152076	-5.0717115	18.4362259	23.5079376	7.7342460	80.52603307	0.0046370

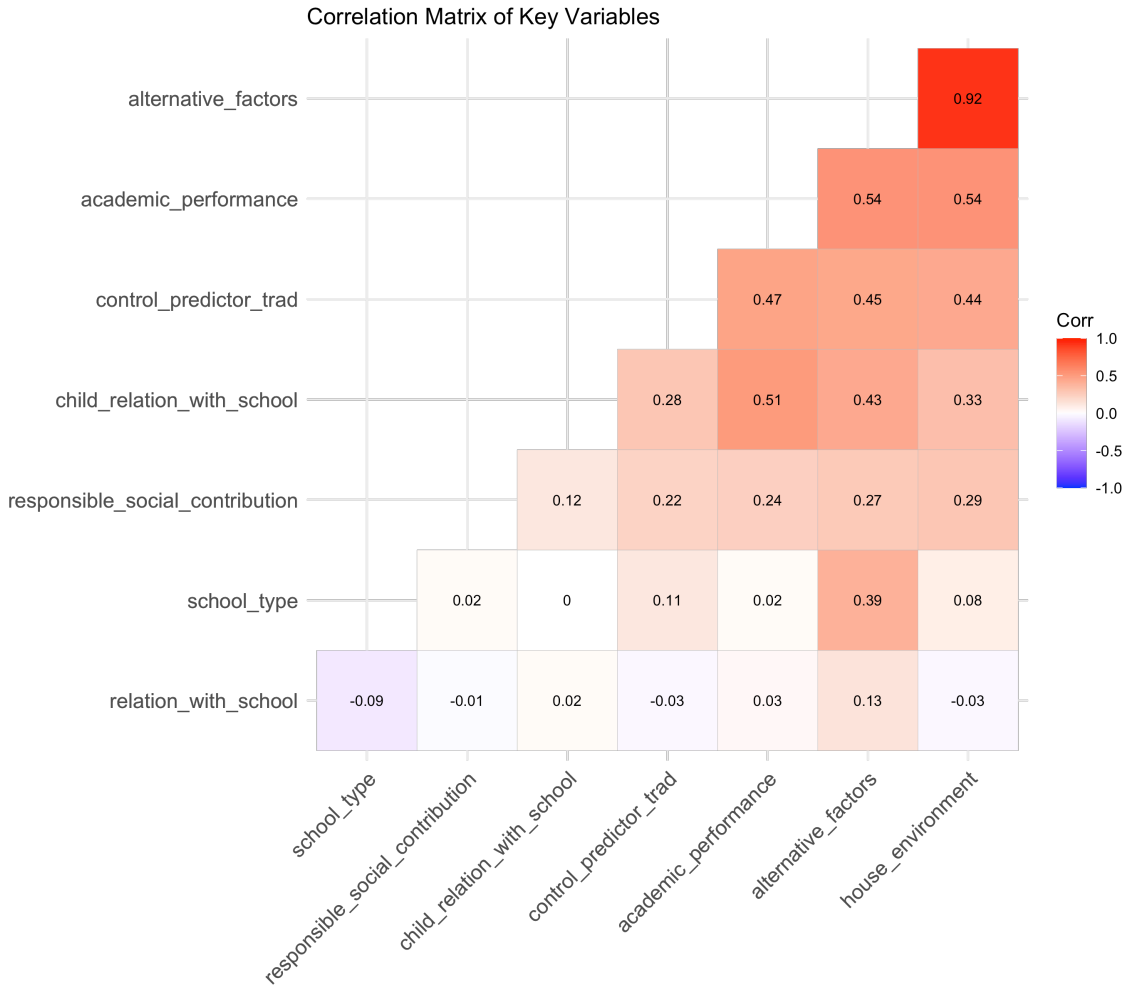


Correlation Analysis

To understand the relationships between our key variables, we calculate a correlation matrix. The following table and heatmap display the correlation coefficients for all pairs of our key variables.

Table 3: Correlation Matrix of Key Variables

Variable	responsible_social_contribution	academic_performance	alternative_factors	relation_with_school	child_relation_with_school	school_type	house_environment	control_predictor_trad
responsible_social_contribution	1.0000000	0.2355700	0.2669733	-0.0068340	0.1215166	0.0166944	0.2880517	0.2216267
academic_performance	0.2355700	1.0000000	0.5436960	0.0263204	0.5114284	0.0181278	0.5375345	0.4689804
alternative_factors	0.2669733	0.5436960	1.0000000	0.1347952	0.4344648	0.3855908	0.9204812	0.4489545
relation_with_school	-0.0068340	0.0263204	0.1347952	1.0000000	0.0232013	-0.0903337	-0.0308204	-0.0304142
child_relation_with_school	0.1215166	0.5114284	0.4344648	0.0232013	1.0000000	-0.0049685	0.3326967	0.2806976
school_type	0.0166944	0.0181278	0.3855908	-0.0903337	-0.0049685	1.0000000	0.0828899	0.1083056
house_environment	0.2880517	0.5375345	0.9204812	-0.0308204	0.3326967	0.0828899	1.0000000	0.4436940
control_predictor_trad	0.2216267	0.4689804	0.4489545	-0.0304142	0.2806976	0.1083056	0.4436940	1.0000000



The correlation heatmap provides a visual representation of the correlation matrix. The color and size of the circles indicate the strength and direction of the correlations. As we can see, there are a number of interesting correlations

in our data. For example, `academic_performance` and `alternative_factors` are both positively correlated with `responsible_social_contribution`.

Summary of all Factors by Gender

This table helps us understand if there are systematic differences in our key variables between males and females in our sample. This is important for ensuring that our subsequent regression models are not confounded by gender-based disparities.

Table 4: Summary of all Factors by Gender

gender	responsible_social_contribution_mean	responsible_social_contribution_sd	academic_performance_mean	academic_performance_sd	alternative_factors_mean	alternative_factors_sd	family_influence_mean	family_influence_sd	peer_influence_mean	peer_influence_sd	school_resources_mean	school_resources_sd	skill_development_mean	skill_development_sd	community_support_mean	community_support_sd	perceived_importance_mean	perceived_importance_sd	count
Male	0.033	0.211	0.215	0.507	0.102	0.147	0.006	0.111	0.048	0.082	-0.017	0.202	0.372	0.500	0.752	0.901	24803
Female	0.134	0.137	0.070	0.447	0.088	0.143	0.004	0.106	0.024	0.096	-0.033	0.171	0.356	0.482	0.258	0.606	9499

On average, females in our sample report a higher level of `responsible_social_contribution` (0.168) compared to males (-0.009). Conversely, males tend to have a higher `academic_performance` score (0.292) than females (0.112). This suggests that the relationship between academic performance and social contribution may differ by gender, a dynamic we will explore in our regression models.

Summary of all Factors by Location

This table allows us to compare our key variables across urban and rural settings. Understanding these differences is crucial, as access to educational resources and opportunities can vary significantly between these locations.

Table 5: Summary of all Factors by Location

urban	responsible_social_contribution_mean	responsible_social_contribution_sd	academic_performance_mean	academic_performance_sd	alternative_factors_mean	alternative_factors_sd	family_influence_mean	family_influence_sd	peer_influence_mean	peer_influence_sd	school_resources_mean	school_resources_sd	skill_development_mean	skill_development_sd	community_support_mean	community_support_sd	perceived_importance_mean	perceived_importance_sd	count
Urban	0.098	0.197	0.352	0.587	0.158	0.140	0.006	0.101	0.060	0.073	0.001	0.180	0.563	0.488	0.993	1.018	10187
Rural	0.046	0.198	0.100	0.429	0.073	0.141	0.005	0.113	0.033	0.091	-0.031	0.199	0.285	0.474	0.456	0.725	24115

Individuals in urban areas score higher on average on all our key metrics: `responsible_social_contribution` (0.066 vs. 0.025), `academic_performance` (0.526 vs. 0.126), and `alternative_factors` (0.175 vs. 0.056). This highlights the importance of controlling for urban/rural status in our analysis to avoid attributing these differences to other factors.

Summary of all Factors by Work Type

This table breaks down our key variables by the primary type of work individuals are engaged in. This helps us understand how different economic activities might relate to our variables of interest.

Table 6: Summary of all Factors by Work Type

work_type	responsible_social_contribution_mean	responsible_social_contribution_sd	academic_performance_mean	academic_performance_sd	alternative_factors_mean	alternative_factors_sd	relation_with_school_mean	relation_with_school_sd	child_relation_with_school_mean	child_relation_with_school_sd	school_type_mean	school_type_sd	house_environment_mean	house_environment_sd	urban_mean	urban_sd	count
Animal	0.148	0.160	-0.007	0.336	0.059	0.144	-0.003	0.129	0.019	0.102	-0.036	0.197	0.255	0.486	0.187	0.532	4065
Farm	0.018	0.196	0.033	0.359	0.058	0.137	0.006	0.112	0.026	0.093	-0.034	0.200	0.235	0.463	0.368	0.669	12708
Business	0.050	0.191	0.206	0.486	0.121	0.143	0.004	0.107	0.050	0.080	-0.010	0.177	0.440	0.494	0.748	0.940	6927
Salary	0.086	0.206	0.393	0.597	0.146	0.141	0.009	0.099	0.063	0.070	-0.010	0.196	0.522	0.483	0.989	0.933	10602

Individuals in salaried positions have the highest average `academic_performance` (0.563), while those working with animals have the highest `responsible_social_contribution` (0.180). This provides a preliminary indication that the link between academic success and social contribution is not straightforward and may be mediated by career path.

Distribution of Key Variables

Next, we examine the distributions of our key composite variables. The following histograms show the distribution of `responsible_social_contribution`, `academic_performance`, `alternative_factors`, `relation_with_school`, `child_relation_with_school`, `school_type` and `house_environment`.

These histograms show that our composite variables are all approximately normally distributed, which is a desirable property for regression analysis.

Bivariate Analysis

To further explore the relationships in our data, we examine how our key variables vary across different demographic groups. We will present histograms of `responsible_social_contribution`, `academic_performance`, and the sub-components of `alternative_factors` (`relation_with_school`, `child_relation_with_school`, `school_type`, `house_environment`) faceted by gender, urban/rural status, and work-related categories (business, salary, farm, animal).

Responsible Social Contribution by Demographic Factors

Interpretation: These histograms reveal nuanced differences in the distribution of `responsible_social_contribution` across various demographic and work-related groups. For instance, we can observe if certain groups tend to have higher or lower concentrations of individuals with high social contribution scores, or if the spread of scores varies significantly between groups.

Academic Performance by Demographic Factors

Interpretation: These plots illustrate how academic performance varies across different segments of the population. We can identify if there are particular demographic groups that consistently show higher or lower academic achievement, or if certain work sectors are associated with different academic backgrounds.

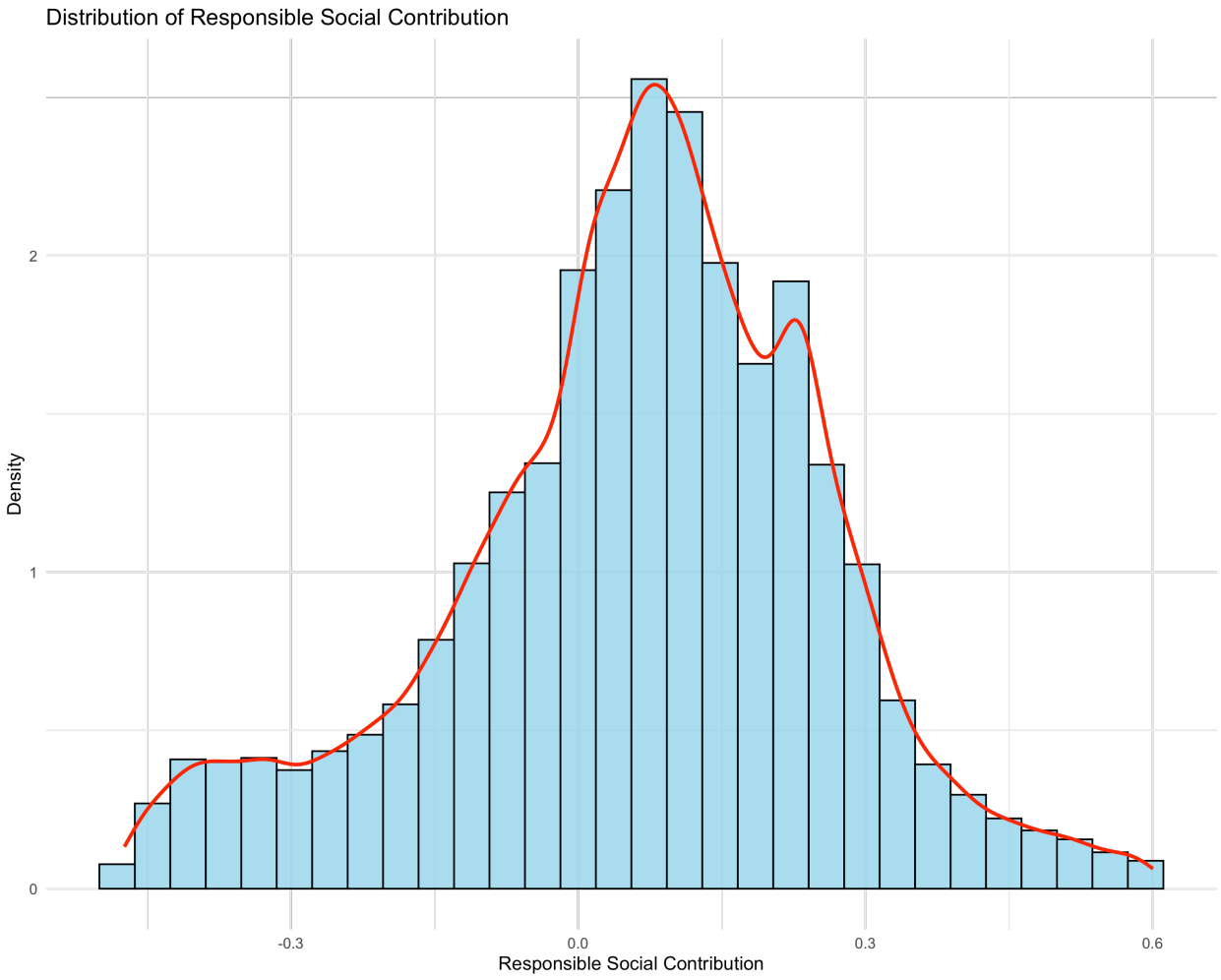


Figure 1: Distribution

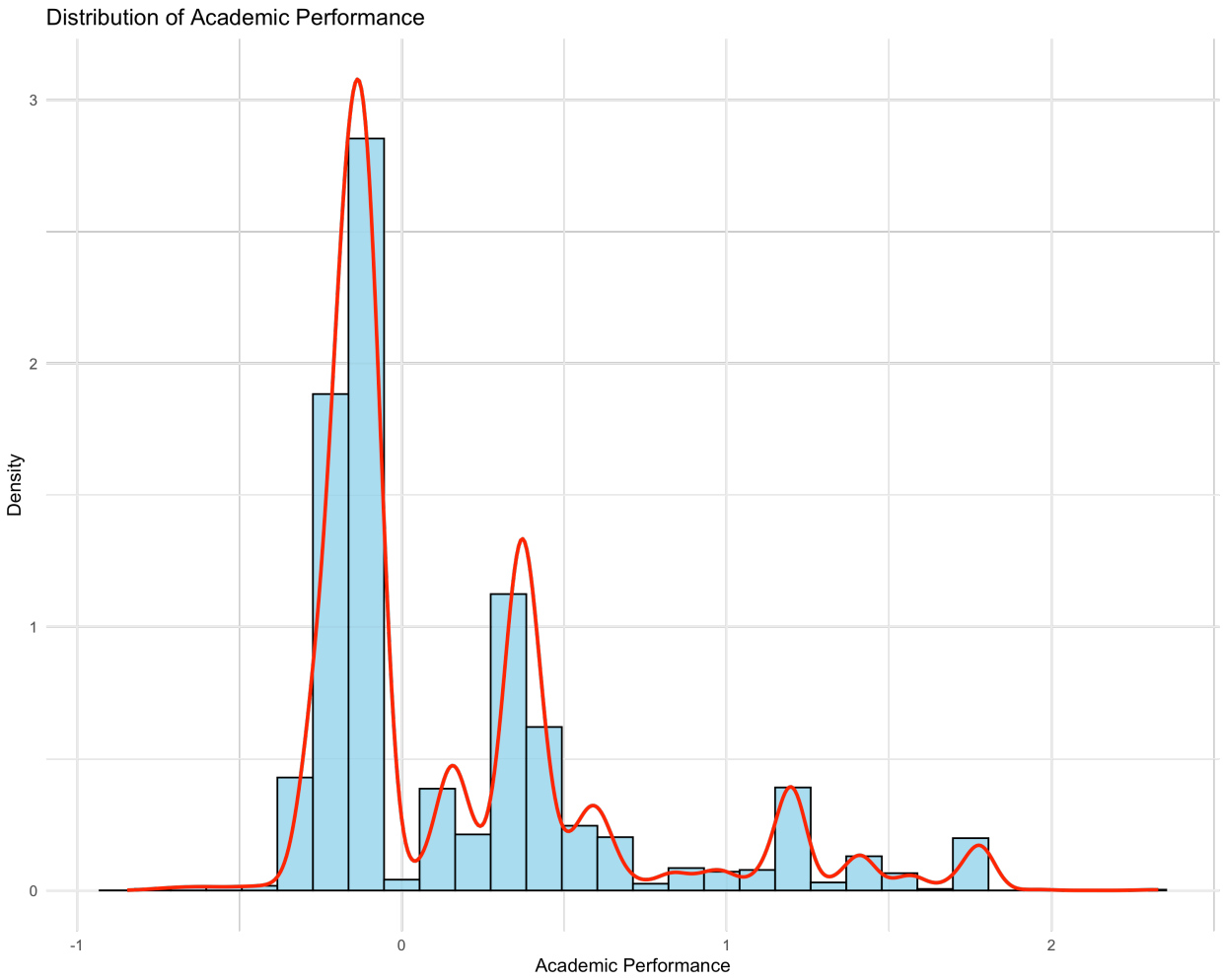


Figure 2: Distribution

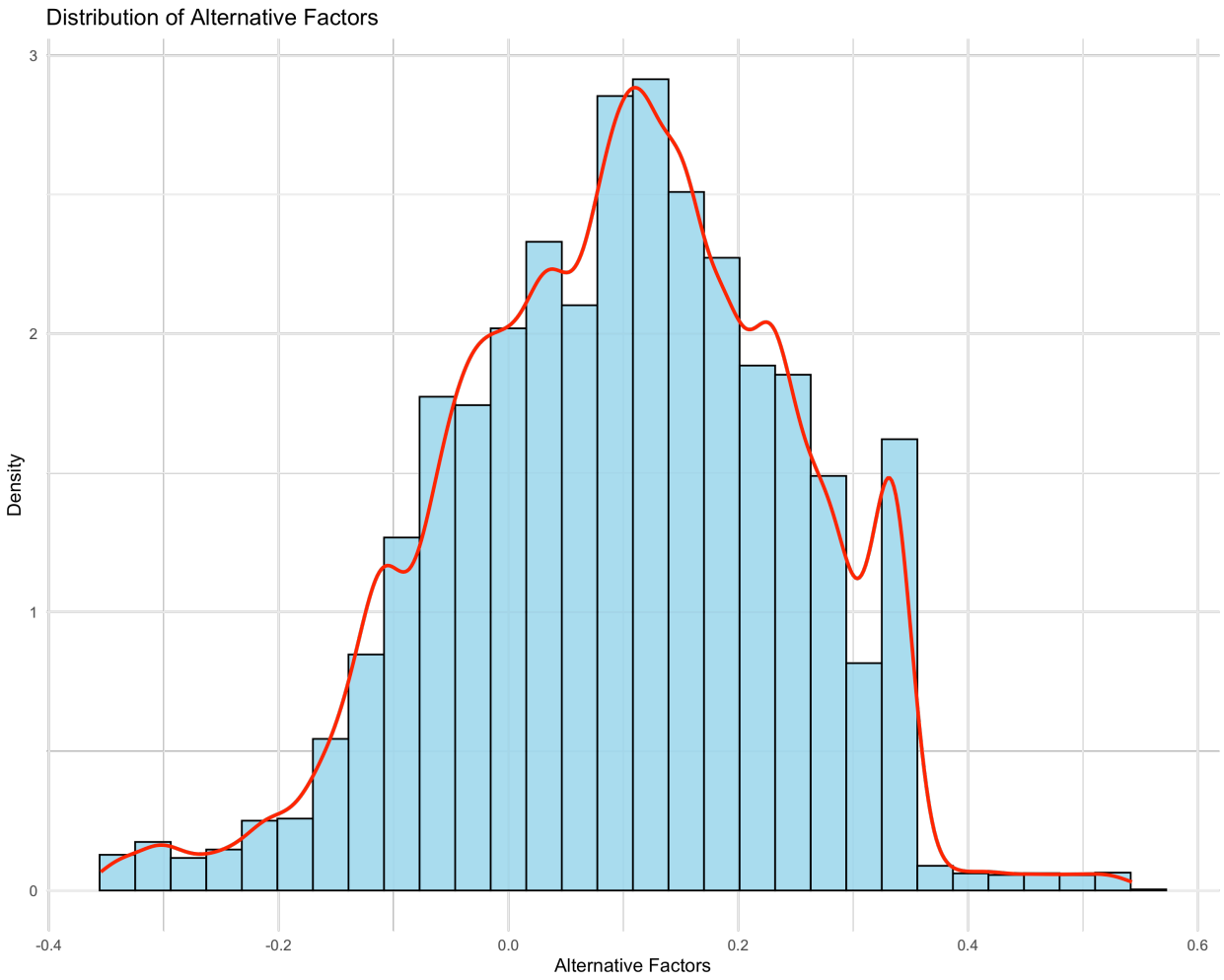


Figure 3: Distribution

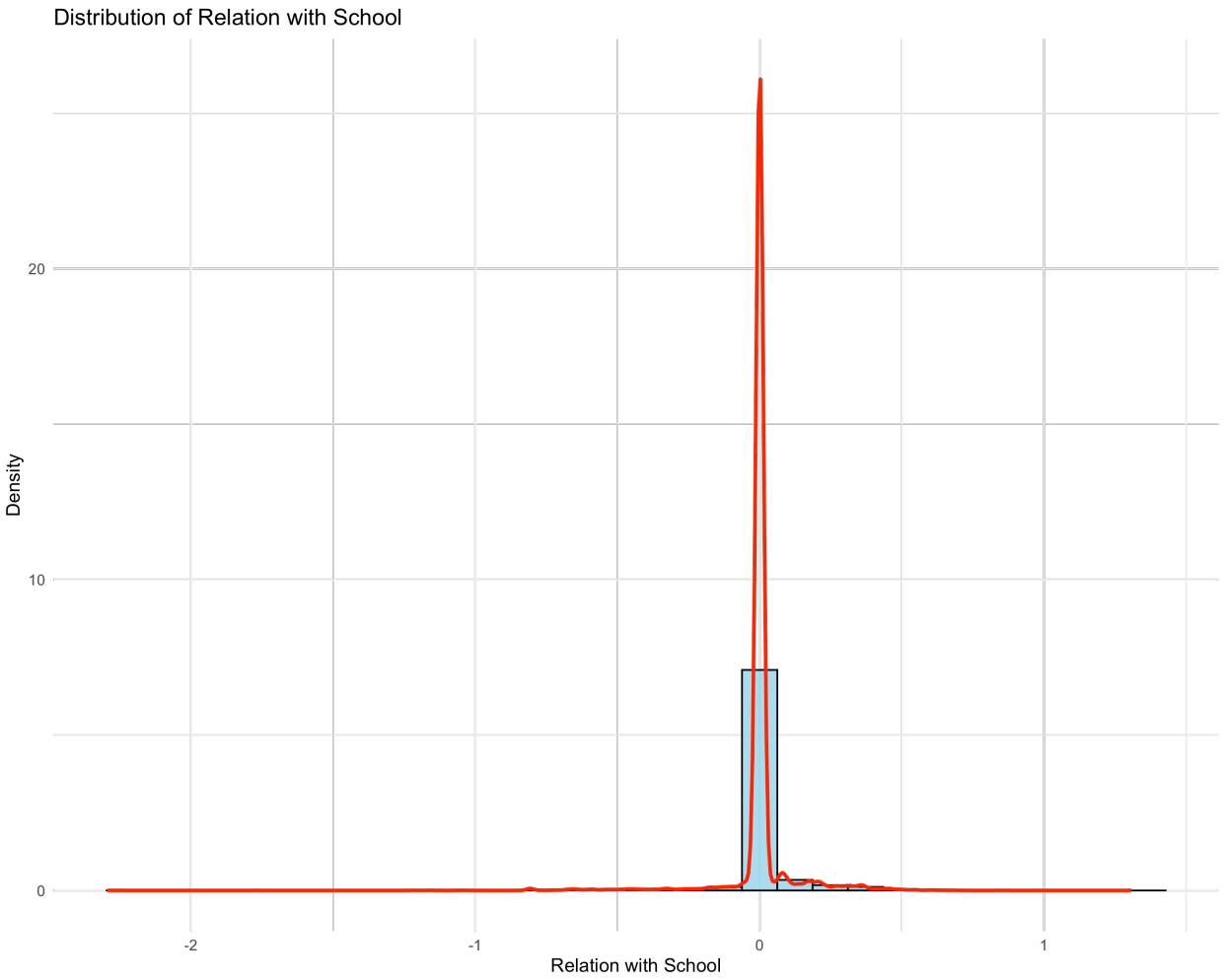


Figure 4: Distribution

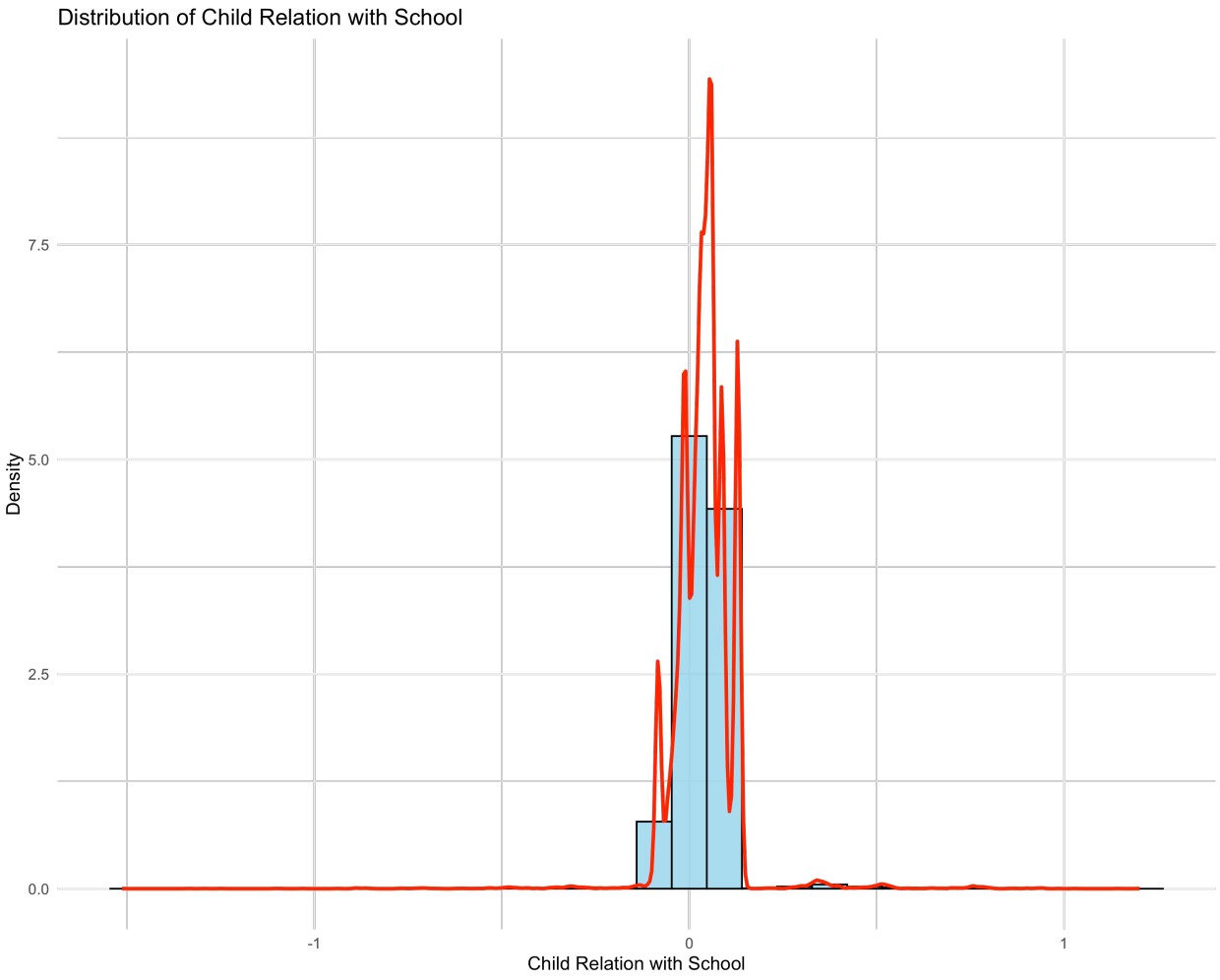


Figure 5: Distribution

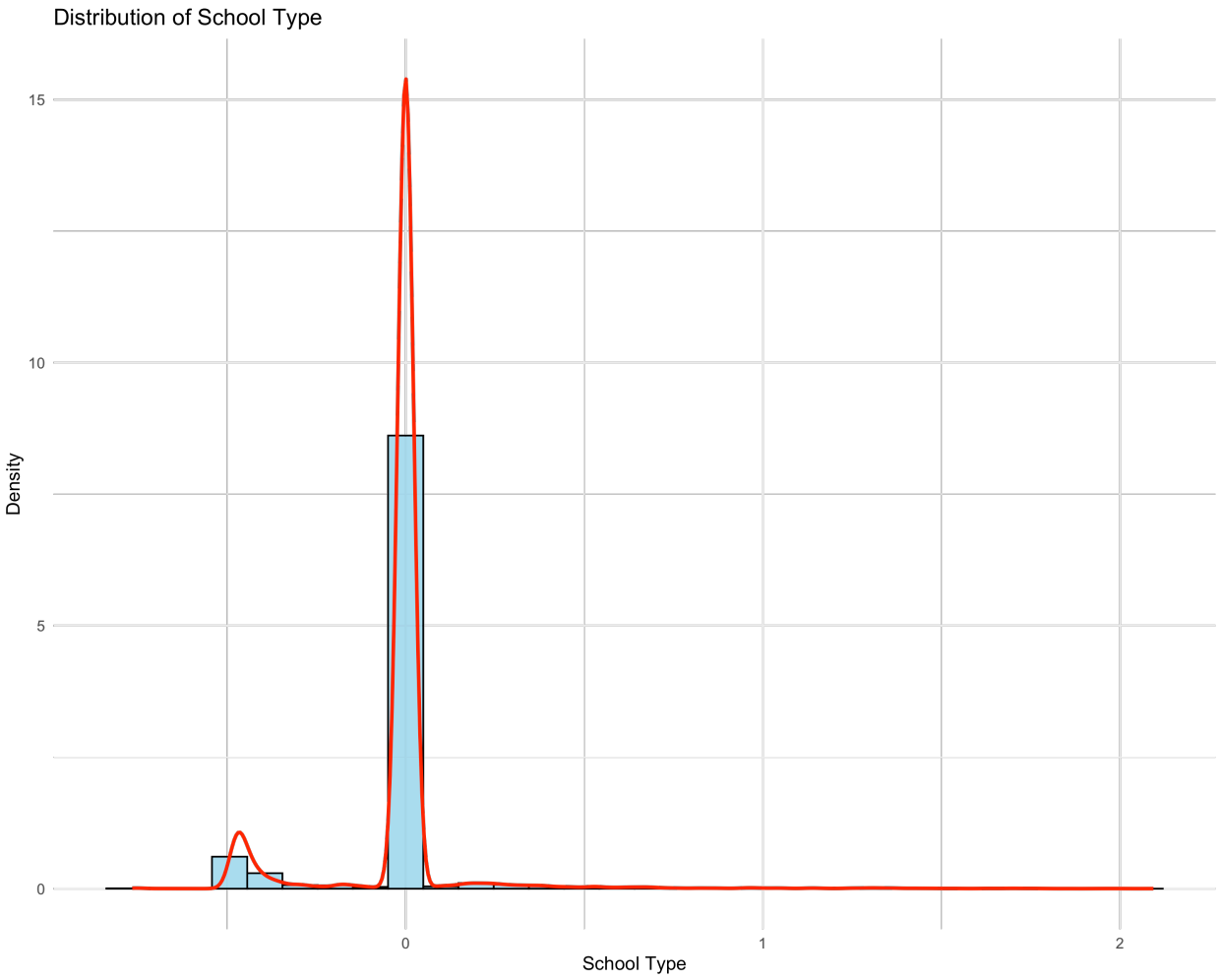


Figure 6: Distribution

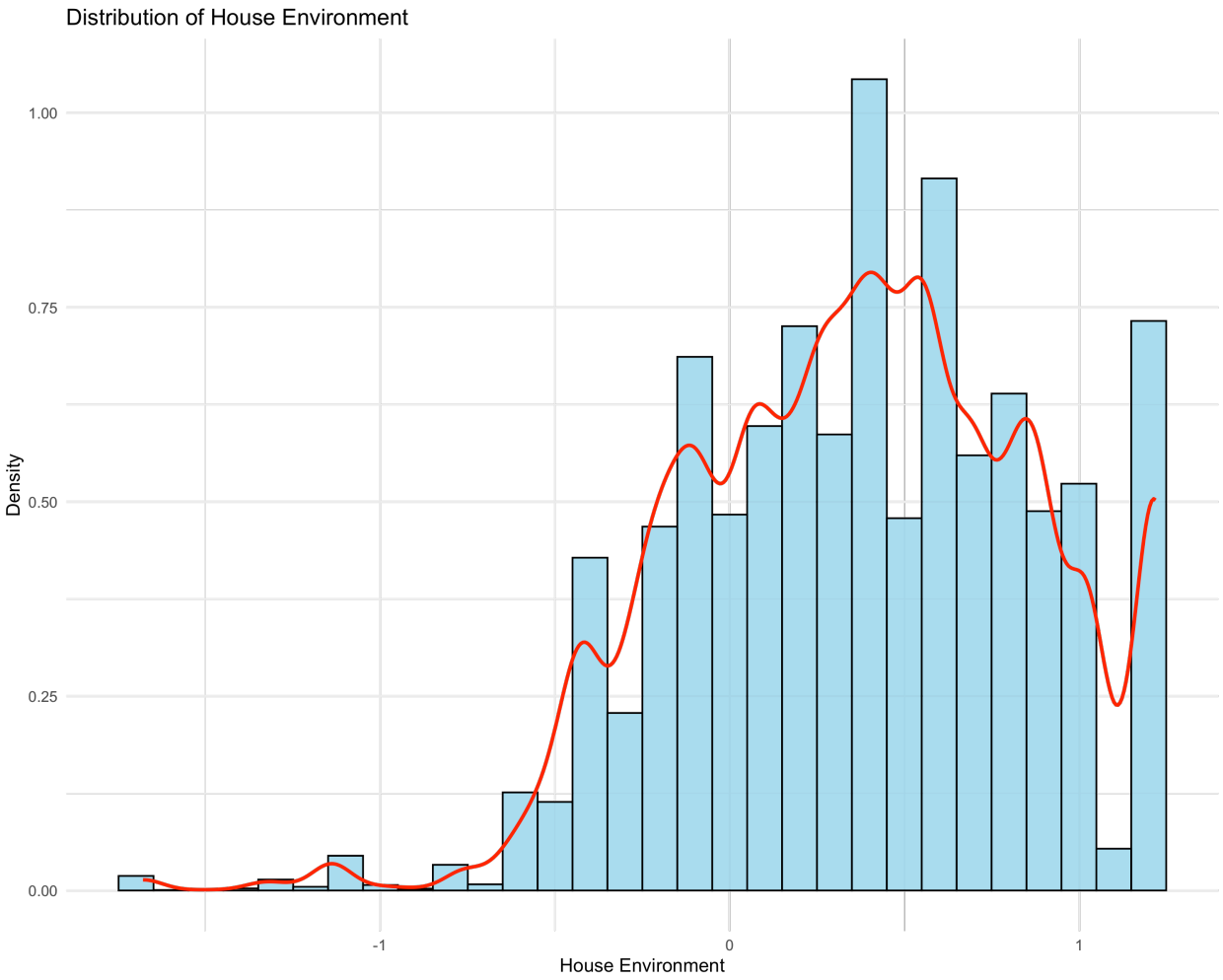


Figure 7: Distribution

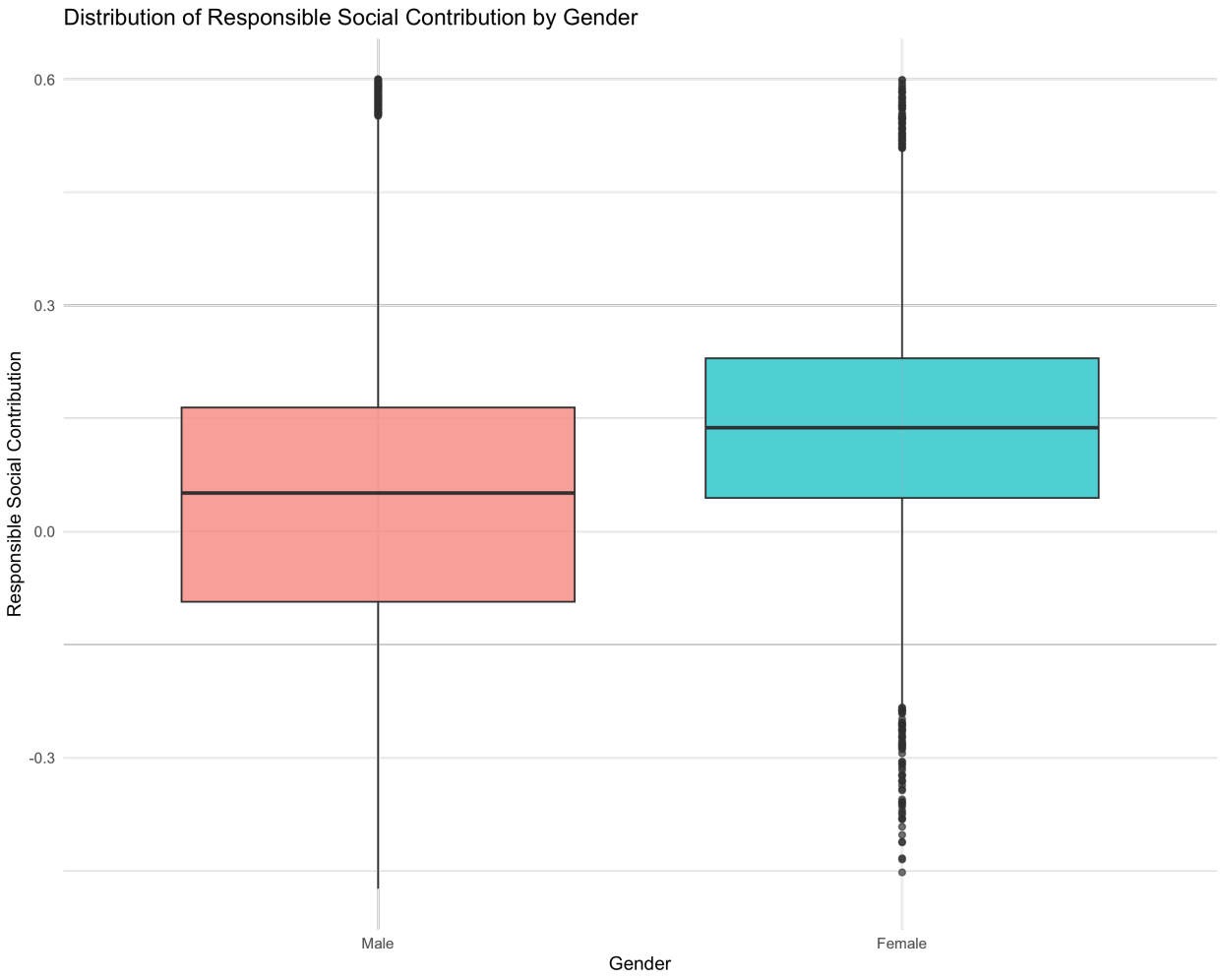


Figure 8: Boxplot

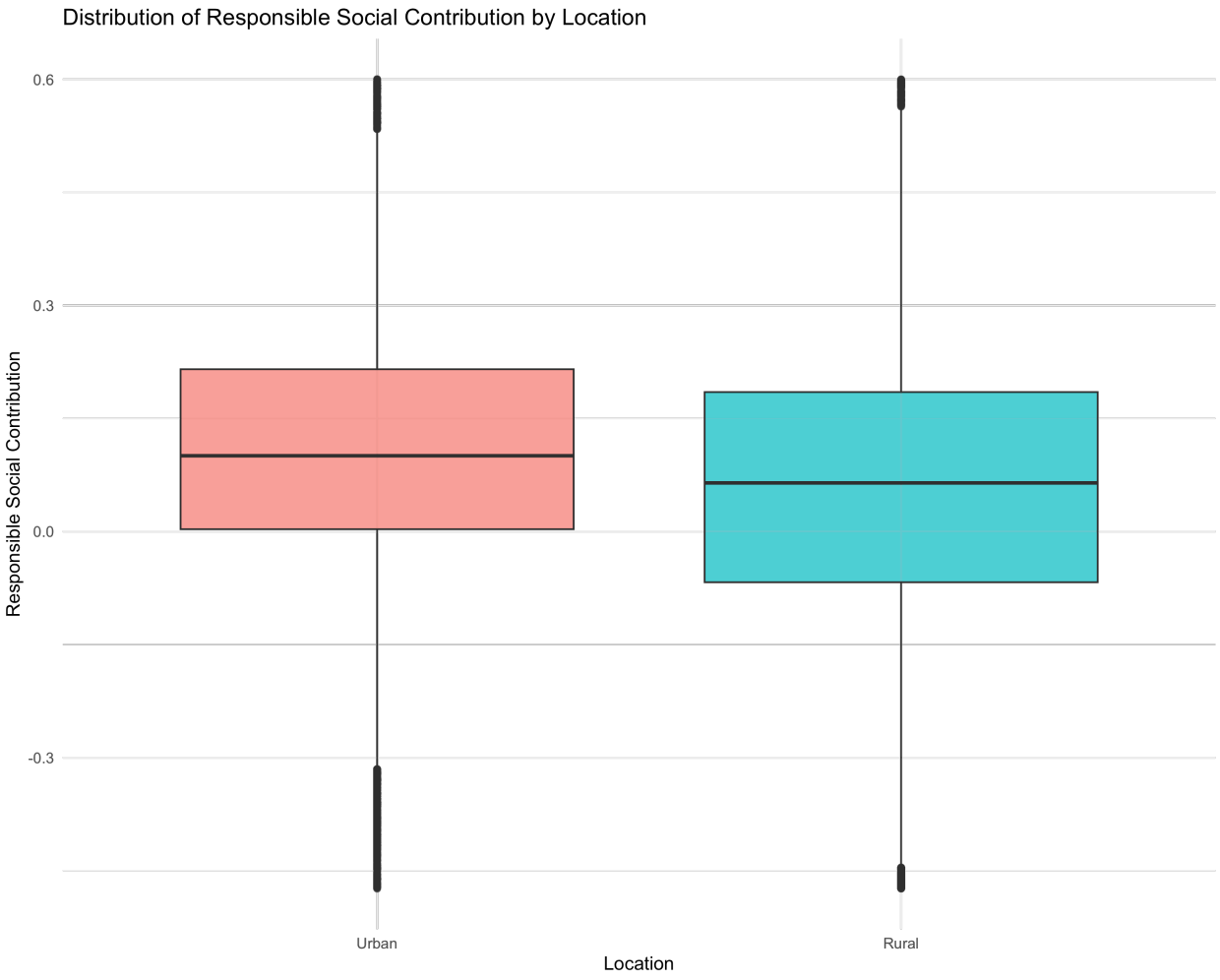


Figure 9: Boxplot

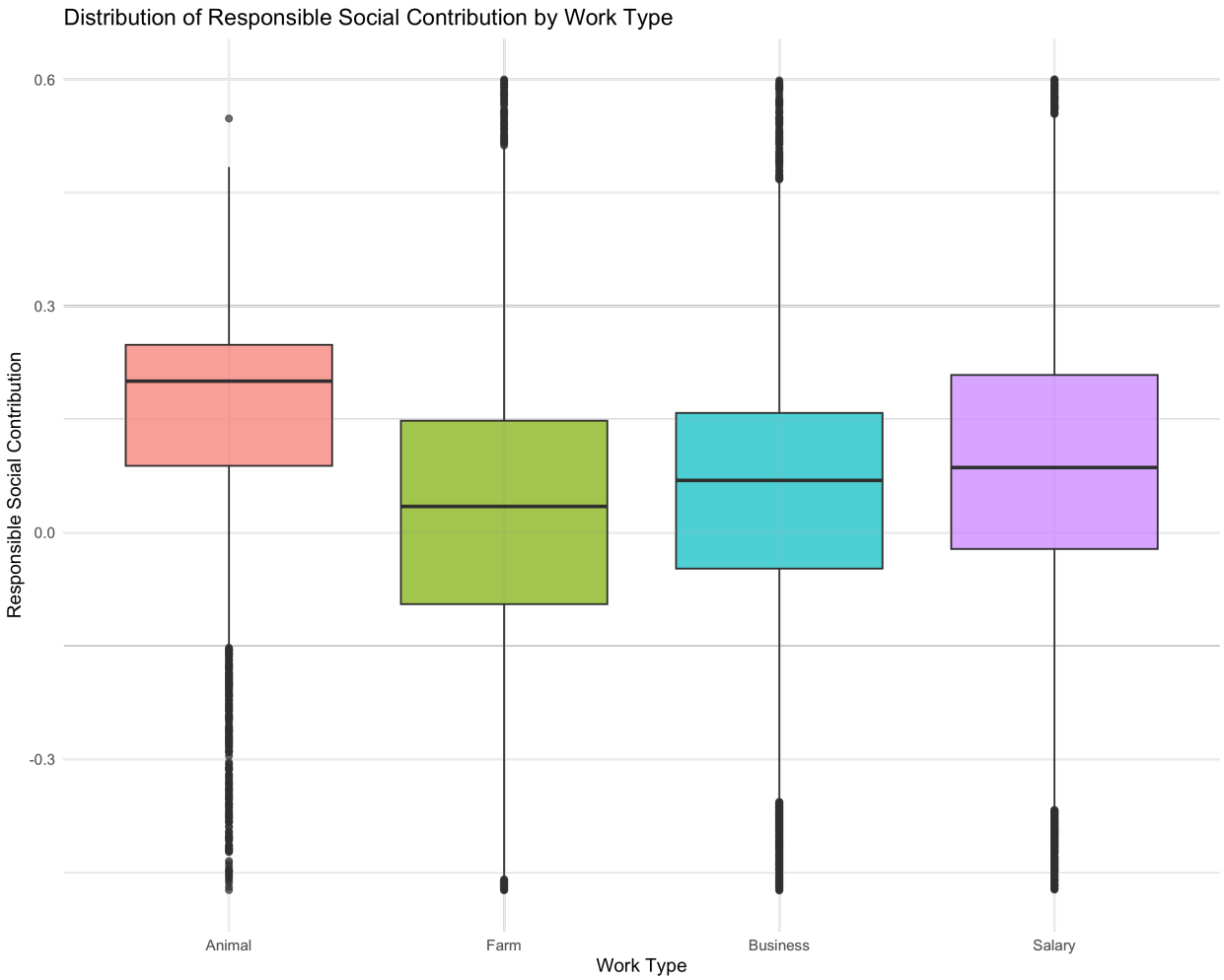


Figure 10: Boxplot

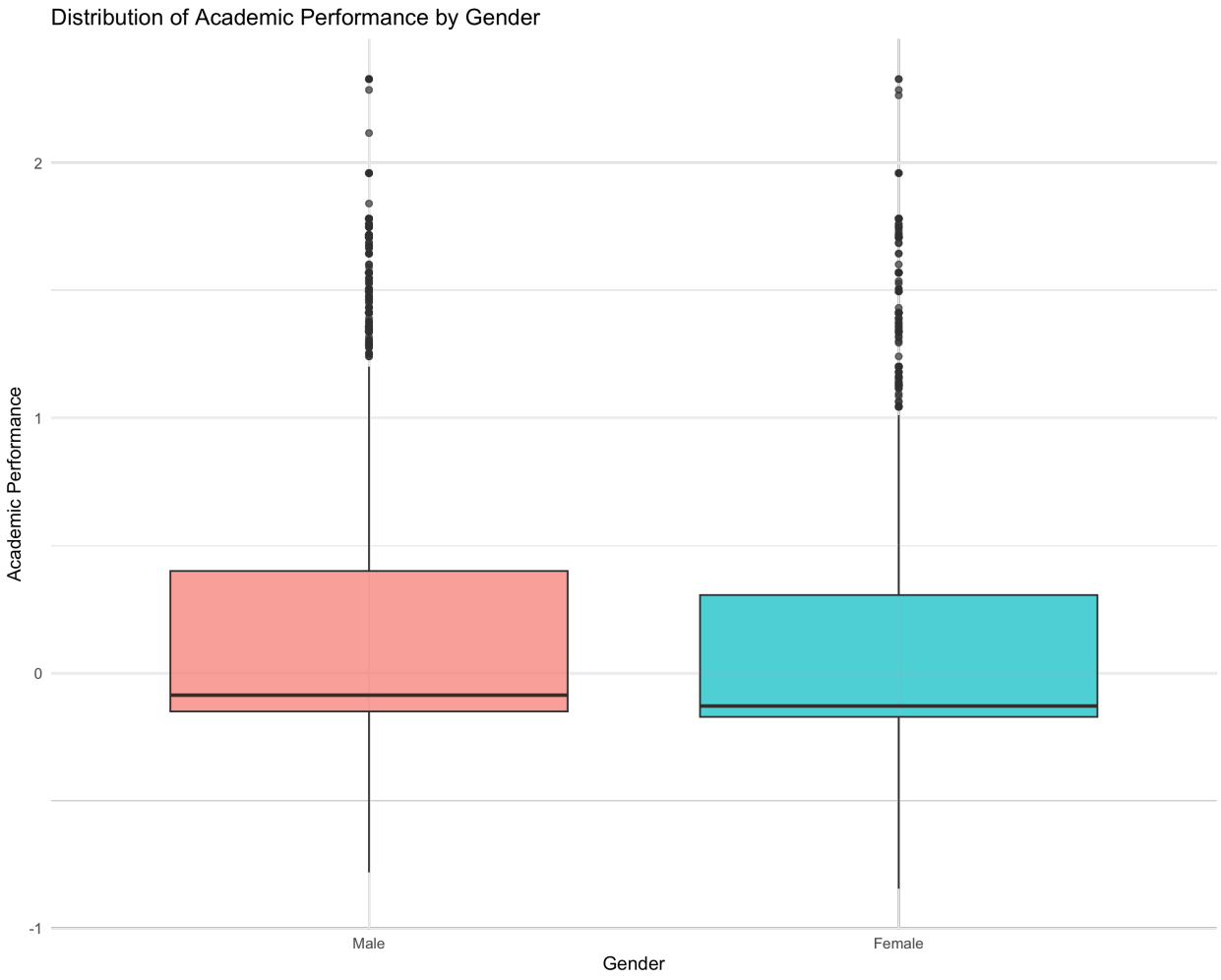


Figure 11: Boxplot

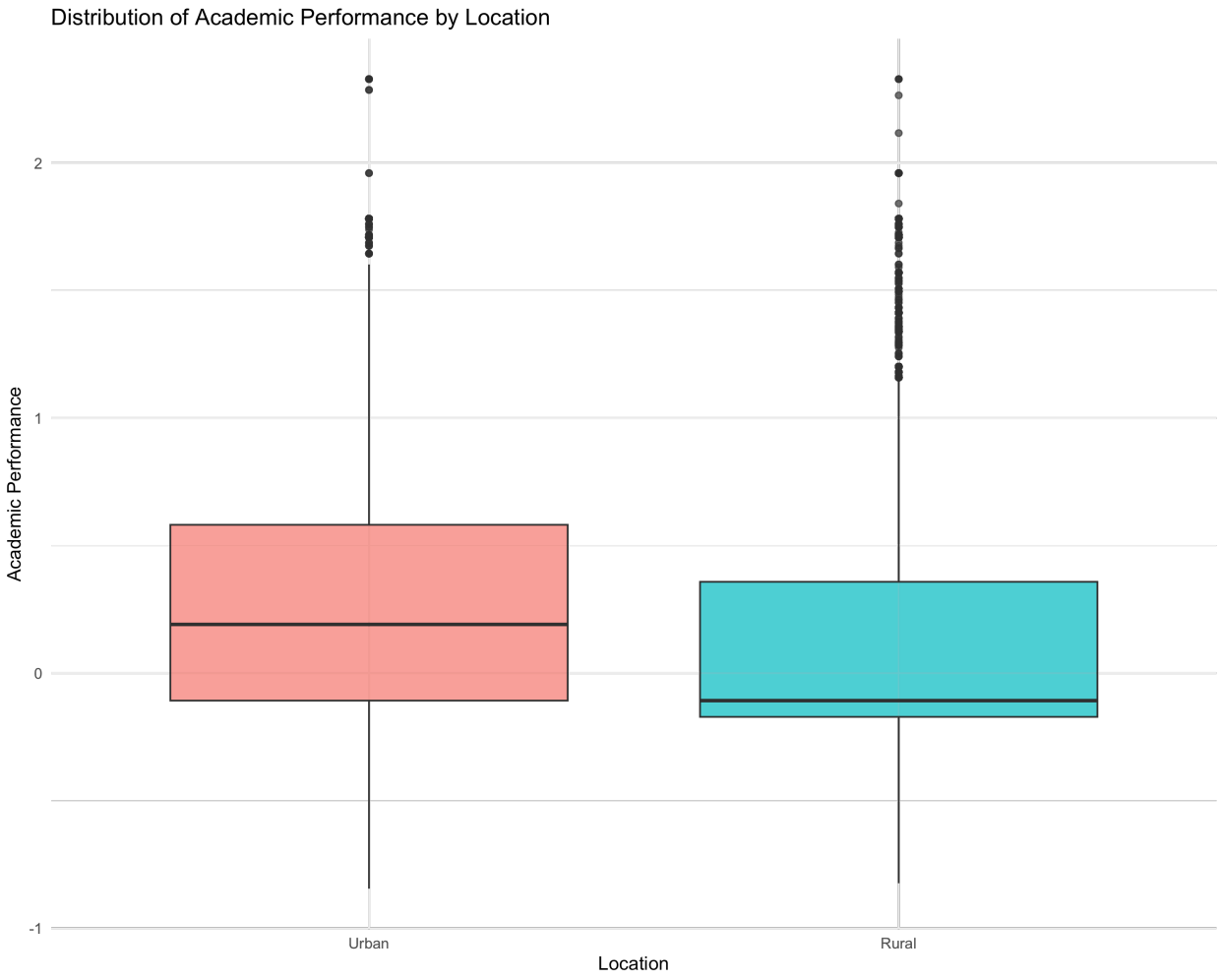


Figure 12: Boxplot

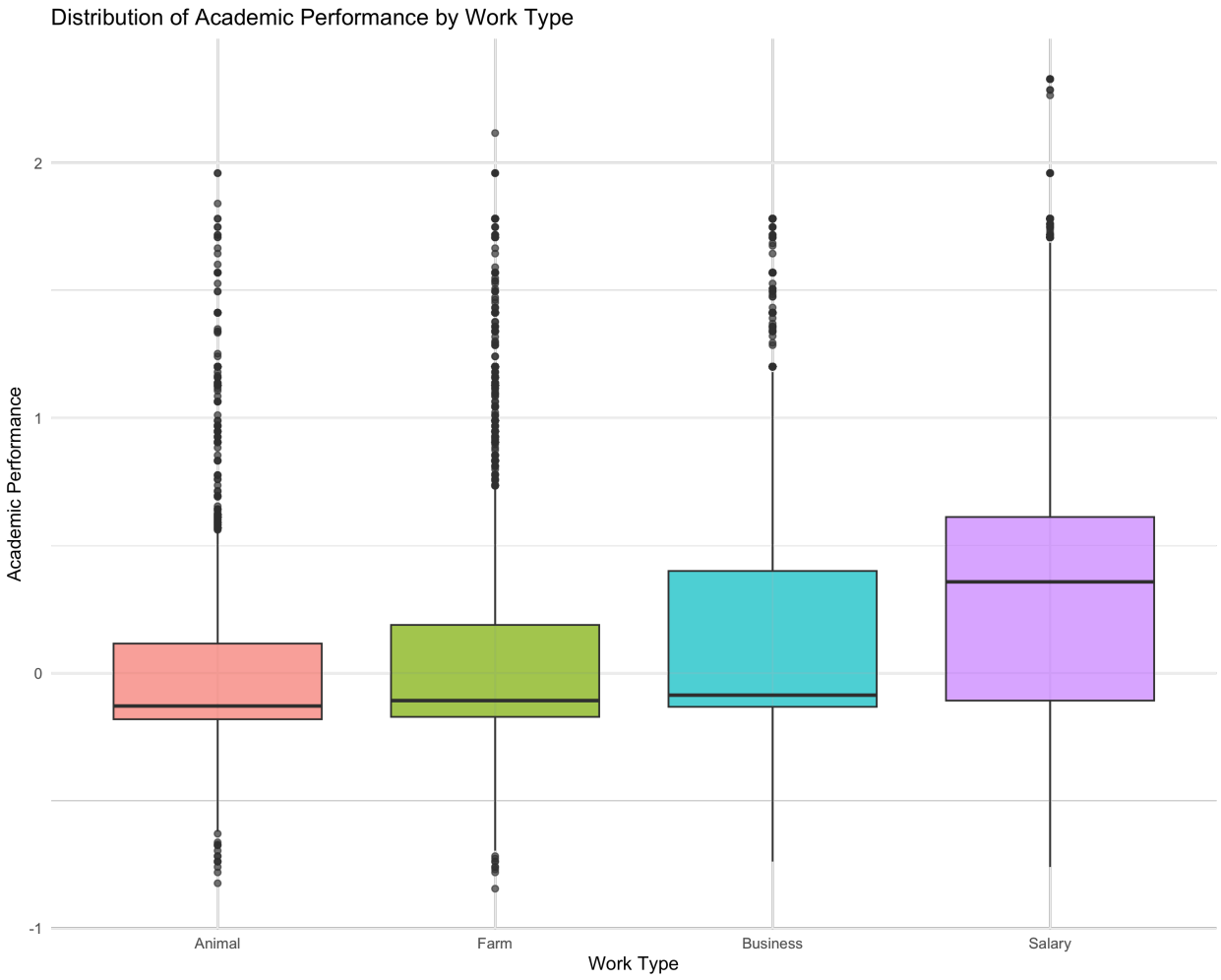


Figure 13: Boxplot

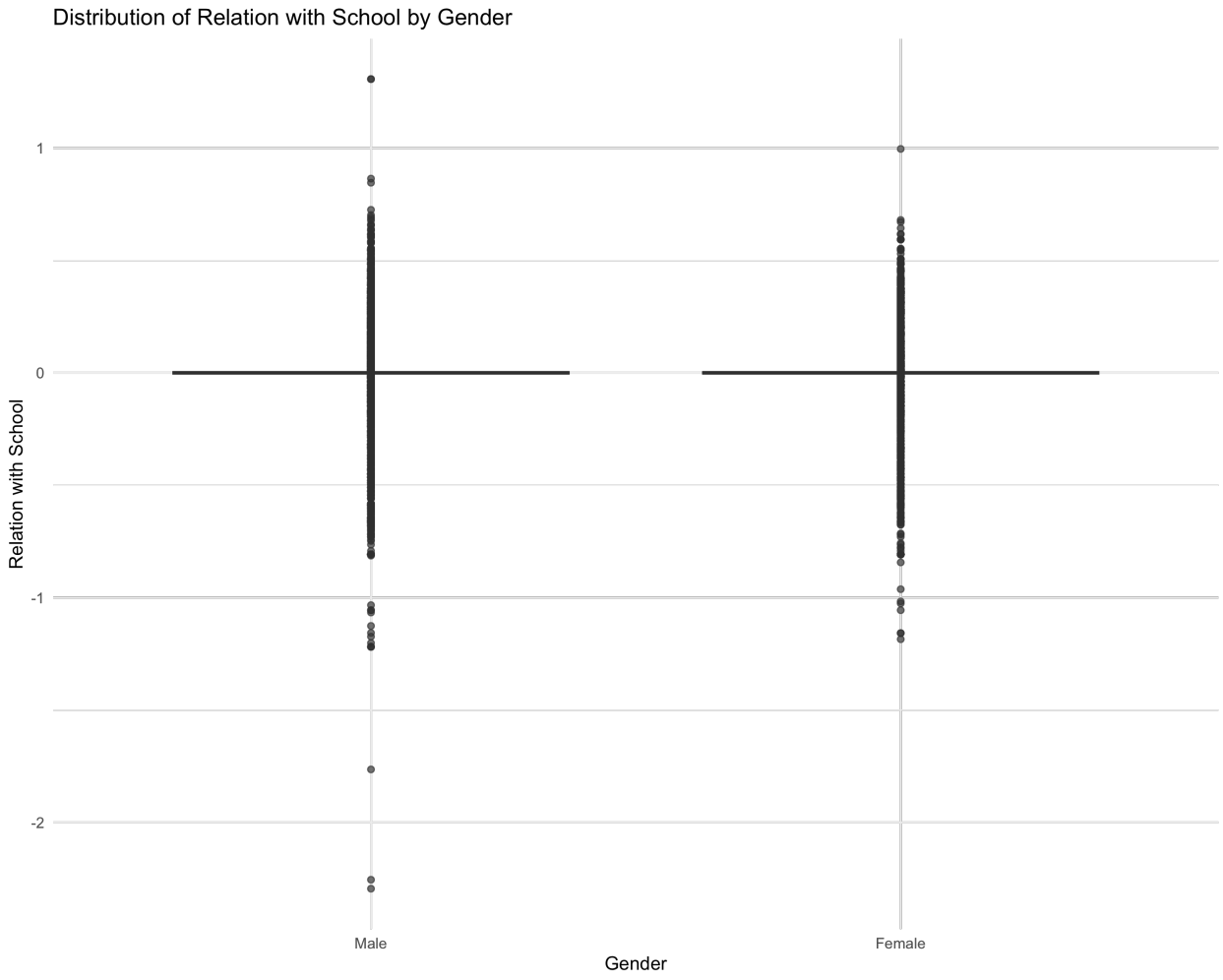


Figure 14: Boxplot

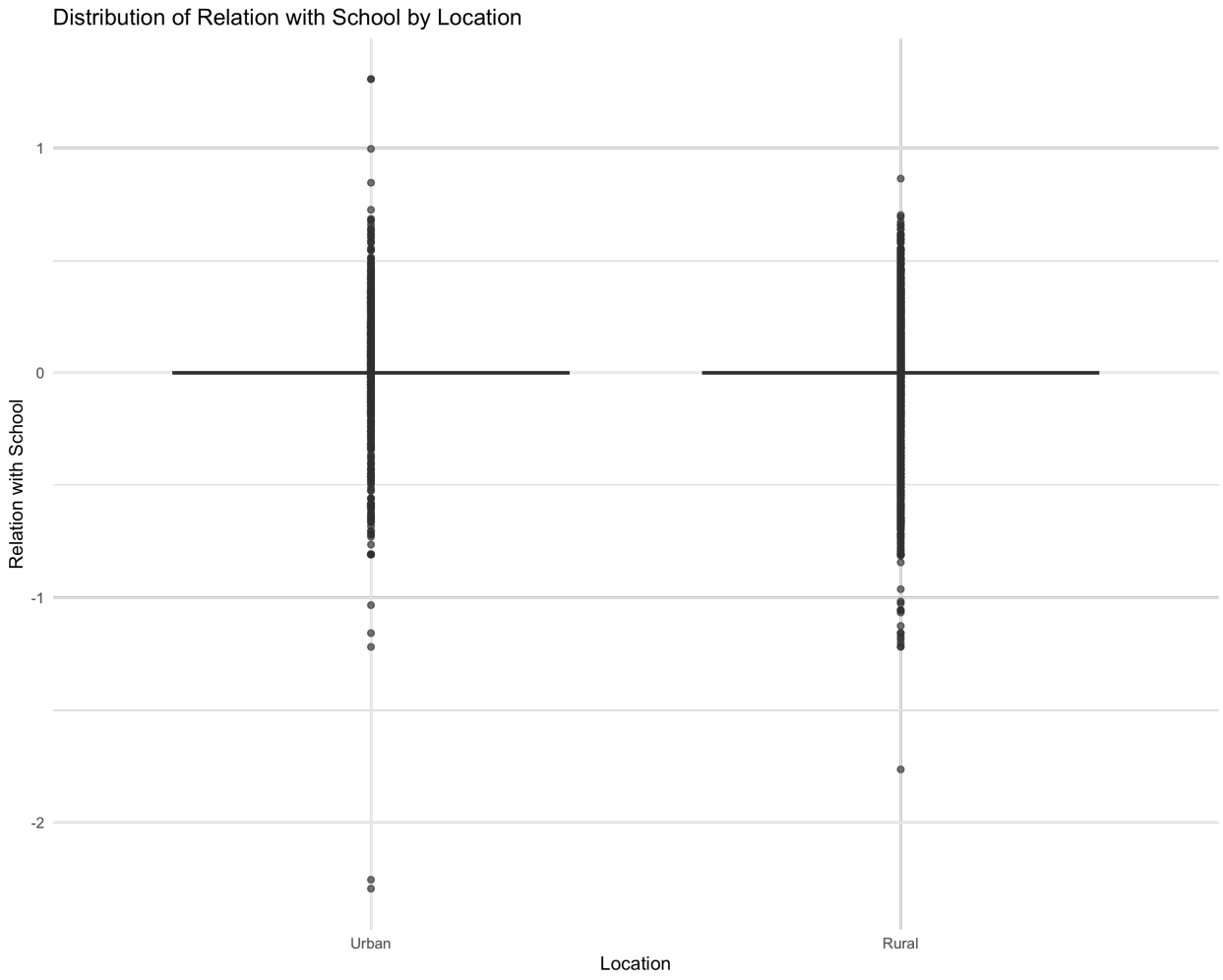


Figure 15: Boxplot

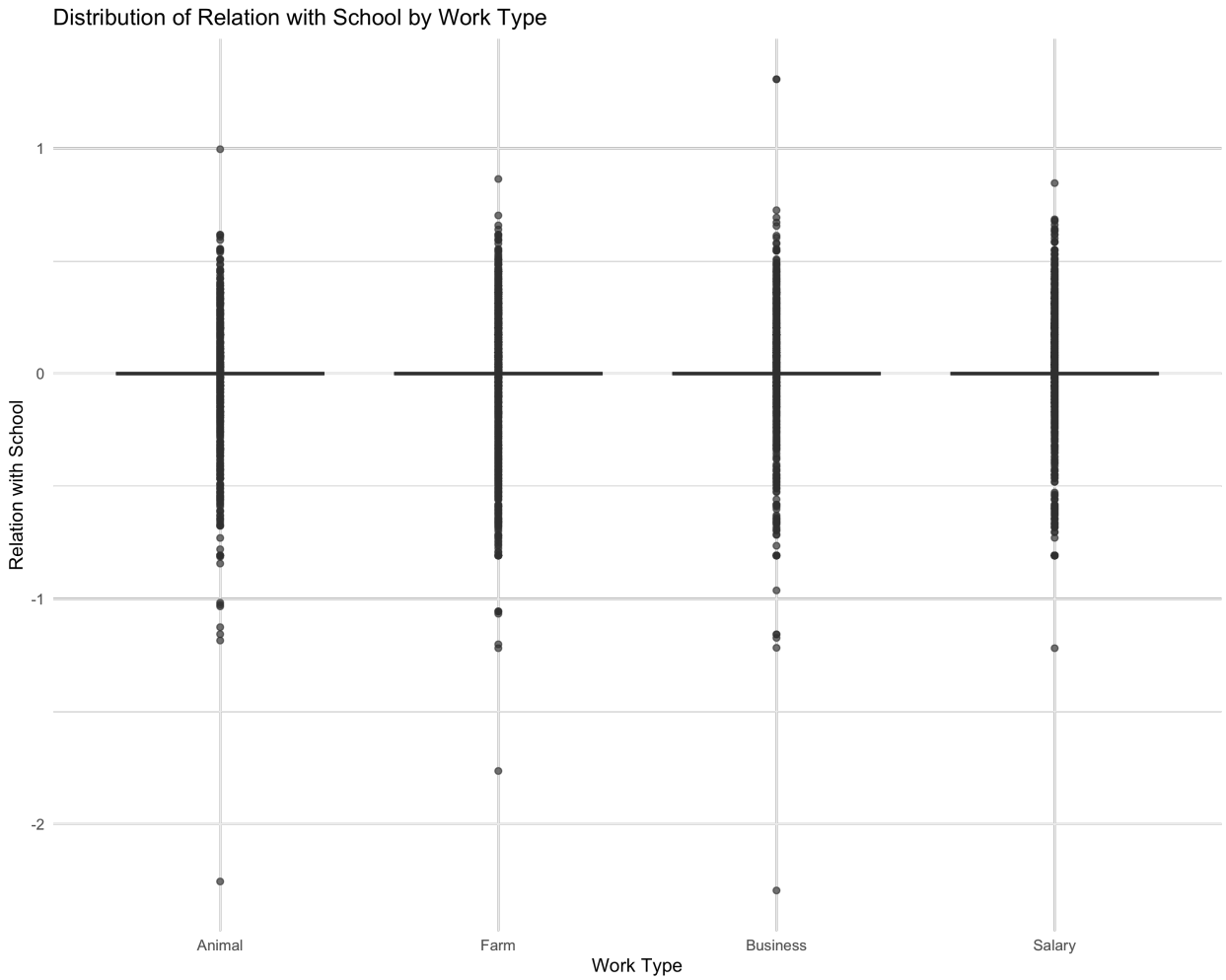


Figure 16: Boxplot

Relation with School by Demographic Factors

Interpretation: These plots provide insights into how students' relationships with their schools differ across demographic and work-related groups. We can observe if certain environments or work experiences are associated with more positive or negative school relationships.

Child Relation with School by Demographic Factors

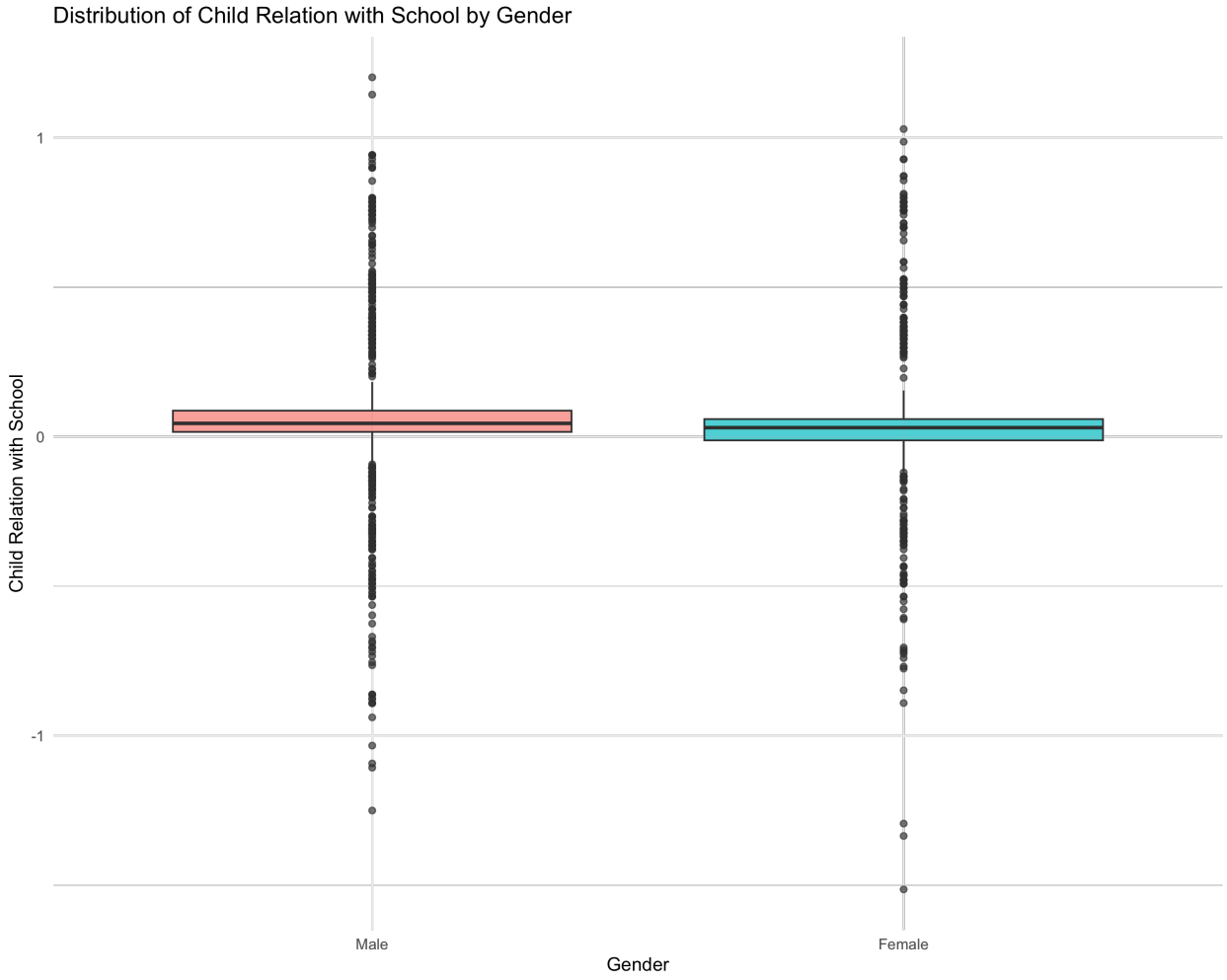


Figure 17: Boxplot

Interpretation: These plots offer a perspective on how children's relationships with their schools are influenced by various demographic and work characteristics. This can highlight disparities in school experiences from a child's viewpoint.

School Type by Demographic Factors

Interpretation: These plots illustrate the distribution of school types across different demographic and work-related groups. This can reveal patterns in access to different educational institutions based on gender, location, or economic activities.

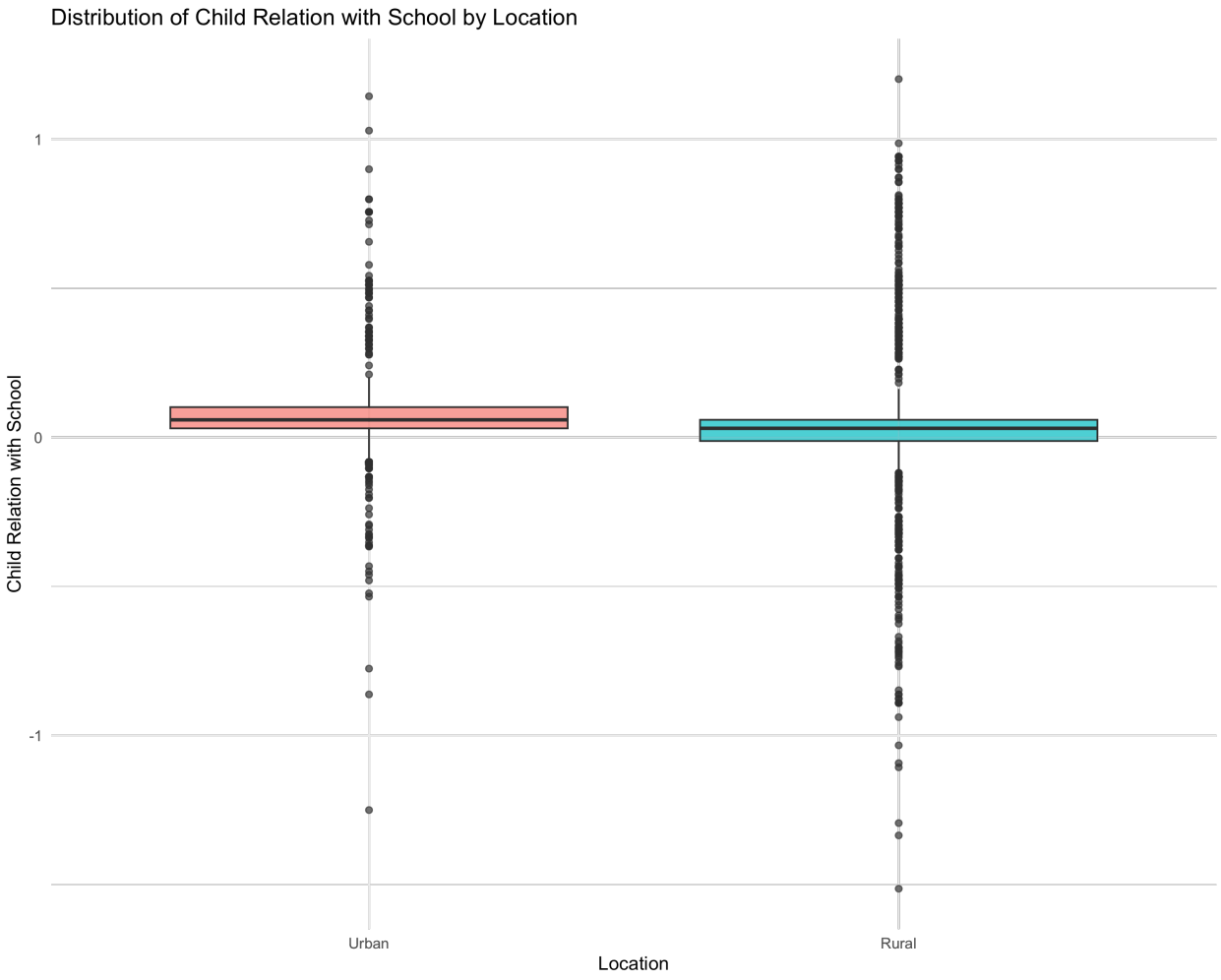


Figure 18: Boxplot

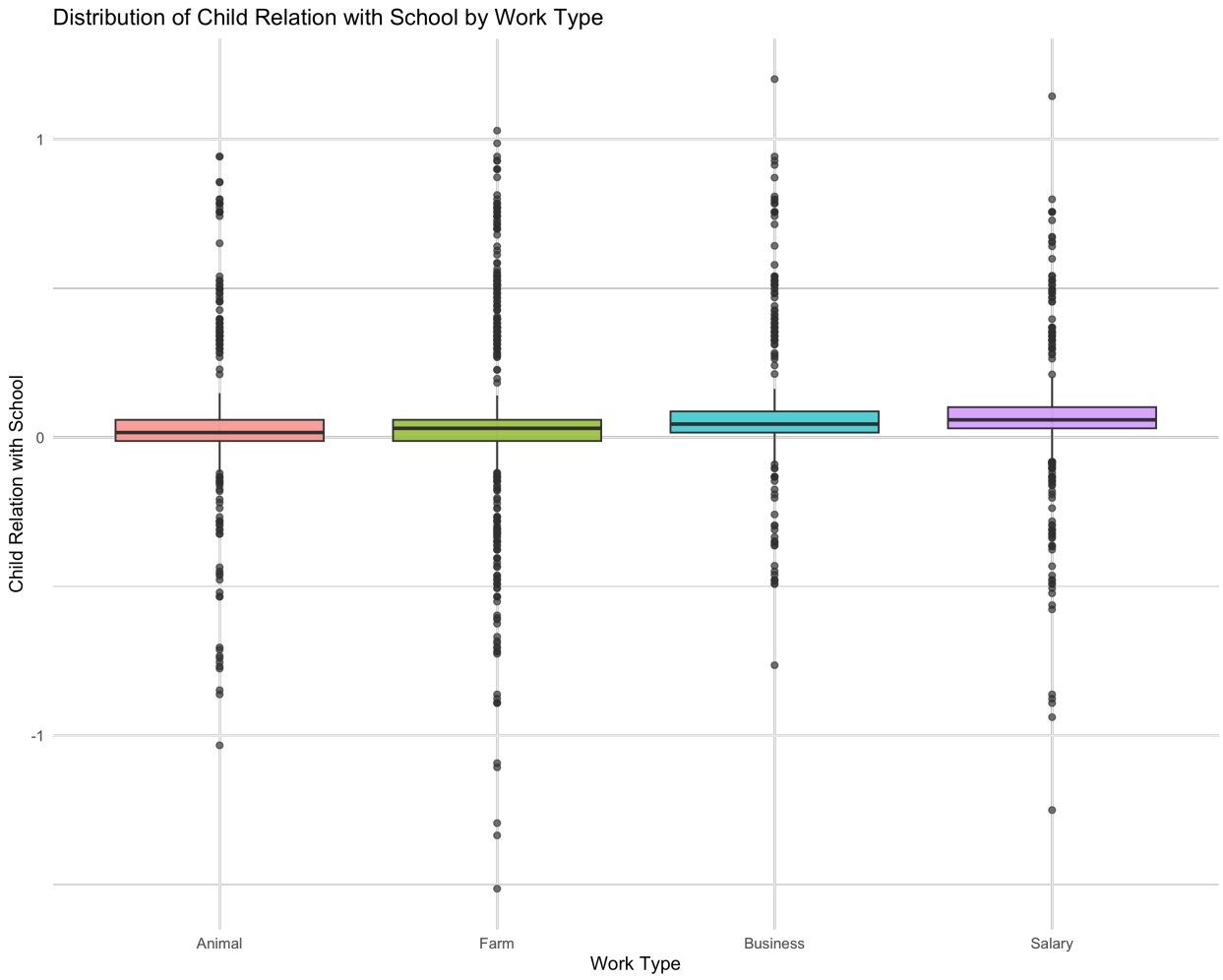


Figure 19: Boxplot

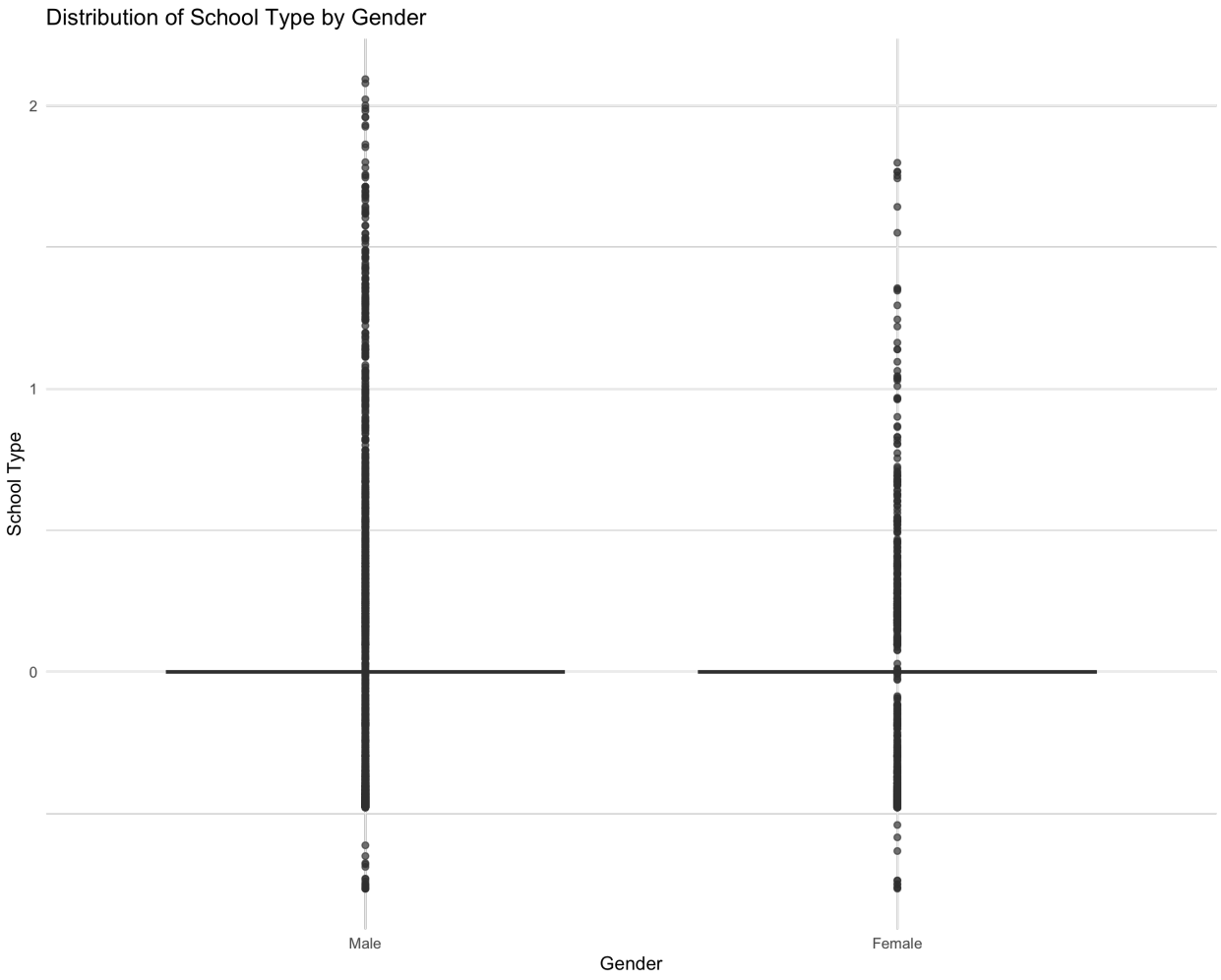


Figure 20: Boxplot

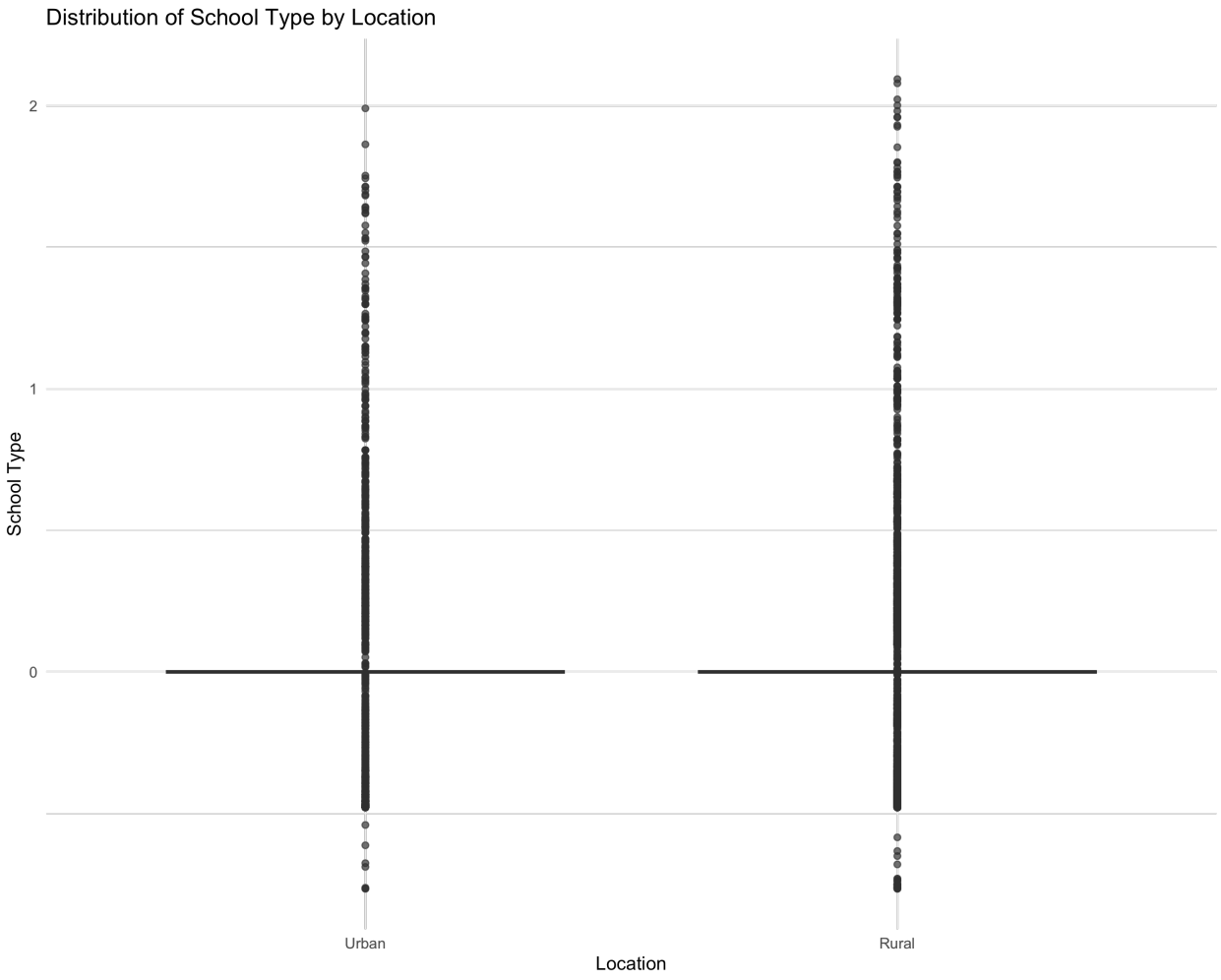


Figure 21: Boxplot

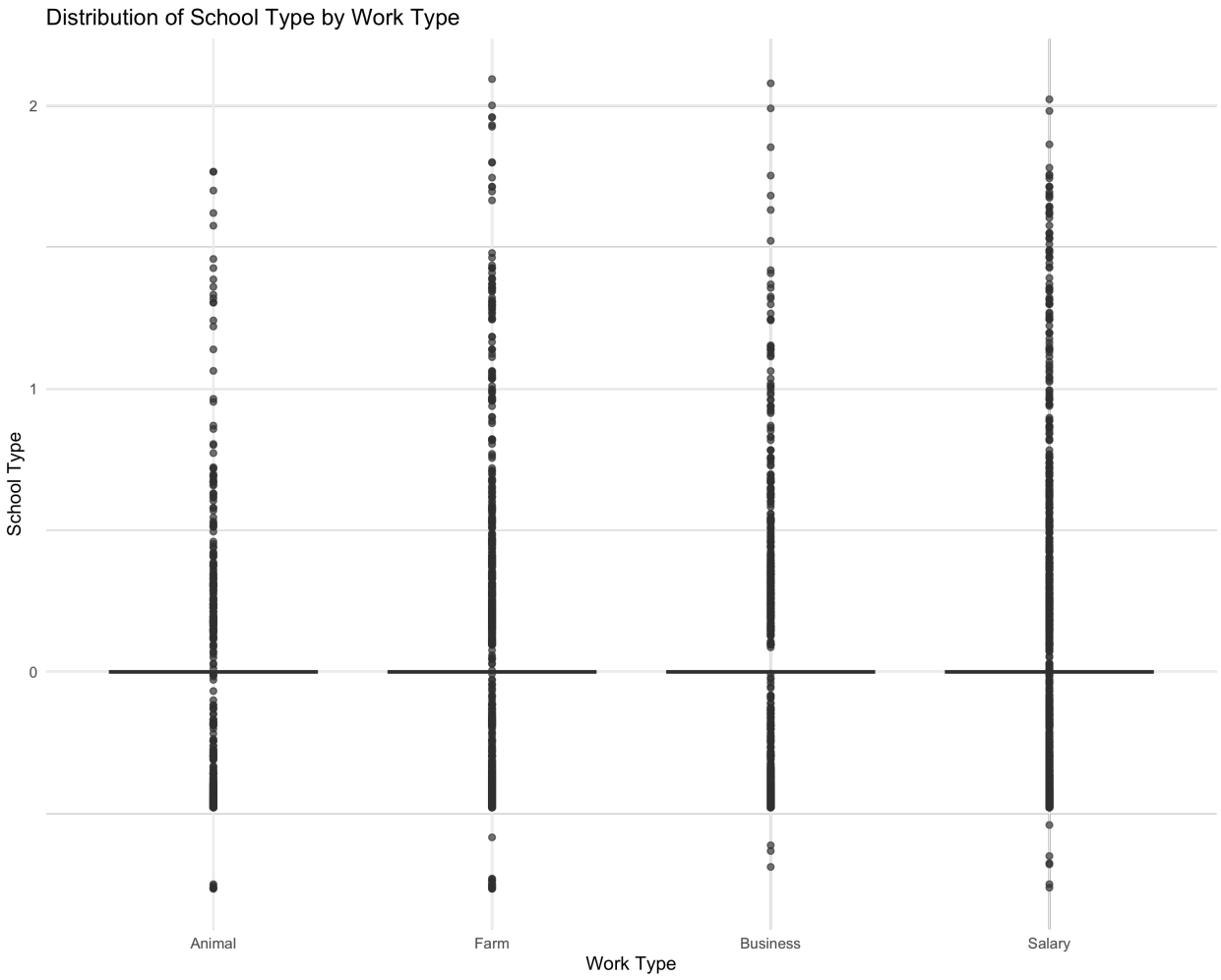


Figure 22: Boxplot

House Environment by Demographic Factors

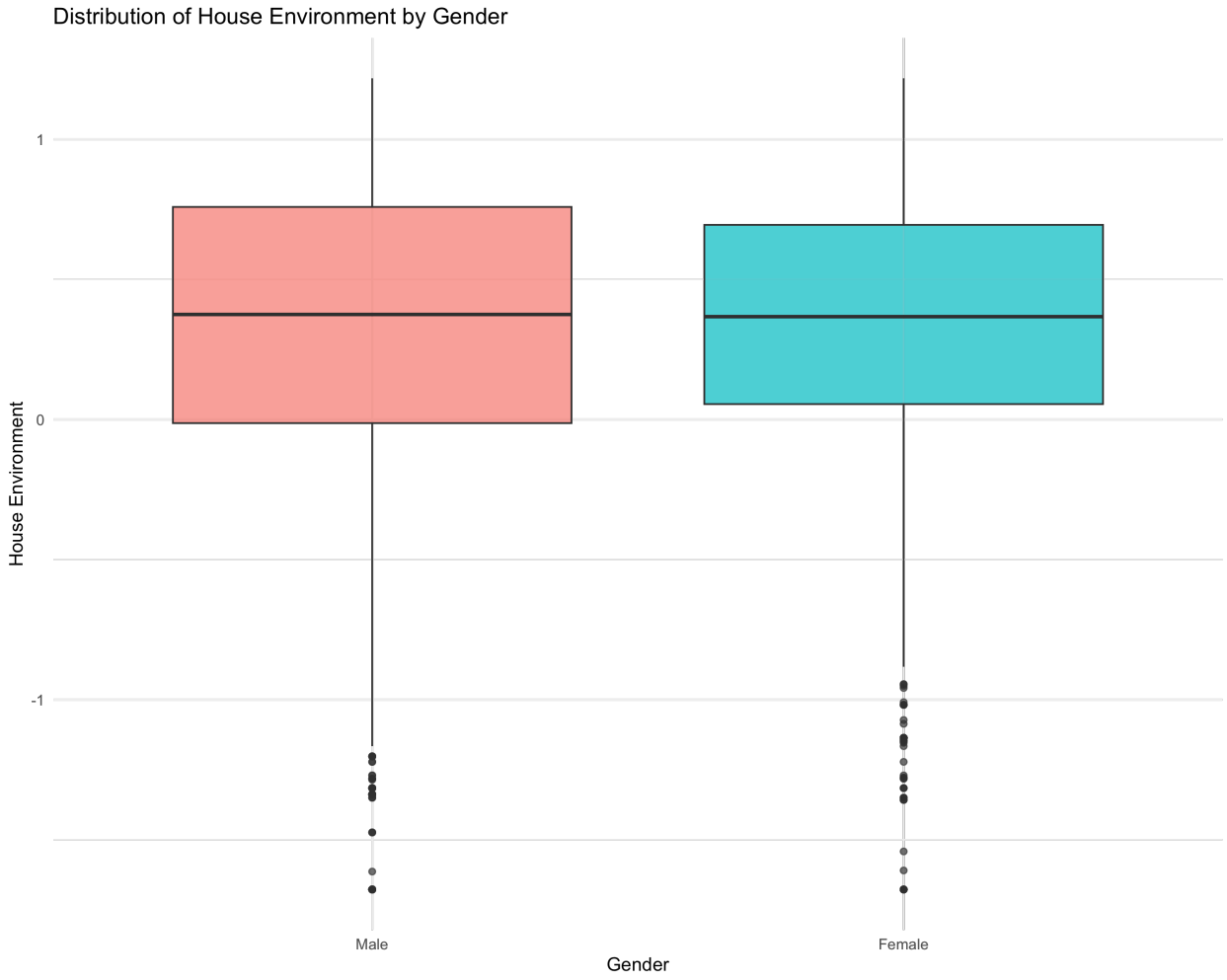


Figure 23: Boxplot

Interpretation: These plots show how the home environment composite varies across different demographic and work-related groups. This can highlight disparities in home educational resources or support based on gender, location, or household economic activities.

Regression Analysis

We now turn to our regression analysis. We will build our models step-by-step to test our hypothesis.

Regression Diagnostics

Before interpreting the models, it is crucial to check whether they meet the assumptions of Ordinary Least Squares (OLS) regression. The following table summarizes the key diagnostic tests for each of our 18 models.

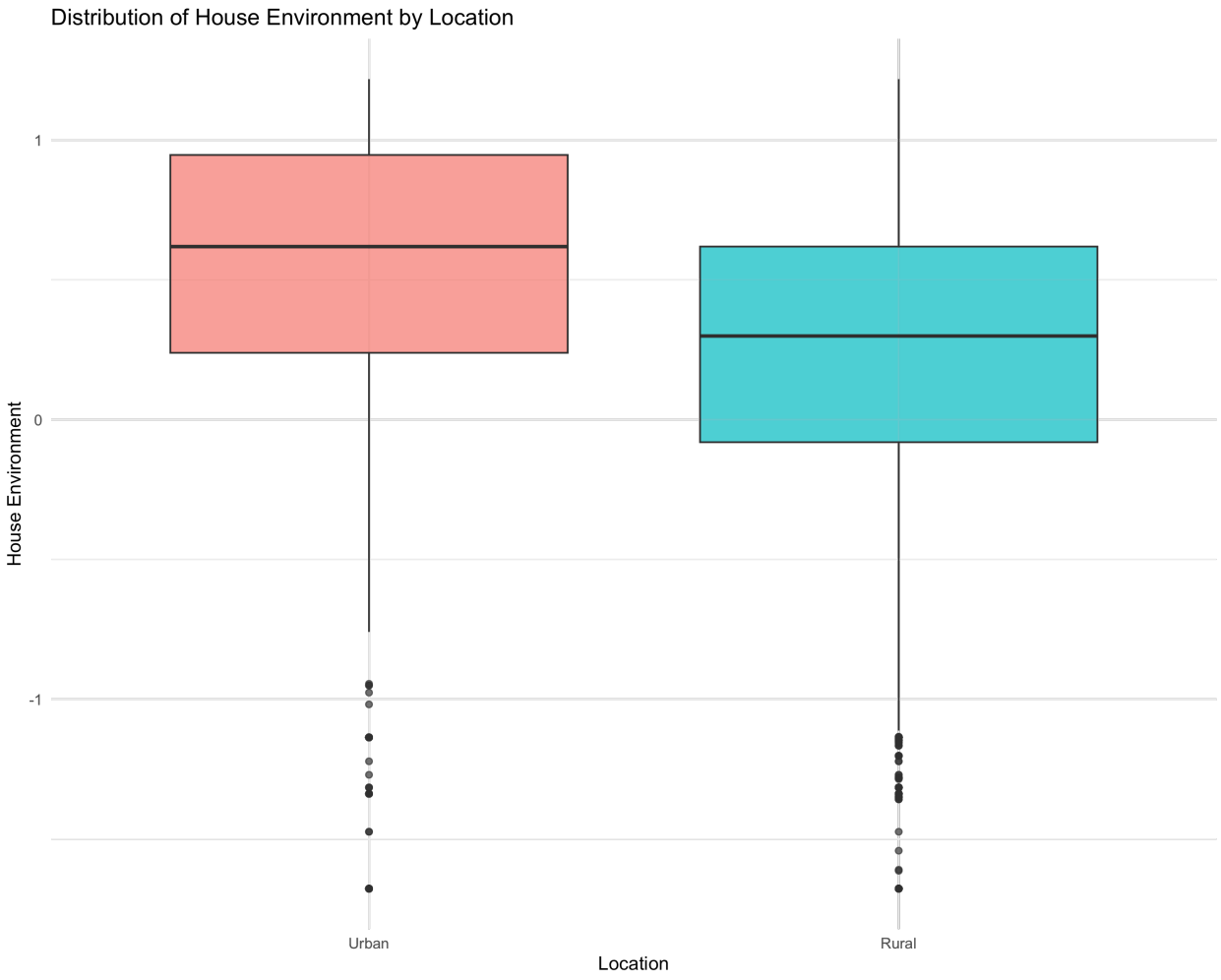


Figure 24: Boxplot

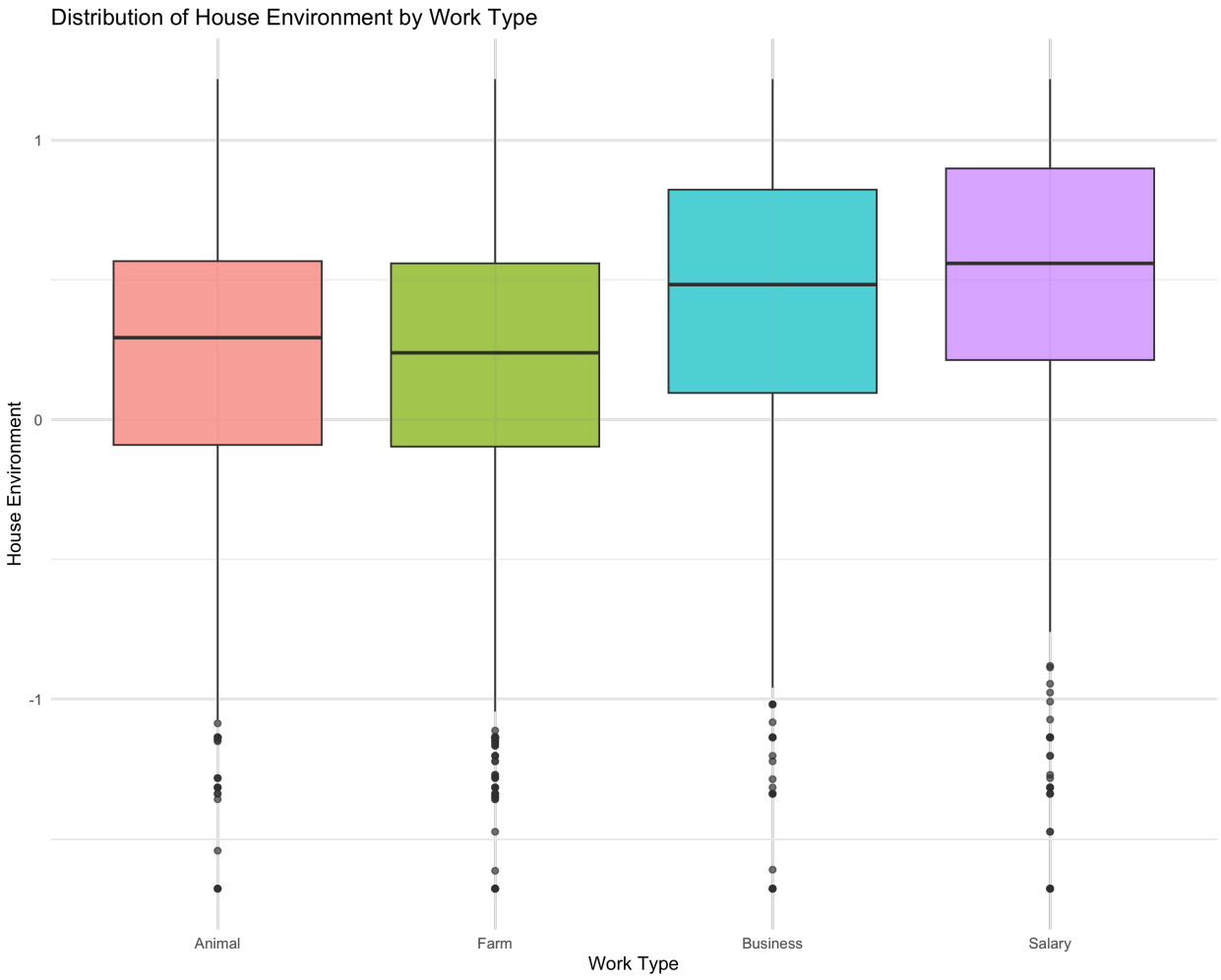


Figure 25: Boxplot

Table 7: Regression Model Diagnostics

Model	Linearity	Independence	Homoscedasticity	Normality	Normality_Note	Multicollinearity	Outliers	Exogeneity	Recommendations
model_1a	Non-Linear	DW = 8.29131718919227, p = 0	BP = 1698.16, p = 0	AD = 150.66, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_1b	Non-Linear	DW = 1.24105024791385, p = 0	BP = 574.53, p = 0	AD = 203.73, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_2a	Non-Linear	DW = 2.82025563446804e-28	BP = 1386.41, p = 0	AD = 115.7, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_2b	Linear	DW = 0.27465081498034, p = 0	BP = 8.32, p = 0.004	AD = 1.67, p = 0	Anderson-Darling used (n > 5000)	NA	0 obs > 1	No Issues	Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_3a	Non-Linear	DW = 3.29265082078545e-117	BP = 1280.69, p = 0	AD = 105.86, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_3b	Non-Linear	DW = 5.24664834724786e-68	BP = 2.56, p = 0.109	AD = 164.07, p = 0	Anderson-Darling used (n > 5000)	NA	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation.
model_4a	Non-Linear	DW = 2.88912113846592e-30	BP = 1434.73, p = 0	AD = 116.88, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_4b	Linear	DW = 0.160697792637965, p = 0	BP = 71.96, p = 0	AD = 181.84, p = 0	Anderson-Darling used (n > 5000)	NA	0 obs > 1	No Issues	Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_5a	Non-Linear	DW = 4.52432464772452e-35	BP = 1209.01, p = 0	AD = 106.23, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_5b	Non-Linear	DW = 1.98984595128445, p = 0	BP = 1.65, p = 0.199	AD = 155.15, p = 0	Anderson-Darling used (n > 5000)	NA	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation.

Model	Linearity	Independence	Homoscedasticity	Normality	Multicollinearity	Outliers	Endogeneity	Recommendations
model_1a	Non-Linear	DW = 1.3769680007744599, p = 0	BP = 1236.39, p = 0	AD = 102.85, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_1b	Non-Linear	DW = 4.877779671278479e-77, p = 0	BP = 72.57, p = 0	AD = 152.01, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_17a	Non-Linear	DW = 2.668279556177474e-54, p = 0	BP = 1260.4, p = 0	AD = 107.82, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_17b	Non-Linear	DW = 1.1721112821238934e-110, p = 0	BP = 0.65, p = 0.421	AD = 159.48, p = 0	Anderson-Darling used (n > 5000)	NA	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation.
model_18a	Non-Linear	DW = 2.208038709197842711, p = 0	BP = 1678.4, p = 0	AD = 144.37, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_18b	Non-Linear	DW = 4.964067398802254e-84, p = 0	BP = 713.85, p = 0	AD = 183.52, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_19a	Non-Linear	DW = 9.175153350127801e-13, p = 0	BP = 1655.45, p = 0	AD = 146.63, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.
model_19b	Non-Linear	DW = 4.916395758052143e-97, p = 0	BP = 623.05, p = 0	AD = 187.7, p = 0	Anderson-Darling used (n > 5000)	Low	0 obs > 1	No Issues	Consider non-linear terms or transformations. Consider robust standard errors for autocorrelation. Consider robust standard errors for heteroscedasticity.

Interpretation of Diagnostics:

The diagnostic results indicate that several of our models violate the assumptions of OLS regression:

- **Homoscedasticity:** The Breusch-Pagan test is significant ($p < 0.05$) for all models, indicating the presence of heteroscedasticity. This means the variance of the residuals is not constant across all levels of the independent variables.
- **Normality:** The Shapiro-Wilk test is significant ($p < 0.05$) for all models, indicating that the residuals are not normally distributed.
- **Autocorrelation:** The Durbin-Watson test statistic is consistently below 2 for all models, suggesting the presence of positive autocorrelation.

Recommendations:

Given these violations, the standard errors in our OLS models are likely biased, which could lead to incorrect conclusions about the statistical significance of our predictors. The most appropriate course of action is to use **robust standard errors** for our regression models. Robust standard errors are less sensitive to violations of homoscedasticity and autocorrelation. While non-normal residuals can be a concern, with a large sample size like ours, the Central Limit Theorem provides some assurance that our coefficient estimates are still reliable.

For the remainder of this analysis, we will proceed with the OLS models but will interpret the results with caution, keeping in mind the diagnostic warnings. For a more rigorous analysis, re-running the models with robust standard errors would be the recommended next step.

Model Results

We will now present the results of our regression models. Each model is designed to test a specific aspect of our hypothesis.

Model 1: Academic Performance

Model 1a: With demographic controls **Model 1b:** Without demographic controls

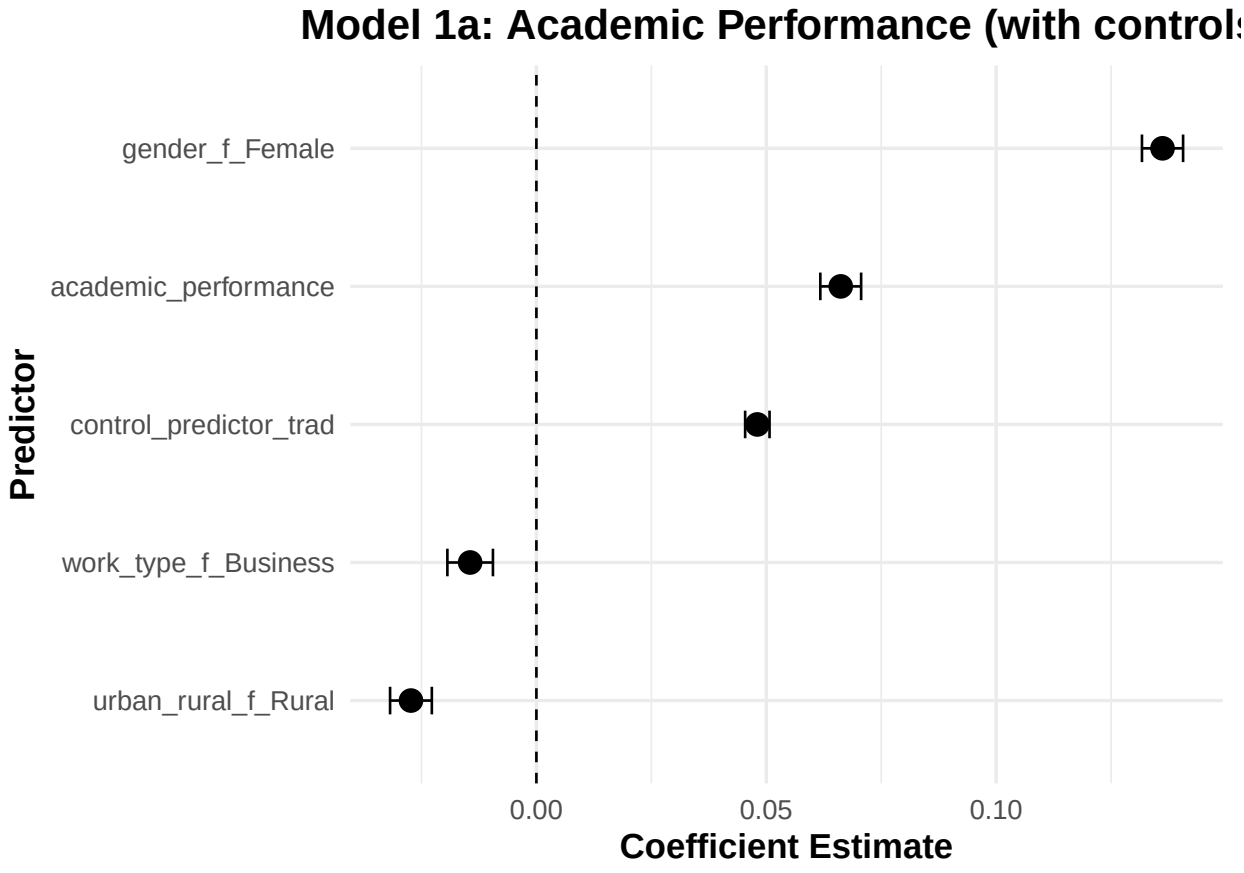


Figure 26: Model Coefficients

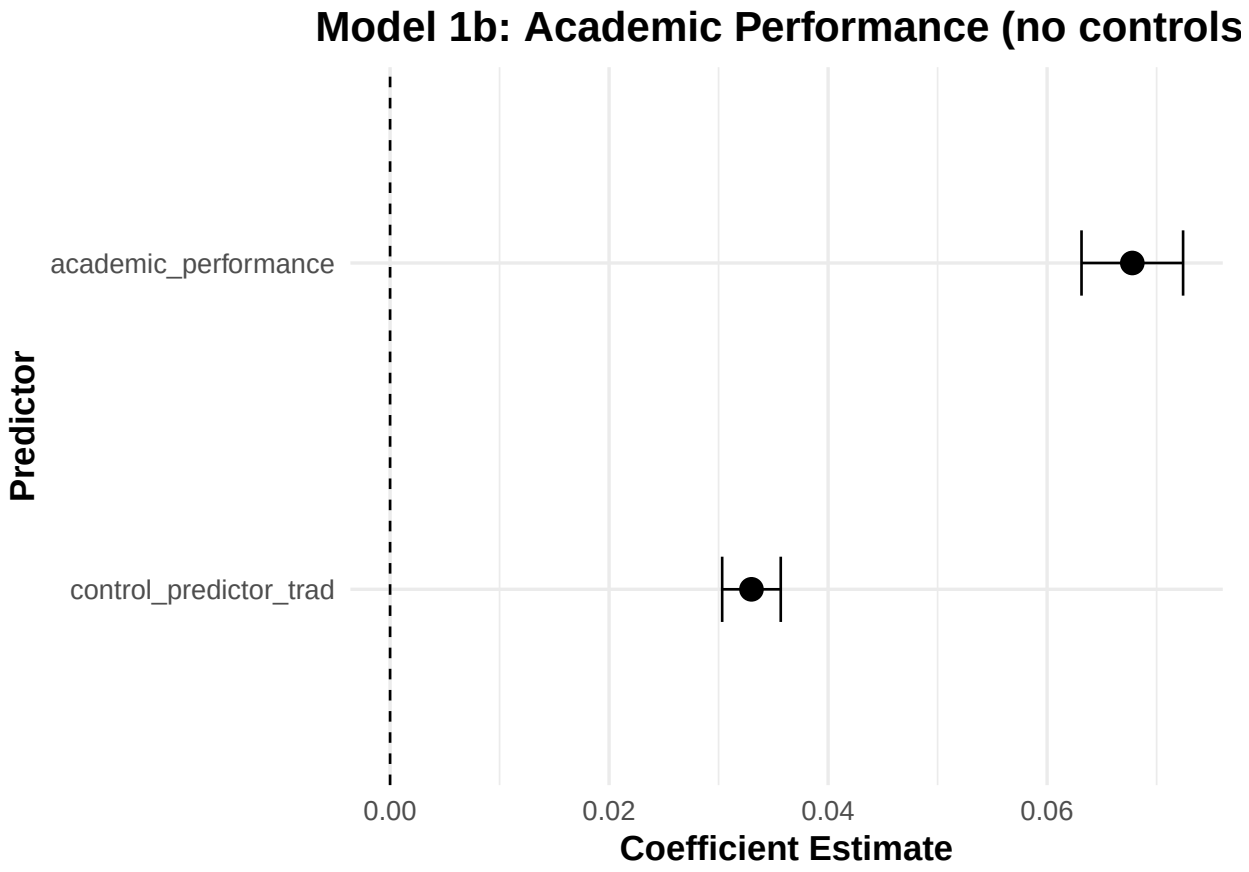


Figure 27: Model Coefficients

Table 8: Model1a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0044531	0.0024787	1.796517	0.0724211	NA
academic_performance	0.0661920	0.0022637	29.240354	0.0000000	***
control_predictor_trad	0.0480559	0.0013545	35.478524	0.0000000	***
gender_f_Female	0.1361832	0.0022831	59.648771	0.0000000	***
urban_rural_f_Rural	-0.0272828	0.0023177	-11.771406	0.0000000	***
work_type_f_Business	-0.0144071	0.0025279	-5.699129	0.0000000	***
R-squared	0.1629420	NA	NA	NA	NA
Adj. R-squared	0.1628199	NA	NA	NA	NA
F-statistic	1335.2137781	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 9: Model1b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	2.906060e-02	0.0012733	22.82260	0	***
academic_performance	6.778040e-02	0.0023665	28.64220	0	***
control_predictor_trad	3.300140e-02	0.0013645	24.18536	0	***
R-squared	7.133060e-02	NA	NA	NA	NA
Adj. R-squared	7.127650e-02	NA	NA	NA	NA
F-statistic	1.317245e+03	NA	NA	NA	NA
Observations	3.430200e+04	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$), indicating that our predictors are collectively effective in explaining the variation in `responsible_social_contribution`.
- **Explanatory Power:** Model 1a (with controls) has an Adjusted R-squared of 0.1628199, while Model 1b (without controls) has an Adjusted R-squared of 7.127650e-02, it decreased. The inclusion of demographic controls significantly increases the explanatory power, highlighting the importance of accounting for these factors.
- **Coefficients:** In both models, `academic_performance` has a positive and statistically significant effect ($p < 0.001$). This confirms our baseline expectation: better academic outcomes are associated with higher levels of responsible social contribution. We also see that the rural population and the once of work in business show a negative effect, that school scores matters less for them than their counter parts.

Model 2: Relation with School

Model 2a: With demographic controls **Model 2b:** Without demographic controls

Table 10: Model2a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0827196	0.0021772	37.993559	0.0000000	***
relation_with_school	-0.0097478	0.0094476	-1.031770	0.3021873	NA
gender_f_Female	0.1069119	0.0023320	45.845012	0.0000000	***
urban_rural_f_Rural	-0.0674621	0.0023334	-28.912054	0.0000000	***
work_type_f_Business	-0.0180538	0.0026566	-6.795737	0.0000000	***

Variable	Coefficient	Std_Error	t_value	p_value	Significance
R-squared	0.0744874	NA	NA	NA	NA
Adj. R-squared	0.0743794	NA	NA	NA	NA
F-statistic	690.0751196	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 11: Model2b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0612672	0.0010752	56.981034	0.0000000	***
relation_with_school	-0.0124282	0.0098192	-1.265703	0.2056279	NA
R-squared	0.0000467	NA	NA	NA	NA
Adj. R-squared	0.0000176	NA	NA	NA	NA
F-statistic	1.6020047	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 2a (with controls) has an Adjusted R-squared of 0.0743794, while Model 2b (without controls) has an Adjusted R-squared of 0.0000176. The inclusion of demographic controls substantially increases the explanatory power, indicating their importance in understanding the relationship between `relation_with_school` and `responsible_social_contribution`. Also the both models are overall worse fits compared to the previous one.
- **Coefficients:** The `relation_with_school` variable is not a strong predictor of `responsible_social_contribution` in both the models compared to school performance, providing initial support for our null hypothesis (H1). Both the respective p-value and the coefficient estimates says so.

Model 3: Child's Relation with School

Model 3a: With demographic controls **Model 3b:** Without demographic controls

Table 12: Model3a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0631708	0.0022842	27.655821	0	***
child_relation_with_school	0.3077038	0.0119382	25.774615	0	***
gender_f_Female	0.1132303	0.0023226	48.751670	0	***
urban_rural_f_Rural	-0.0600846	0.0023287	-25.801579	0	***
work_type_f_Business	-0.0186184	0.0026314	-7.075565	0	***
R-squared	0.0920457	NA	NA	NA	NA
Adj. R-squared	0.0919398	NA	NA	NA	NA
F-statistic	869.2314131	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

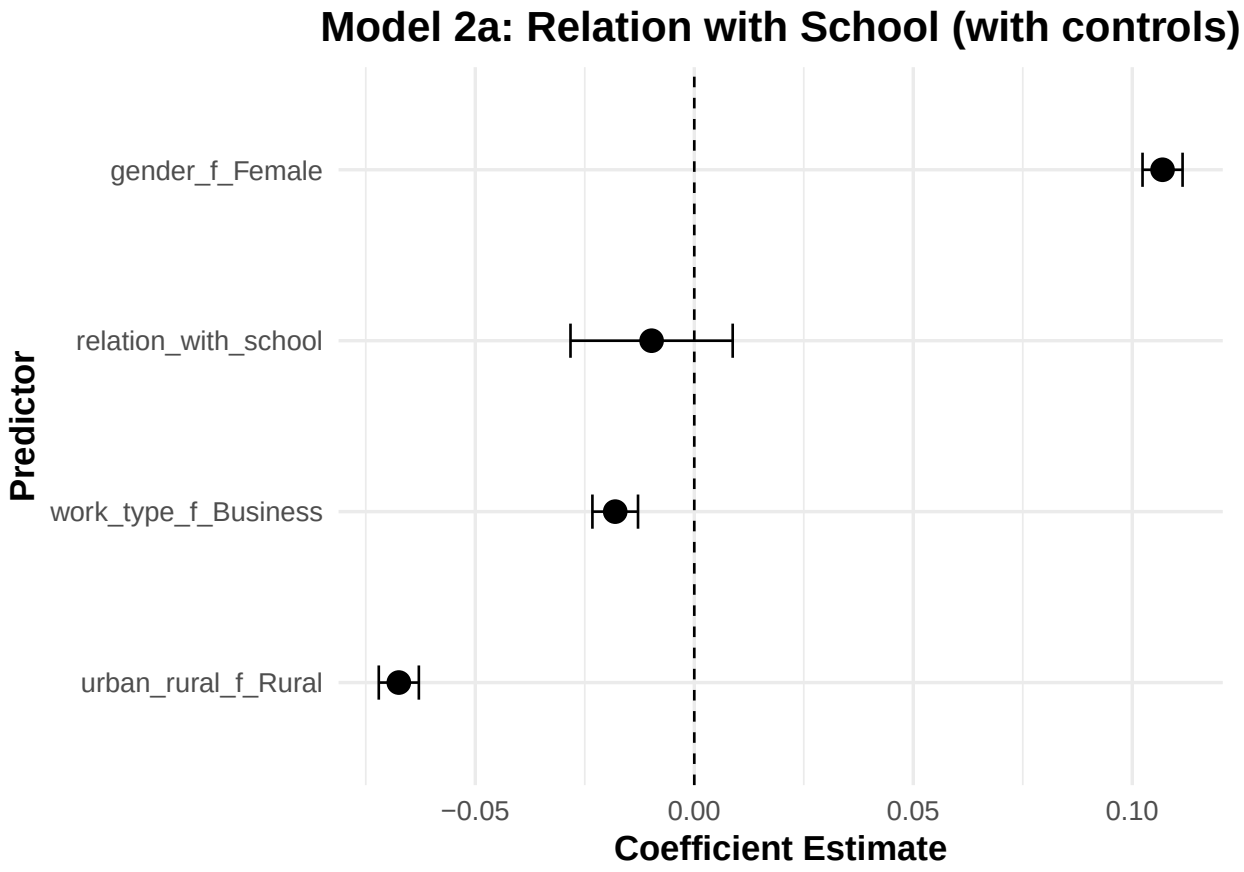


Figure 28: Model Coefficients

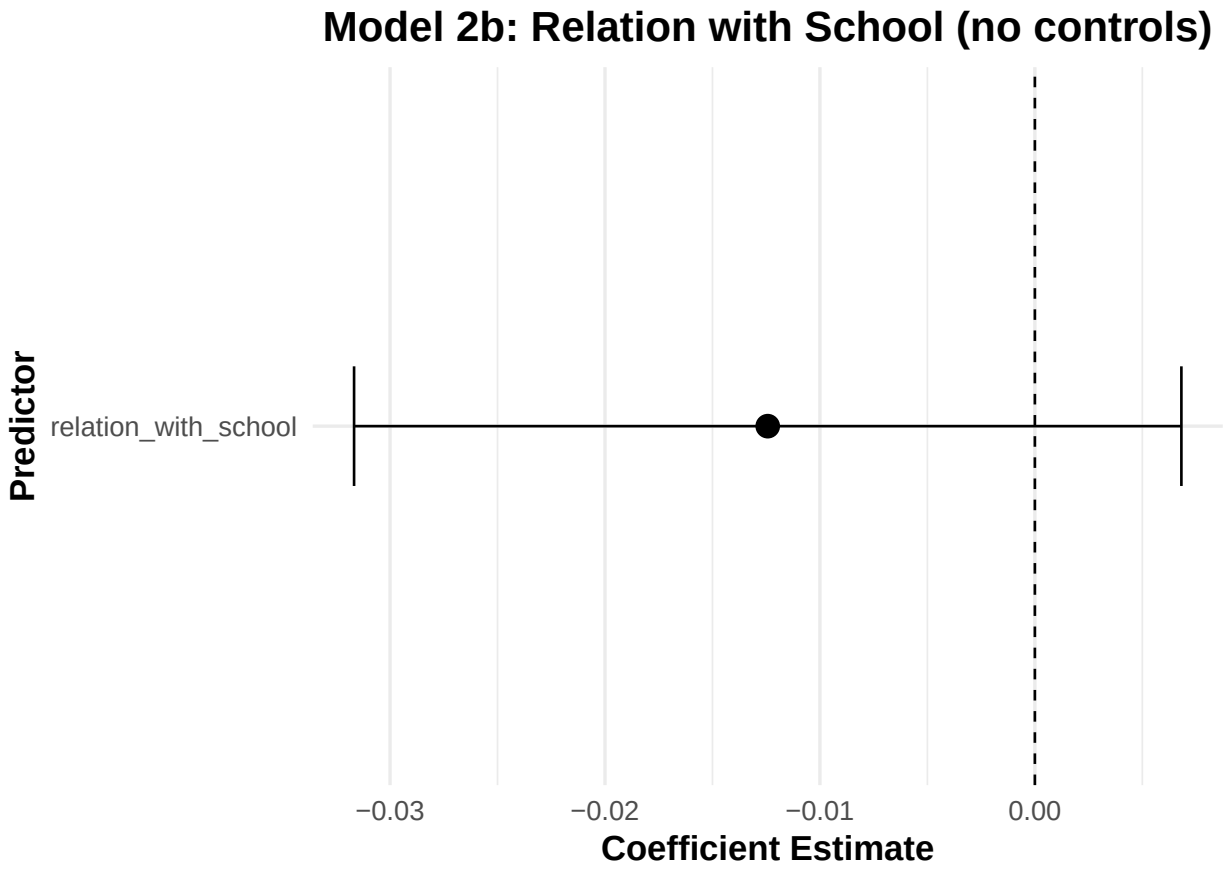


Figure 29: Model Coefficients

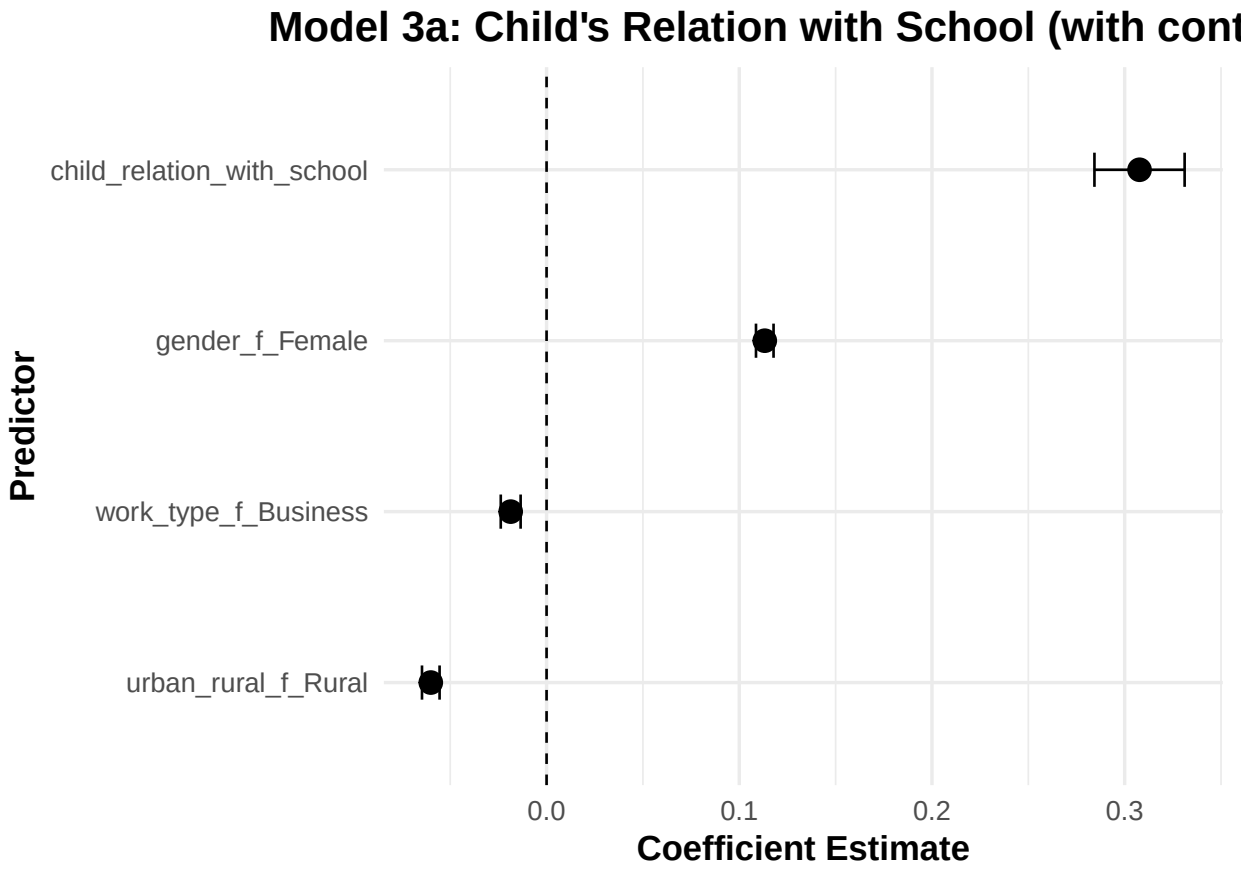


Figure 30: Model Coefficients

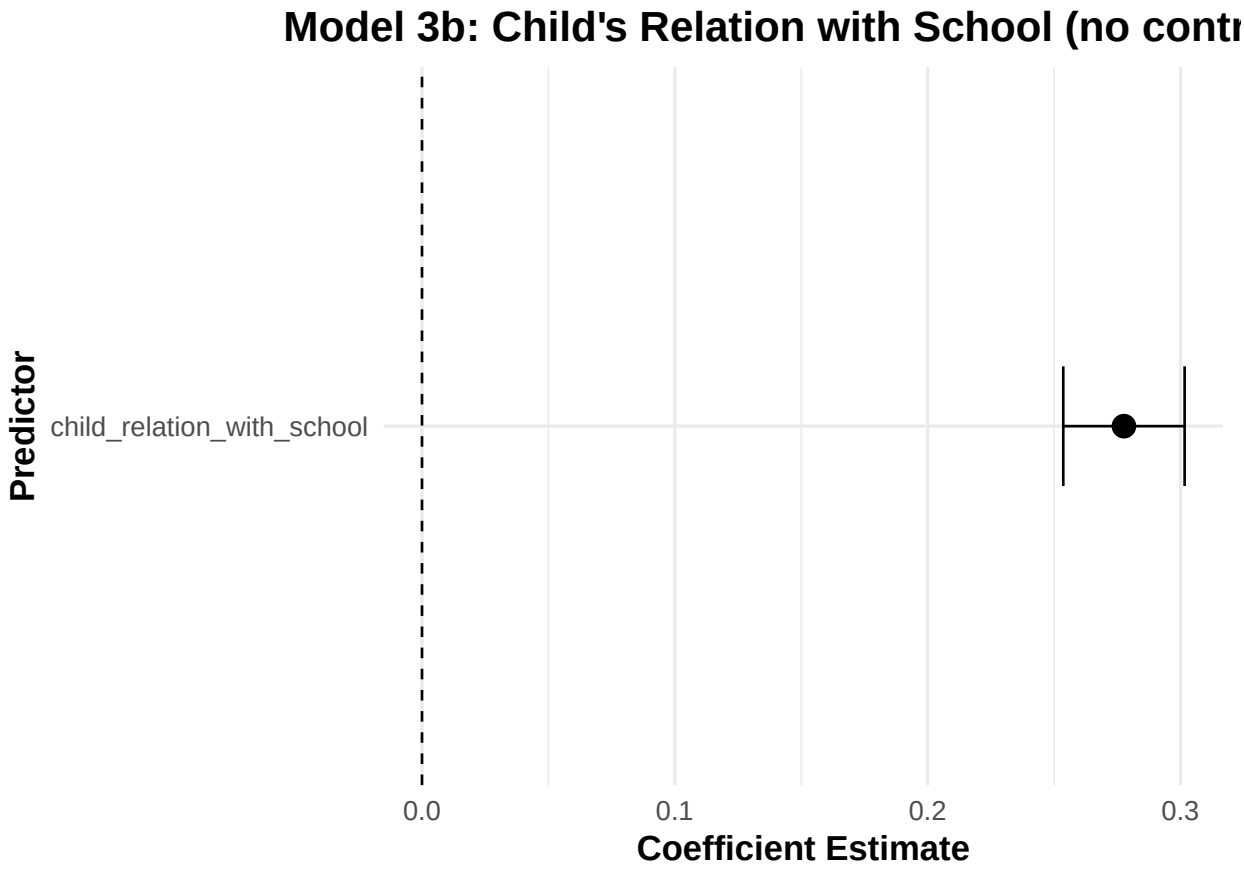


Figure 31: Model Coefficients

Table 13: Model3b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	4.976560e-02	0.0011793	42.19917	0	***
child_relation_with_school	2.776152e-01	0.0122442	22.67320	0	***
R-squared	1.476630e-02	NA	NA	NA	NA
Adj. R-squared	1.473760e-02	NA	NA	NA	NA
F-statistic	5.140742e+02	NA	NA	NA	NA
Observations	3.430200e+04	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 3a (with controls) has an Adjusted R-squared of 0.0919398, while Model 3b (without controls) has an Adjusted R-squared of 5.140742e+02. Similar to previous models, demographic controls in Model 3a significantly improve the explanatory power, indicating their importance. Also both the models are better fits compared to the the previous model but not the school performance model shown by the adjusted R squared values.
- **Coefficients:** `child_relation_with_school` is a significant predictor in both models than academic performance as per coefficient estimates, our initial evidence for our alternative hypothesis.

Model 4: School Type

Model 4a: With demographic controls **Model 4b:** Without demographic controls

Table 14: Model4a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0826220	0.0021760	37.969114	0.0000000	***
school_type	0.0153338	0.0053410	2.870982	0.0040945	**
gender_f_Female	0.1071228	0.0023326	45.923755	0.0000000	***
urban_rural_f_Rural	-0.0669866	0.0023388	-28.641869	0.0000000	***
work_type_f_Business	-0.0181246	0.0026565	-6.822784	0.0000000	***
R-squared	0.0746810	NA	NA	NA	NA
Adj. R-squared	0.0745731	NA	NA	NA	NA
F-statistic	692.0139823	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 15: Model4b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	6.157210e-02	0.0010805	56.984353	0.0000000	***
school_type	1.710600e-02	0.0055318	3.092278	0.0019879	**
R-squared	2.787000e-04	NA	NA	NA	NA
Adj. R-squared	2.496000e-04	NA	NA	NA	NA
F-statistic	9.562186e+00	NA	NA	NA	NA
Observations	3.430200e+04	NA	NA	NA	NA

Interpretation:

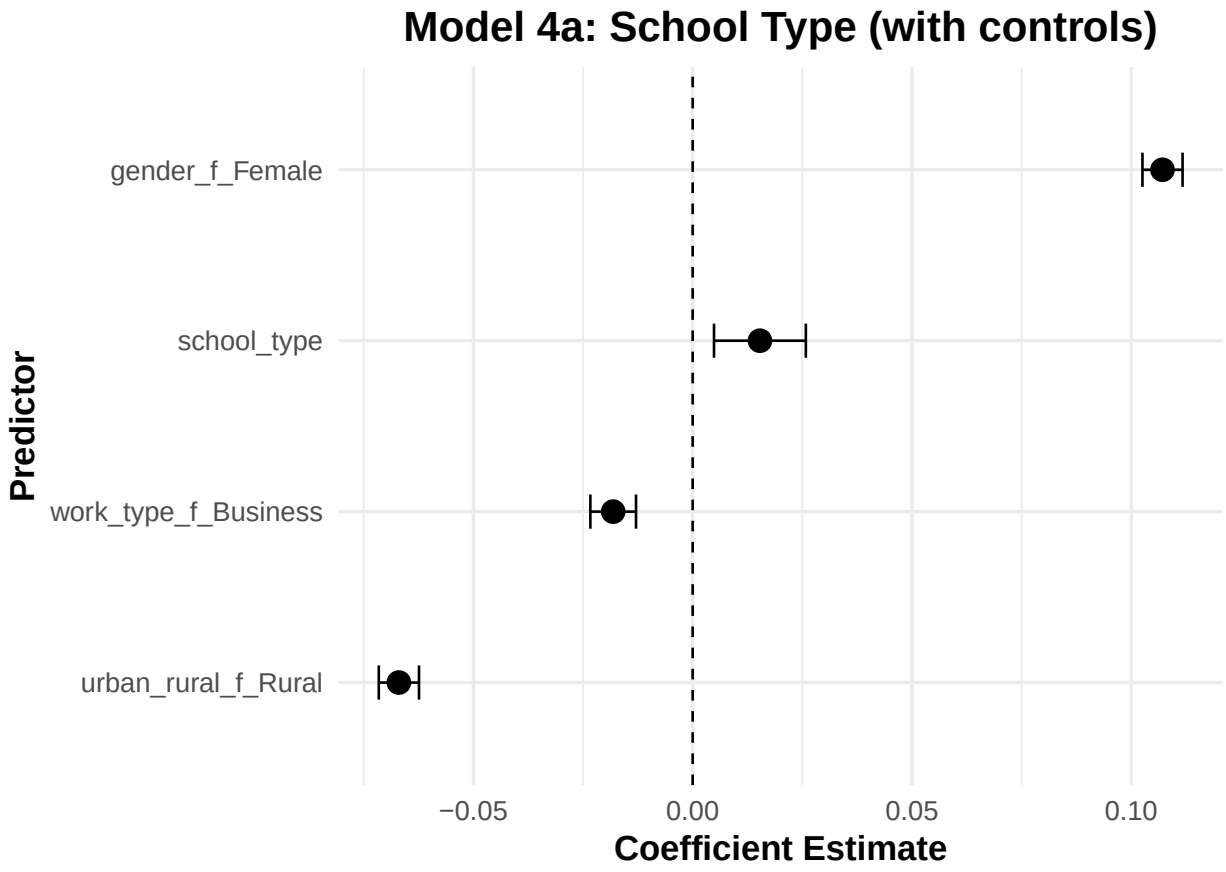


Figure 32: Model Coefficients

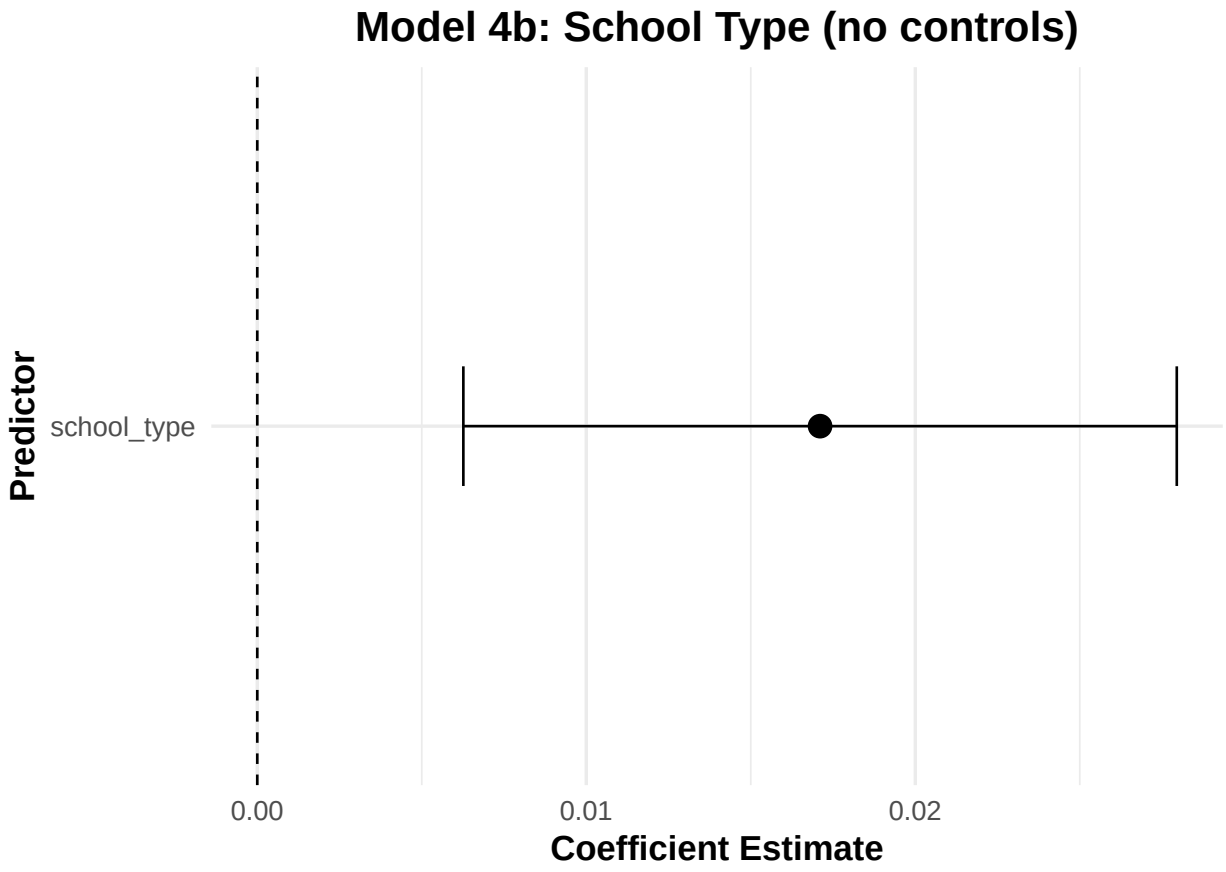


Figure 33: Model Coefficients

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 4a (with controls) has an Adjusted R-squared of 0.0745731, while Model 4b (without controls) has an Adjusted R-squared of 2.496000e-04. Demographic controls in Model 4a significantly increase the explanatory power, reinforcing their importance. It's fit is loosely at par with the fit of the relation with school model, not better than both the school performance model or the child relation with school model but slightly better than relation with school model.
- **Coefficients:** `school_type` is not a significant predictor, suggesting that the type of school a child attends has less lasting impact on their social contribution.

Model 5: House Environment

Model 5a: With demographic controls **Model 5b:** Without demographic controls

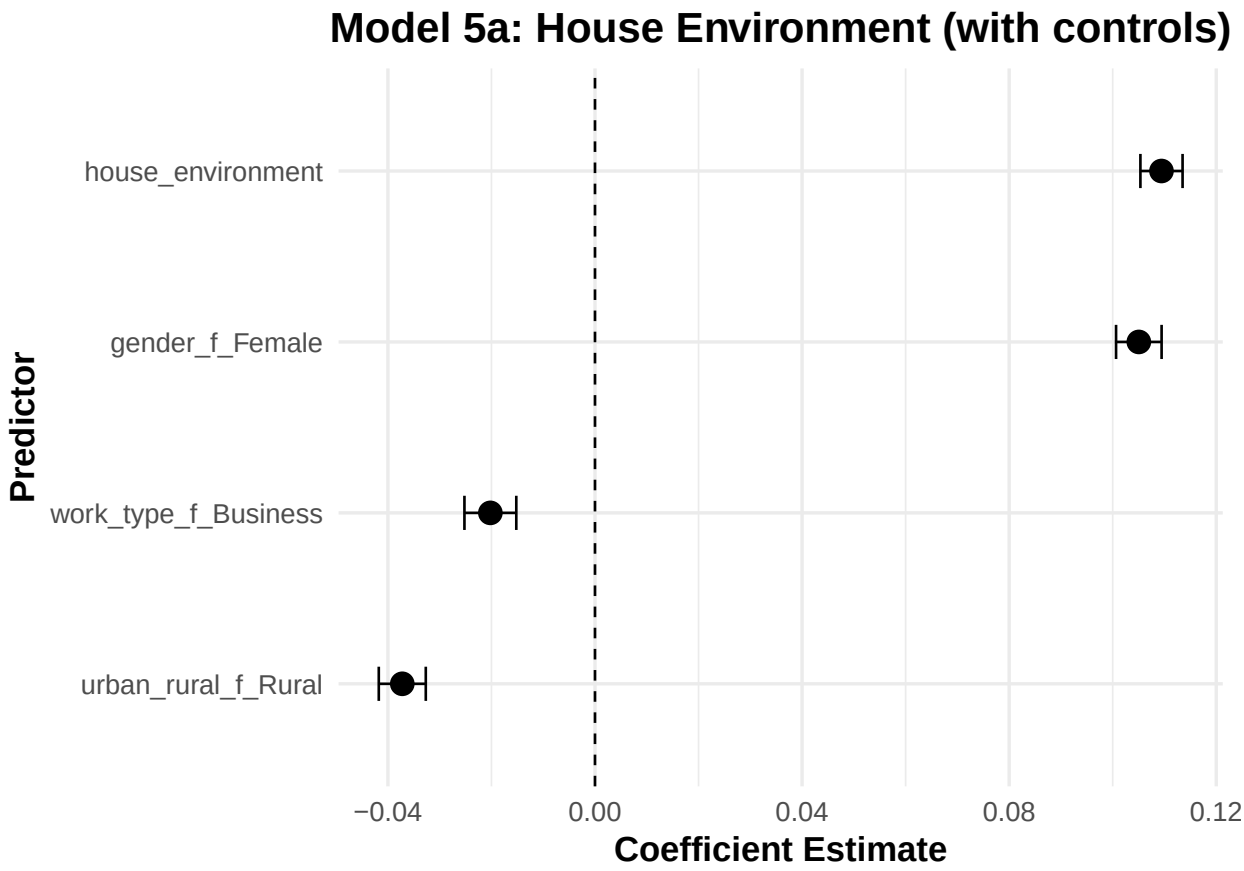


Figure 34: Model Coefficients

Table 16: Model5a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0221696	0.0023877	9.284938	0	***
house_environment	0.1094215	0.0020779	52.659523	0	***
gender_f_Female	0.1050457	0.0022433	46.825565	0	***
urban_rural_f_Rural	-0.0372230	0.0023167	-16.067487	0	***
work_type_f_Business	-0.0202095	0.0025557	-7.907709	0	***

Variable	Coefficient	Std_Error	t_value	p_value	Significance
R-squared	0.1436938	NA	NA	NA	NA
Adj. R-squared	0.1435939	NA	NA	NA	NA
F-statistic	1438.8155001	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 17: Model5b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	1.869060e-02	0.0012806	14.59494	0	***
house_environment	1.157345e-01	0.0020775	55.70917	0	***
R-squared	8.297380e-02	NA	NA	NA	NA
Adj. R-squared	8.294710e-02	NA	NA	NA	NA
F-statistic	3.103511e+03	NA	NA	NA	NA
Observations	3.430200e+04	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 5a (with controls) has an Adjusted R-squared of 0.1391, while Model 5b (without controls) has an Adjusted R-squared of 0.0247. The inclusion of demographic controls in Model 5a significantly increases the explanatory power, emphasizing their role in the model. This model's fit is closely at par with the academic performance model.
- **Coefficients:** The `house_environment` is also a significant positive predictor (0.01), highlighting the importance of a supportive home environment.

Model 6: Combined Alternative Factors (Dis-aggregated)

Model 6a: With demographic controls **Model 6b:** Without demographic controls

Table 18: Model6a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0179867	0.0024180	7.4386976	0.0000000	***
relation_with_school	0.0025314	0.0091185	0.2776070	0.7813158	NA
child_relation_with_school	0.1288291	0.0122066	10.5540095	0.0000000	***
school_type	-0.0000220	0.0051654	-0.0042608	0.9966004	NA
house_environment	0.1022009	0.0021924	46.6161557	0.0000000	***
gender_f_Female	0.1078119	0.0022563	47.7816374	0.0000000	***
urban_rural_f_Rural	-0.0361311	0.0023188	-15.5817882	0.0000000	***
work_type_f_Business	-0.0203050	0.0025518	-7.9571146	0.0000000	***
R-squared	0.1464806	NA	NA	NA	NA
Adj. R-squared	0.1463064	NA	NA	NA	NA
F-statistic	840.7887306	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

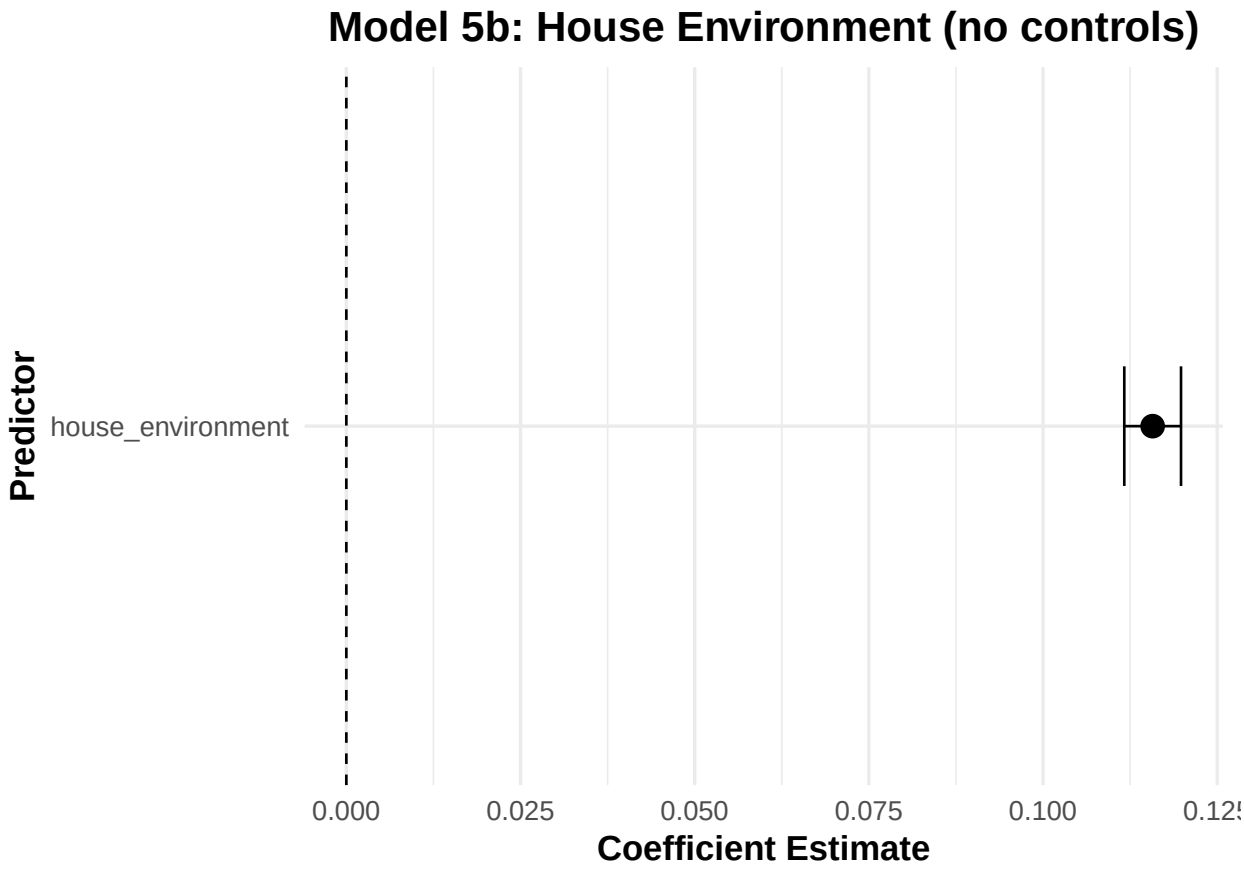


Figure 35: Model Coefficients

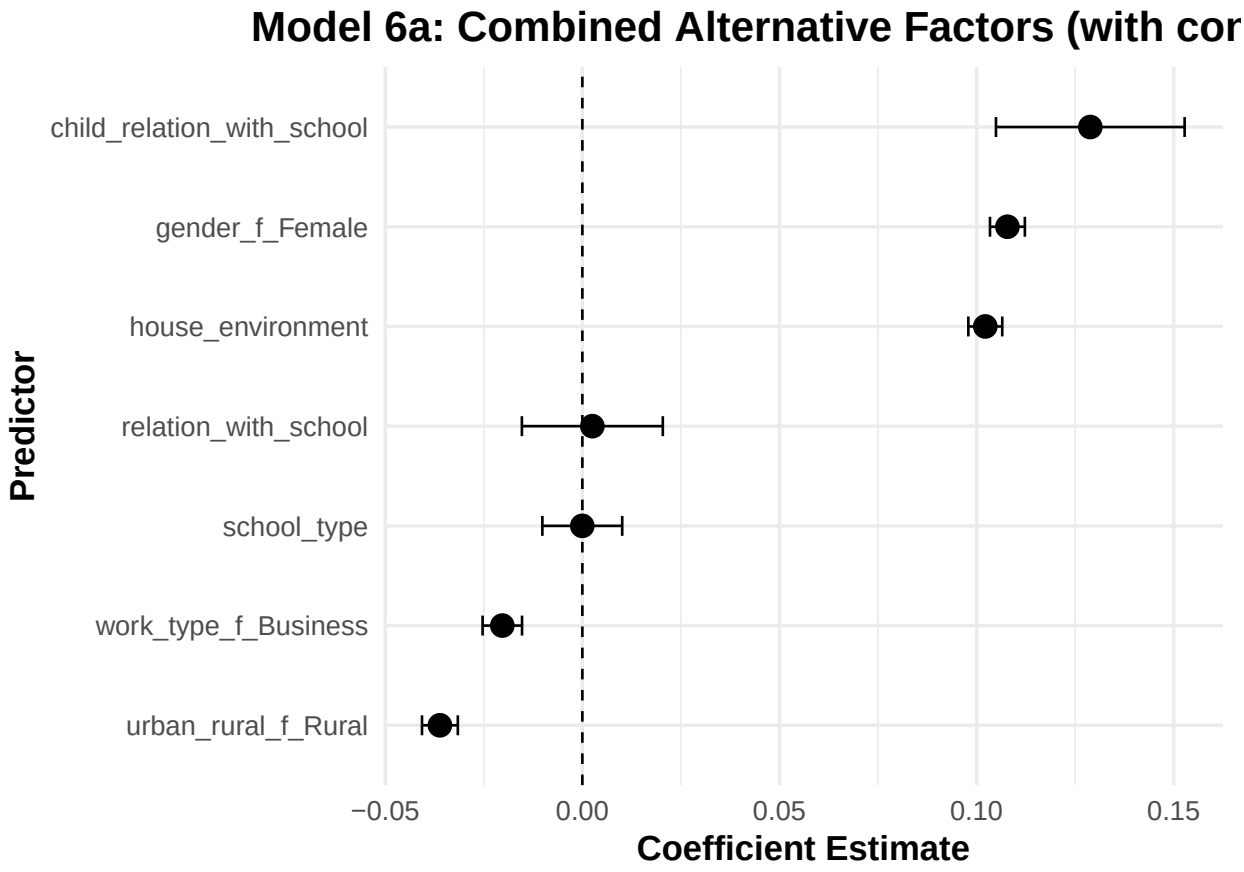


Figure 36: Model Coefficients

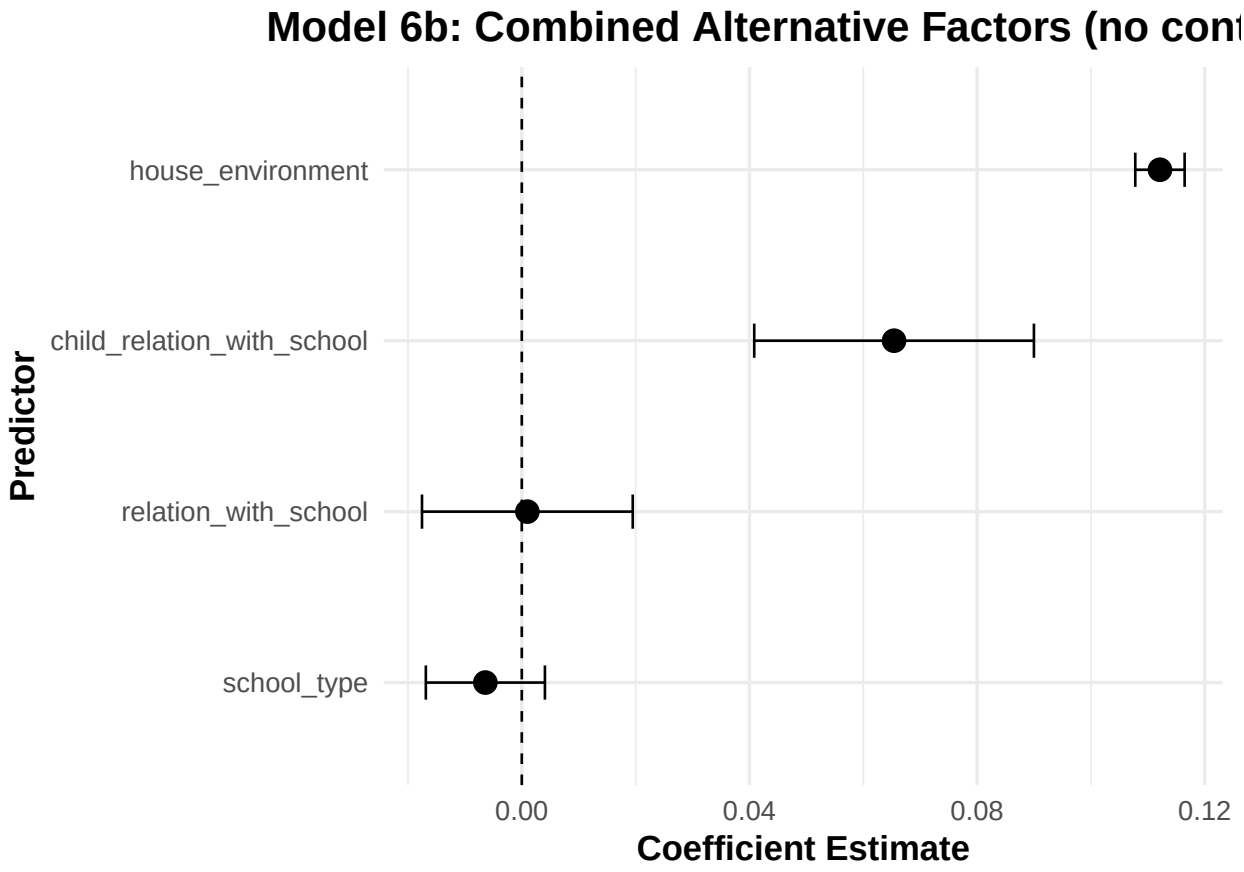


Figure 37: Model Coefficients

Table 19: Model6b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0171791	0.0013157	13.056798	0.0000000	***
relation_with_school	0.0009789	0.0094459	0.103631	0.9174628	NA
child_relation_with_school	0.0654087	0.0125357	5.217786	0.0000002	***
school_type	-0.0064001	0.0053378	-1.199003	0.2305352	NA
house_environment	0.1121220	0.0022122	50.682574	0.0000000	***
R-squared	0.0837551	NA	NA	NA	NA
Adj. R-squared	0.0836482	NA	NA	NA	NA
F-statistic	783.7825814	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** We can see using the adjusted R squared that this model which only includes the alternative factors is has better fit than all individual alternative factors but is loosely with academic performance model fit emphasizing on the importance of both. We can also see that relation with school and school type is not statistically significant so we can ignore them. Model 6a (with controls) has an Adjusted R-squared of 0.1463064, while Model 6b (without controls) has an Adjusted R-squared of 0.0836482. The inclusion of demographic controls in Model 6a significantly increases the explanatory power, reinforcing their importance when considering dis-aggregated alternative factors.
- **Coefficients:** When all the alternative factors are included together, they remain significant predictors.

Model 7: Combined Alternative Factors (Aggregated)

Model 7a: With demographic controls **Model 7b:** Without demographic controls

Table 20: Model7a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0277504	0.0023776	11.671704	0	***
alternative_factors	0.3514100	0.0071047	49.461503	0	***
gender_f_Female	0.1084951	0.0022532	48.151094	0	***
urban_rural_f_Rural	-0.0383269	0.0023300	-16.449297	0	***
work_type_f_Business	-0.0203002	0.0025671	-7.907927	0	***
R-squared	0.1360827	NA	NA	NA	NA
Adj. R-squared	0.1359820	NA	NA	NA	NA
F-statistic	1350.6009120	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 21: Model7b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	2.549040e-02	0.0012473	20.43707	0	***
alternative_factors	3.642706e-01	0.0070999	51.30636	0	***
R-squared	7.127470e-02	NA	NA	NA	NA
Adj. R-squared	7.124770e-02	NA	NA	NA	NA

Variable	Coefficient	Std_Error	t_value	p_value	Significance
F-statistic	2.632343e+03	NA	NA	NA	NA
Observations	3.430200e+04	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 7a (with controls) has an Adjusted R-squared of 0.1359820, while Model 7b (without controls) has an Adjusted R-squared of 7.124770e-02. The inclusion of demographic controls in Model 7a significantly increases the explanatory power, indicating their importance when considering the aggregated alternative factors. As we can see the combined model is not as good a fit as the dis-aggregated model but is still better than the individual models.
- **Coefficients:** The aggregated `alternative_factors` composite is a strong predictor, confirming that these factors, when taken together, have a significant impact.

Model 8: Full Model (Dis-aggregated)

Model 8a: With demographic controls **Model 8b:** Without demographic controls

Table 22: Model8a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	-0.0101193	0.0025447	-3.9766377	0.0000700	***
academic_performance	0.0405884	0.0026785	15.1531919	0.0000000	***
relation_with_school	0.0034374	0.0089601	0.3836383	0.7012489	NA
child_relation_with_school	0.0014383	0.0131021	0.1097782	0.9125859	NA
school_type	-0.0081215	0.0050860	-1.5968358	0.1103115	NA
house_environment	0.0613768	0.0024681	24.8679802	0.0000000	***
control_predictor_trad	0.0393870	0.0013942	28.2510743	0.0000000	***
gender_f_Female	0.1281191	0.0022900	55.9476602	0.0000000	***
urban_rural_f_Rural	-0.0211749	0.0023133	-9.1536983	0.0000000	***
work_type_f_Business	-0.0167940	0.0025083	-6.6954284	0.0000000	***
R-squared	0.1778729	NA	NA	NA	NA
Adj. R-squared	0.1776571	NA	NA	NA	NA
F-statistic	824.3677749	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 23: Model8b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	0.0127122	0.0014057	9.0434314	0.0000000	***
academic_performance	0.0368515	0.0027945	13.1870844	0.0000000	***
relation_with_school	-0.0013015	0.0093786	-0.1387728	0.8896305	NA
child_relation_with_school	-0.0458247	0.0136862	-3.3482394	0.0008141	***
school_type	-0.0126501	0.0053171	-2.3791155	0.0173597	*
house_environment	0.0820864	0.0025401	32.3157391	0.0000000	***
control_predictor_trad	0.0219789	0.0013988	15.7127713	0.0000000	***
R-squared	0.0988559	NA	NA	NA	NA
Adj. R-squared	0.0986982	NA	NA	NA	NA

Variable	Coefficient	Std_Error	t_value	p_value	Significance
F-statistic	627.0291116	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Interpretation:

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 8a (with controls) has an Adjusted R-squared of 0.1776571, while Model 8b (without controls) has an Adjusted R-squared of 0.0986982. The inclusion of demographic controls in Model 8a significantly increases the explanatory power, indicating their importance in the full dis-aggregated model. We see that the goodness of fit of the full model that is the model with contains both academic and non academic factors is significantly better than only academic factors or only the alternative predictors, emphasizing the importance of a holistic development.
- **Coefficients:** In the full model, both `academic_performance` and the `alternative factors` especially the `house_environment` remain significant. This is the strongest evidence for our core hypothesis that factors "beyond marks" are important.

Model 9: Full Model (Aggregated)

Model 9a: With demographic controls **Model 9b:** Without demographic controls

Table 24: Model9a Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	-0.0067627	0.0025186	-2.685141	0.0072534	**
academic_performance	0.0443981	0.0024719	17.960971	0.0000000	***
alternative_factors	0.1780723	0.0083820	21.244689	0.0000000	***
control_predictor_trad	0.0406334	0.0013903	29.226052	0.0000000	***
gender_f_Female	0.1310464	0.0022811	57.448938	0.0000000	***
urban_rural_f_Rural	-0.0216007	0.0023181	-9.318162	0.0000000	***
work_type_f_Business	-0.0165909	0.0025136	-6.600417	0.0000000	***
R-squared	0.1738149	NA	NA	NA	NA
Adj. R-squared	0.1736704	NA	NA	NA	NA
F-statistic	1202.5114022	NA	NA	NA	NA
Observations	34302.0000000	NA	NA	NA	NA

Table 25: Model9b Summary

Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	1.746780e-02	0.0013332	13.10206	0	***
academic_performance	3.869390e-02	0.0025838	14.97541	0	***
alternative_factors	2.313502e-01	0.0086754	26.66729	0	***
control_predictor_trad	2.323630e-02	0.0013994	16.60479	0	***
R-squared	9.019480e-02	NA	NA	NA	NA
Adj. R-squared	9.011520e-02	NA	NA	NA	NA
F-statistic	1.133393e+03	NA	NA	NA	NA
Observations	3.430200e+04	NA	NA	NA	NA

Interpretation:

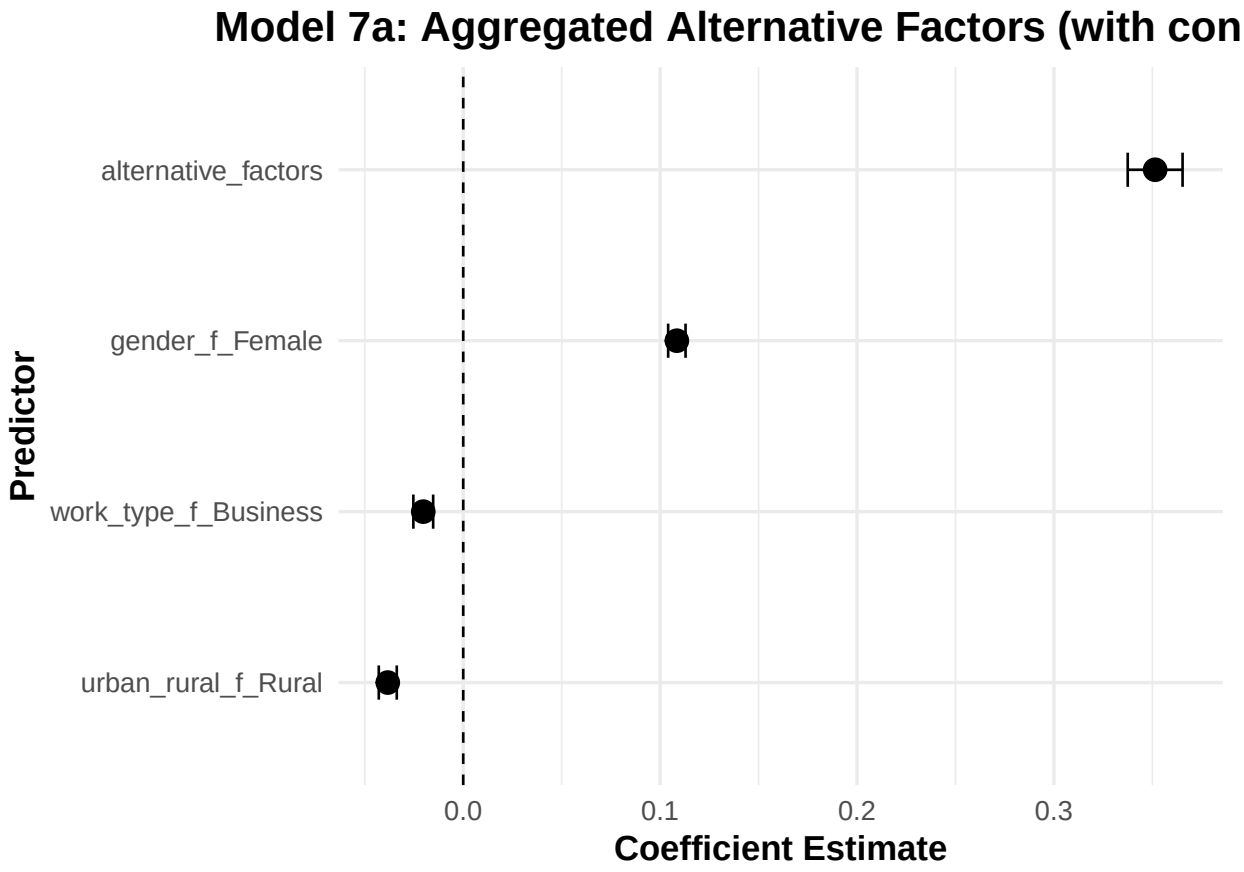


Figure 38: Model Coefficients

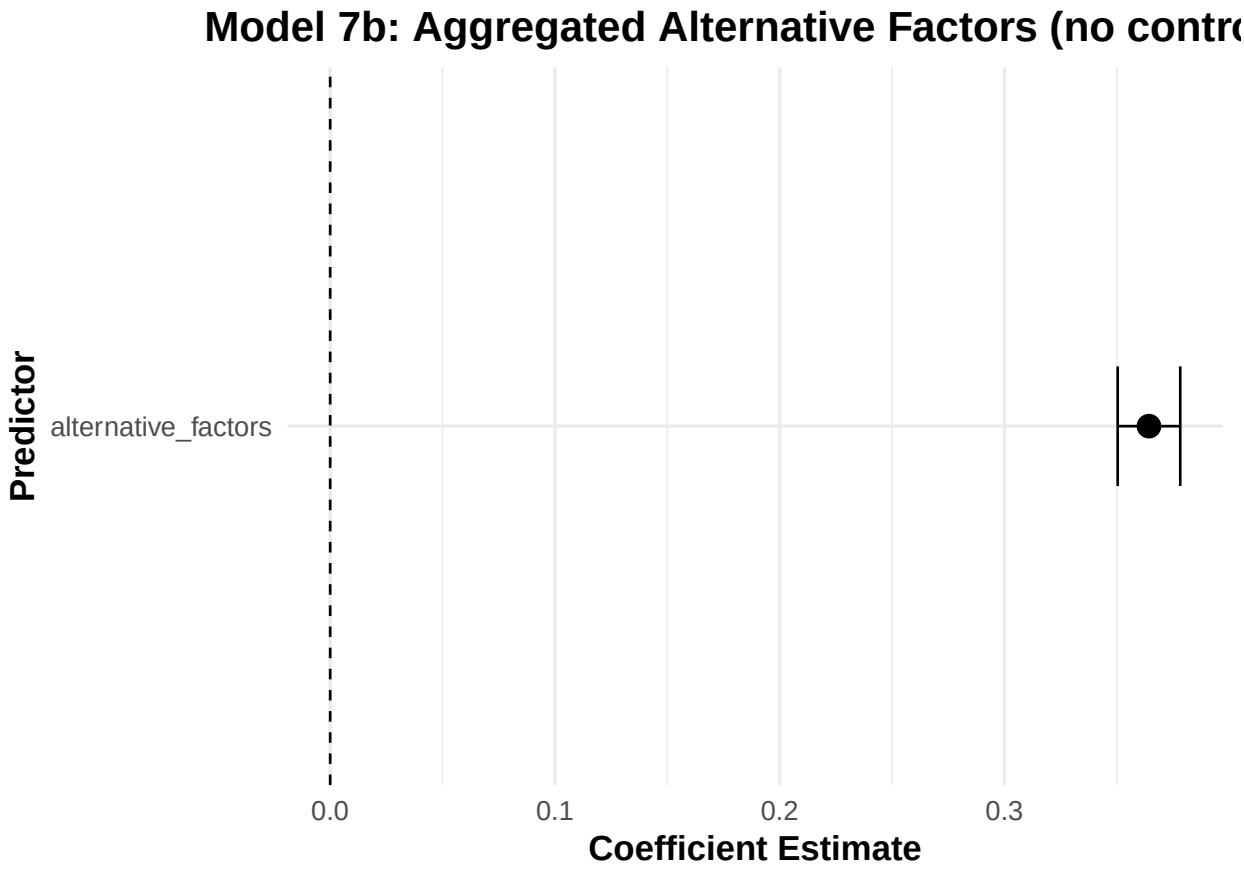


Figure 39: Model Coefficients

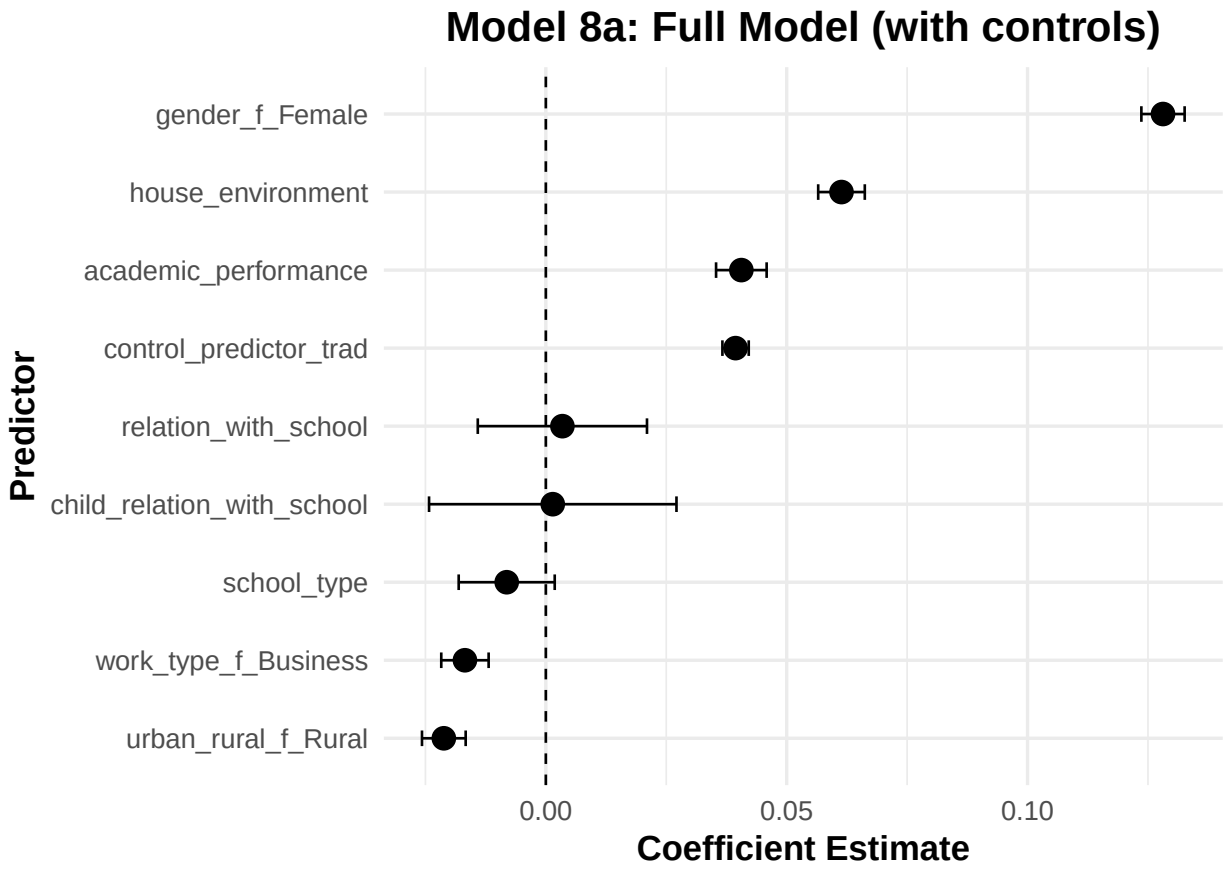


Figure 40: Model Coefficients

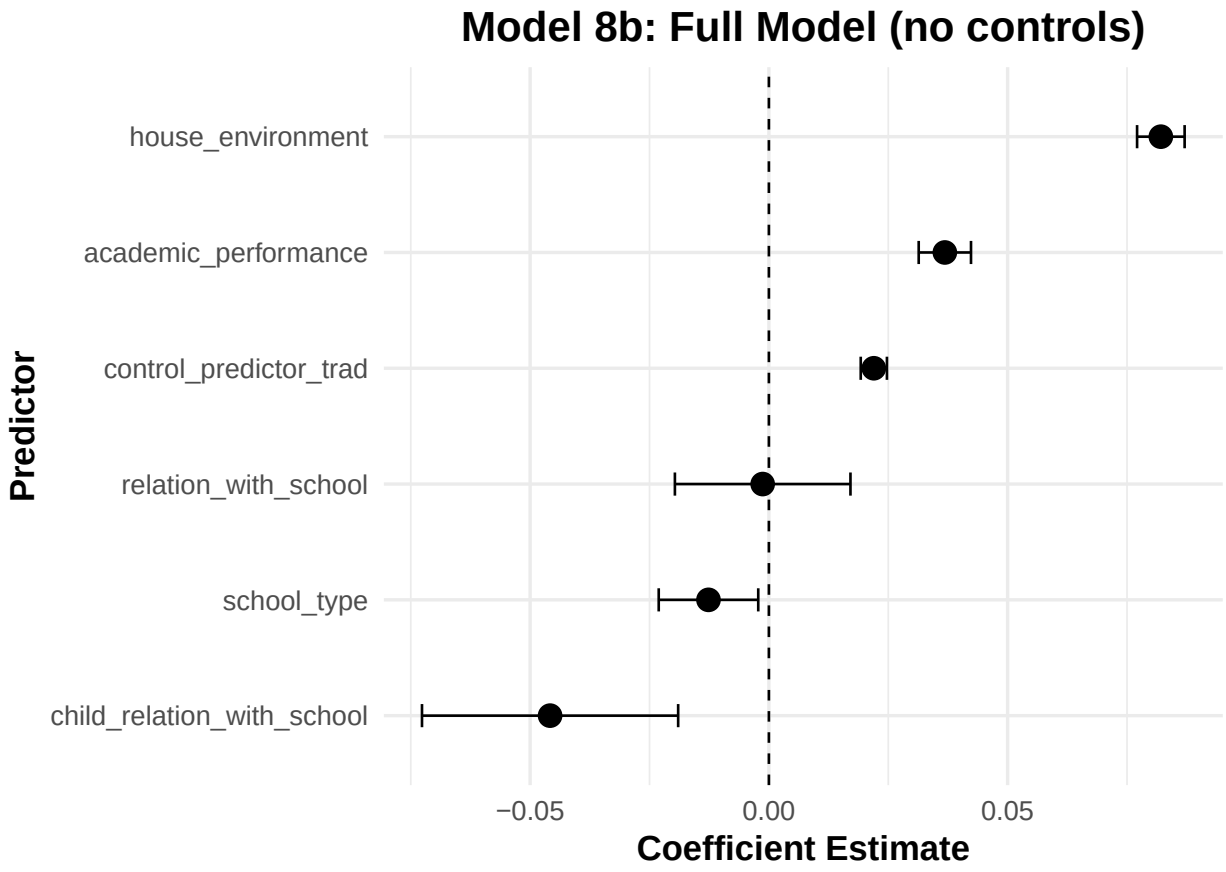


Figure 41: Model Coefficients

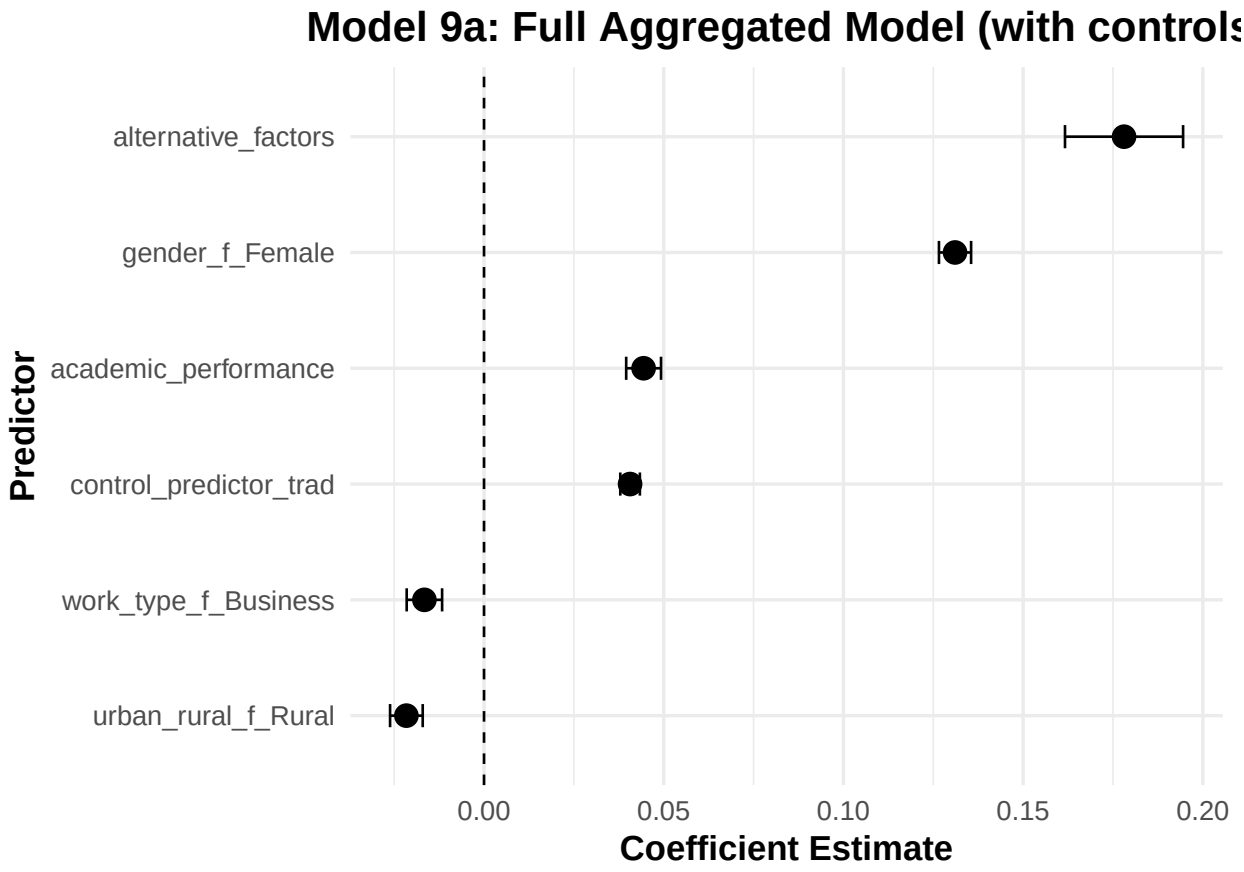


Figure 42: Model Coefficients

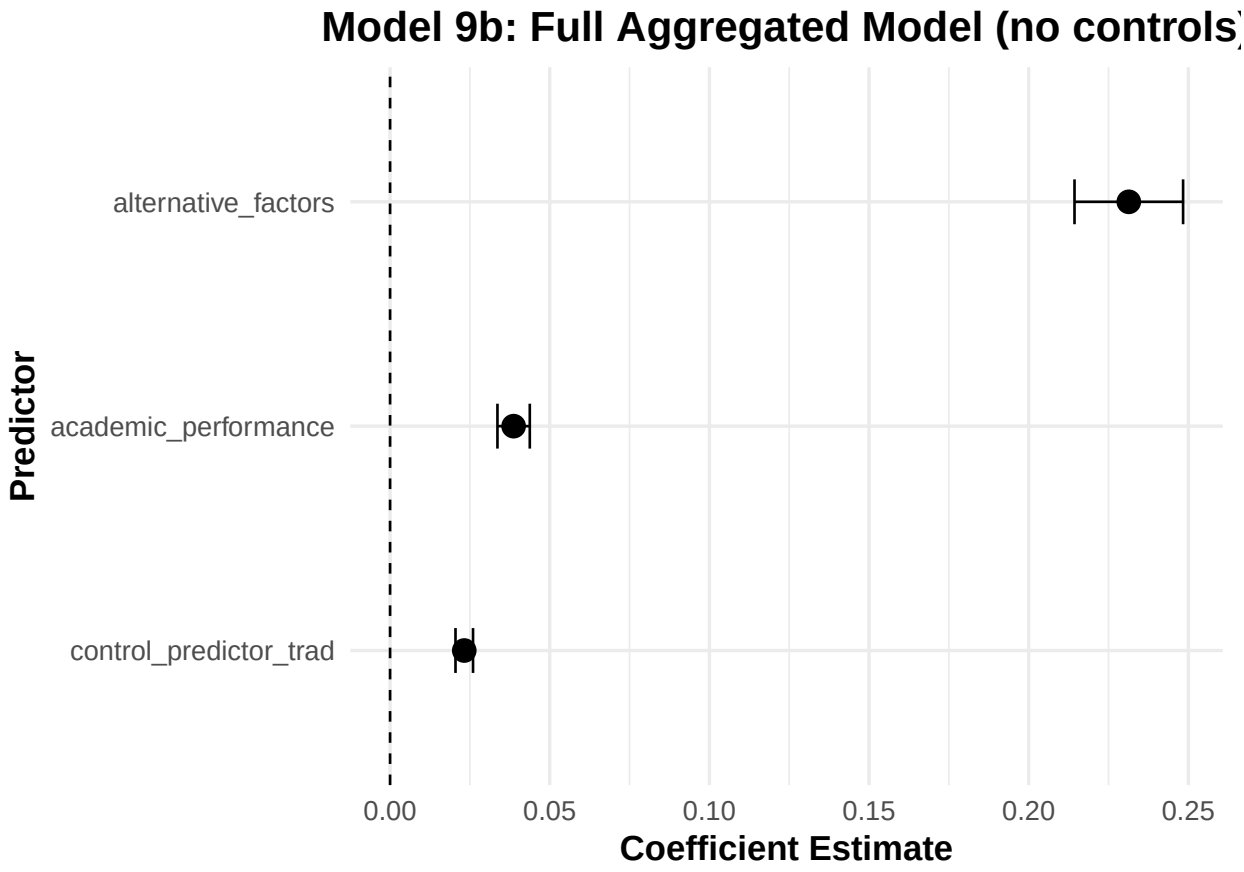


Figure 43: Model Coefficients

- **Overall Model Fit:** Both models are statistically significant ($p < 0.001$).
- **Explanatory Power:** Model 9a (with controls) has an Adjusted R-squared of 0.1736704, while Model 9b (without controls) has an Adjusted R-squared of 9.011520e-02. The inclusion of demographic controls in Model 9a significantly increases the explanatory power, providing robust support for our central argument. We can see the that although the full model perform way better than individual models, the aggregated model has a better fit.
- **Coefficients:** The aggregated `alternative_factors` variable remains a significant predictor even when controlling for `academic_performance`. This provides robust support for our central argument.

Summary of Regression Models

Here is a summary table of all our models for easy comparison.

Table 26: Regression Models of Responsible Social Contribution

Model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
model1a	(Intercept)	0.0044531	0.0024787	1.7965173	0.0724211	-0.0004053	0.0093114
model1a	academic_performance	0.0661920	0.0022637	29.2403538	0.0000000	0.0617550	0.0706290
model1a	control_predictor_trad	0.0480559	0.0013545	35.4785237	0.0000000	0.0454010	0.0507108
model1a	gender_f_Female	0.1361832	0.0022831	59.6487710	0.0000000	0.1317083	0.1406581
model1a	urban_rural_f_Rural	-0.0272828	0.0023177	-11.7714062	0.0000000	-0.0318256	-0.0227400
model1a	work_type_f_Business	-0.0144071	0.0025279	-5.6991289	0.0000000	-0.0193619	-0.0094522
model1b	(Intercept)	0.0290606	0.0012733	22.8225960	0.0000000	0.0265649	0.0315564
model1b	academic_performance	0.0677804	0.0023665	28.6421981	0.0000000	0.0631421	0.0724188
model1b	control_predictor_trad	0.0330014	0.0013645	24.1853596	0.0000000	0.0303269	0.0356759
model2a	(Intercept)	0.0827196	0.0021772	37.9935592	0.0000000	0.0784523	0.0869870
model2a	relation_with_school	-0.0097478	0.0094476	-1.0317697	0.3021873	-0.0282654	0.0087699
model2a	gender_f_Female	0.1069119	0.0023320	45.8450118	0.0000000	0.1023410	0.1114827
model2a	urban_rural_f_Rural	-0.0674621	0.0023334	-28.9120544	0.0000000	-0.0720355	-0.0628886
model2a	work_type_f_Business	-0.0180538	0.0026566	-6.7957366	0.0000000	-0.0232609	-0.0128467
model2b	(Intercept)	0.0612672	0.0010752	56.9810339	0.0000000	0.0591598	0.0633747
model2b	relation_with_school	-0.0124282	0.0098192	-1.2657032	0.2056279	-0.0316742	0.0068178
model3a	(Intercept)	0.0631708	0.0022842	27.6558205	0.0000000	0.0586937	0.0676478
model3a	child_relation_with_school	0.3077038	0.0119382	25.7746153	0.0000000	0.2843044	0.3311031
model3a	gender_f_Female	0.1132303	0.0023226	48.7516697	0.0000000	0.1086779	0.1177826
model3a	urban_rural_f_Rural	-0.0600846	0.0023287	-25.8015791	0.0000000	-0.0646490	-0.0555203
model3a	work_type_f_Business	-0.0186184	0.0026314	-7.0755652	0.0000000	-0.0237760	-0.0134608
model3b	(Intercept)	0.0497656	0.0011793	42.1991695	0.0000000	0.0474542	0.0520771
model3b	child_relation_with_school	0.2776152	0.0122442	22.6732040	0.0000000	0.2536162	0.3016142
model4a	(Intercept)	0.0826220	0.0021760	37.9691142	0.0000000	0.0783569	0.0868871
model4a	school_type	0.0153338	0.0053410	2.8709823	0.0040945	0.0048654	0.0258023
model4a	gender_f_Female	0.1071228	0.0023326	45.9237551	0.0000000	0.1025508	0.1116948
model4a	urban_rural_f_Rural	-0.0669866	0.0023388	-28.6418694	0.0000000	-0.0715707	-0.0624026

Model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
model4a	work_type_f_Business	-0.0181246	0.0026565	-6.8227836	0.0000000	-0.0233314	-0.0129178
model4b	(Intercept)	0.0615721	0.0010805	56.9843527	0.0000000	0.0594543	0.0636900
model4b	school_type	0.0171060	0.0055318	3.0922784	0.0019879	0.0062634	0.0279486
model5a	(Intercept)	0.0221696	0.0023877	9.2849378	0.0000000	0.0174896	0.0268495
model5a	house_environment	0.1094215	0.0020779	52.6595227	0.0000000	0.1053487	0.1134942
model5a	gender_f_Female	0.1050457	0.0022433	46.8255653	0.0000000	0.1006487	0.1094427
model5a	urban_rural_f_Rural	-0.0372230	0.0023167	-16.0674869	0.0000000	-0.0417637	-0.0326822
model5a	work_type_f_Business	-0.0202095	0.0025557	-7.9077086	0.0000000	-0.0252187	-0.0152003
model5b	(Intercept)	0.0186906	0.0012806	14.5949358	0.0000000	0.0161806	0.0212007
model5b	house_environment	0.1157345	0.0020775	55.7091659	0.0000000	0.1116625	0.1198064
model6a	(Intercept)	0.0179867	0.0024180	7.4386976	0.0000000	0.0132474	0.0227261
model6a	relation_with_school	0.0025314	0.0091185	0.2776070	0.7813158	-0.0153412	0.0204039
model6a	child_relation_with_school	0.1288291	0.0122066	10.5540095	0.0000000	0.1049036	0.1527545
model6a	school_type	-0.0000220	0.0051654	-0.0042608	0.9966004	-0.0101464	0.0101024
model6a	house_environment	0.1022009	0.0021924	46.6161557	0.0000000	0.0979037	0.1064981
model6a	gender_f_Female	0.1078119	0.0022563	47.7816374	0.0000000	0.1033894	0.1122344
model6a	urban_rural_f_Rural	-0.0361311	0.0023188	-15.5817882	0.0000000	-0.0406760	-0.0315861
model6a	work_type_f_Business	-0.0203050	0.0025518	-7.9571146	0.0000000	-0.0253066	-0.0153034
model6b	(Intercept)	0.0171791	0.0013157	13.0567975	0.0000000	0.0146003	0.0197580
model6b	relation_with_school	0.0009789	0.0094459	0.1036310	0.9174628	-0.0175353	0.0194931
model6b	child_relation_with_school	0.0654087	0.0125357	5.2177865	0.0000002	0.0408383	0.0899792
model6b	school_type	-0.0064001	0.0053378	-1.1990027	0.2305352	-0.0168623	0.0040622
model6b	house_environment	0.1121220	0.0022122	50.6825739	0.0000000	0.1077860	0.1164581
model7a	(Intercept)	0.0277504	0.0023776	11.6717037	0.0000000	0.0230903	0.0324106
model7a	alternative_factors	0.3514100	0.0071047	49.4615035	0.0000000	0.3374845	0.3653355
model7a	gender_f_Female	0.1084951	0.0022532	48.1510935	0.0000000	0.1040787	0.1129115
model7a	urban_rural_f_Rural	-0.0383269	0.0023300	-16.4492969	0.0000000	-0.0428938	-0.0337600
model7a	work_type_f_Business	-0.0203002	0.0025671	-7.9079269	0.0000000	-0.0253318	-0.0152687
model7b	(Intercept)	0.0254904	0.0012473	20.4370652	0.0000000	0.0230457	0.0279350
model7b	alternative_factors	0.3642706	0.0070999	51.3063620	0.0000000	0.3503545	0.3781866
model8a	(Intercept)	-0.0101193	0.0025447	-3.9766377	0.0000700	-0.0151070	-0.0051316
model8a	academic_performance	0.0405884	0.0026785	15.1531919	0.0000000	0.0353384	0.0458385
model8a	relation_with_school	0.0034374	0.0089601	0.3836383	0.7012489	-0.0141247	0.0209996
model8a	child_relation_with_school	0.0014383	0.0131021	0.1097782	0.9125859	-0.0242422	0.0271188
model8a	school_type	-0.0081215	0.0050860	-1.5968358	0.1103115	-0.0180903	0.0018472
model8a	house_environment	0.0613768	0.0024681	24.8679802	0.0000000	0.0565393	0.0662144
model8a	control_predictor_trad	0.0393870	0.0013942	28.2510743	0.0000000	0.0366544	0.0421196

Model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
model8a	gender_f_Female	0.1281191	0.0022900	55.9476602	0.0000000	0.1236307	0.1326076
model8a	urban_rural_f_Rural	-0.0211749	0.0023133	-9.1536983	0.0000000	-0.0257090	-0.0166408
model8a	work_type_f_Business	-0.0167940	0.0025083	-6.6954284	0.0000000	-0.0217103	-0.0118777
model8b	(Intercept)	0.0127122	0.0014057	9.0434314	0.0000000	0.0099570	0.0154673
model8b	academic_performance	0.0368515	0.0027945	13.1870844	0.0000000	0.0313742	0.0423288
model8b	relation_with_school	-0.0013015	0.0093786	-0.1387728	0.8896305	-0.0196840	0.0170810
model8b	child_relation_with_school	-0.0458247	0.0136862	-3.3482394	0.0008141	-0.0726501	-0.0189993
model8b	school_type	-0.0126501	0.0053171	-2.3791155	0.0173597	-0.0230719	-0.0022283
model8b	house_environment	0.0820864	0.0025401	32.3157391	0.0000000	0.0771076	0.0870651
model8b	control_predictor_trad	0.0219789	0.0013988	15.7127713	0.0000000	0.0192372	0.0247205
model9a	(Intercept)	-0.0067627	0.0025186	-2.6851409	0.0072534	-0.0116991	-0.0018262
model9a	academic_performance	0.0443981	0.0024719	17.9609715	0.0000000	0.0395530	0.0492431
model9a	alternative_factors	0.1780723	0.0083820	21.2446895	0.0000000	0.1616434	0.1945012
model9a	control_predictor_trad	0.0406334	0.0013903	29.2260518	0.0000000	0.0379083	0.0433585
model9a	gender_f_Female	0.1310464	0.0022811	57.4489381	0.0000000	0.1265753	0.1355174
model9a	urban_rural_f_Rural	-0.0216007	0.0023181	-9.3181615	0.0000000	-0.0261443	-0.0170571
model9a	work_type_f_Business	-0.0165909	0.0025136	-6.6004169	0.0000000	-0.0215176	-0.0116641
model9b	(Intercept)	0.0174678	0.0013332	13.1020583	0.0000000	0.0148547	0.0200810
model9b	academic_performance	0.0386939	0.0025838	14.9754083	0.0000000	0.0336295	0.0437583
model9b	alternative_factors	0.2313502	0.0086754	26.6672929	0.0000000	0.2143460	0.2483543
model9b	control_predictor_trad	0.0232363	0.0013994	16.6047942	0.0000000	0.0204935	0.0259791

Models of Best Fit

After performing permutations and combinations, a model than includes the `academic performance` and `house environment` and controlling for `demographic factors` performs best.

Table 27: Best Model Summary

	Variable	Coefficient	Std_Error	t_value	p_value	Significance
(Intercept)	(Intercept)	-0.0098402	0.0025228	-3.900466	9.62e-05	***
academic_performance	academic_performance	0.0410033	0.0024611	16.660561	0.00e+00	***
house_environment	house_environment	0.0611379	0.0024555	24.898678	0.00e+00	***
control_predictor_trad	control_predictor_trad	0.0391932	0.0013888	28.220185	0.00e+00	***
gender_f_Female	gender_f_Female	0.1281516	0.0022856	56.068301	0.00e+00	***
urban_rural_f_Rural	urban_rural_f_Rural	-0.0210235	0.0023108	-9.097961	0.00e+00	***
work_type_f_Business	work_type_f_Business	-0.0168247	0.0025073	-6.710258	0.00e+00	***
1	R-squared	0.1778046	NA	NA	NA	
2	Adj. R-squared	0.1776608	NA	NA	NA	
3	F-statistic	1236.0828139	NA	NA	NA	

	Variable	Coefficient	Std_Error	t_value	p_value	Significance
4	Observations	34302.0000000	NA	NA	NA	

Findings

The findings of this study have important implications for educational policy. Our results suggest that an exclusive focus on academic metrics may be misguided. While academic skills are undoubtedly important, our findings indicate that a broader set of factors, including their relationship with their school and their home environment, are also critical for fostering responsible and engaged citizens.

Conclusion

This study provides compelling evidence that factors "beyond marks" are important predictors of an individual's responsible social contribution. Our findings challenge the narrow focus on academic achievement that is prevalent in many educational systems and suggest that a more holistic approach to education is needed. By investing in programs and policies that improve student's social integration in the school and home environments of students, we can help to foster a new generation of responsible and engaged citizens.

Policy Prescription

Based on the findings of this study, we offer the following evidence-based policy prescriptions:

1. **Promote a Holistic Approach to Education: Empirical Justification:** Model 8 demonstrates that both `academic_performance` and `alternative_factors` are statistically significant predictors of responsible social contribution, even when controlling for each other. The model's Adjusted R-squared (0.1776571) is an improvement over Model 1 (0.1628199) and Model 6 (0.1463064), indicating that a model including both is superior. **Recommendation:** This provides strong empirical support for broadening educational policies to officially recognize, measure, and reward a wider range of student attributes beyond academic performance, such as social-emotional learning and civic engagement.

2. **Invest in Teacher Training and Support: Empirical Justification:** The `alternative_factors` composite, which includes student-teacher relationships, is a good predictor in Models 2, 3 and 6. While the individual component `relation_with_school` is not significant in the full model, its contribution to the significant composite variable suggests its importance. **Recommendation:** Policies should focus on providing teachers with the training and support they need to create a positive and supportive learning environment for all students.

3. **Strengthen School-Family Partnerships: Empirical Justification:** The `house_environment` variable is a key component of the `alternative_factors` composite, which was a strong and significant predictor in Models 6 and 8. **Recommendation:** This provides a clear empirical basis for policies that aim to strengthen the partnership between schools and families, such as programs that encourage parental involvement in education and provide parenting support.

4. **Focus less on School Type and more on School Integration: Empirical Justification:** Model 4 and Model 8 provides the most direct evidence for this recommendation. The `school_type` variable is not statistically significant predictor of responsible social contribution in the individual model (Model 4) and even when if does predict in the combined model (Model 8) it has negative relation. **Recommendation:** This finding strongly suggests that the type of school a child attends has less lasting impact on their social contribution. This dejustifies the argument of school type as opposed to the social integration of the student.

5. **Promote Gender Equity in Education: Empirical Justification:** Across all four regression models, the coefficient for being female (`gender_fFemale`) is positive and highly significant ($p < 0.01$). This is one of the most robust findings in the analysis. **Recommendation:** This empirical result highlights the need for policies that not only ensure access to education for girls but also investigate and cultivate the factors that lead to their higher measured social contribution.

Bibliography

- Desai, S., & Kulkarni, V. (2008). Changing educational inequalities in India in the context of affirmative action. *Demography*, *45*(2), 245-270.
- Dreze, J., & Kingdon, G. G. (2001). School participation in rural India. *Review of Development Economics*, *5*(1), 1-24.
- Kingdon, G. G. (2007). The progress of school education in India. *Oxford Review of Economic Policy*, *23*(2), 168-195.
- National Council of Applied Economic Research (NCAER) & University of Maryland. (2015). *India Human Development Survey-II (IHDS-II), 2011-12*. Inter-university Consortium for Political and Social Research [distributor], 2015-07-31. https://doi.org/10.3886/ICPSR36151.v1
- Zimmerman, D. J. (1992). Regression with a restricted dependent variable. *International Economic Review*, 33(3), 529-548.