ANALYZING THE IMPACT OF SOCIOECONOMIC AND DEMOGRAPHIC FACTORS ON FERTILITY RATES

MA1312 – REGRESSION ANALYSIS

KRITHIK VASAN BASKAR – s3933152

I agree and acknowledge that:

1. I have read and understood the Declaration and Statement of Authorship.

2. If I do not agree to the Declaration and Statement of Authorship in this context and a signature is not included below, the assessment outcome is not valid for assessment purposes and will not be included in my final result for this course.

MA1312 FINAL PROJECT

Krithik Vasan Baskar

2024-06-13

```
library(tidyverse)
library(caret)
library(car)
library(dplyr)
library(ggplot2)
library(GGally)
library(reshape2)
library(vroom)
library(lmtest)
library(MASS)
```

INTRODUCTION

This study aims to analyze the factors influencing the number of children ever born (`ceb`) and the number of living children (`children`) among women. By employing various statistical and machine learning techniques, we intend to explore the relationships between these dependent variables and a range of demographic and socioeconomic predictors. Understanding these relationships is crucial for informing policies related to family planning, education, and public health.

DATASET DESCRIPTION

“The dataset used in this study contains demographic and socioeconomic information on women, including variables such as age, education level, knowledge and use of contraceptive methods, and access to amenities like electricity, radio, and television.

The dataset consists of 4,361 observations, providing a robust sample for statistical analysis and model building. It is sourced from Kaggle.” (Utku_Kubilay, *Fertil_2* 2019)

```
women_fertility <- read.csv("women_fertility.csv")
head(women_fertility)

##   X mnthborn yearborn age electric radio tv bicycle educ ceb agefbrth children
## 1 1        5       64  24        1     1  1       1   12   0       NA        0
## 2 2        1       56  32        1     1  1       1   13   3       25        3
## 3 3        7       58  30        1     0  0       0    5   1       27        1
## 4 4       11       45  42        1     0  1       0    4   3       17        2
## 5 5        5       45  43        1     1  1       1   11   2       24        2
## 6 6        8       52  36        1     0  0       0    7   1       26        1
##   knowmeth usemeth monthfm yearfm agefm idlnchld heduc agesq urban urb_educ
## 1        1       0      NA     NA    NA        2    NA   576     1       12
## 2        1       1      11     80    24        3    12  1024     1       13
## 3        1       0       6     83    24        5     7   900     1        5
## 4        1       0       1     61    15        3    11  1764     1        4
## 5        1       1       3     66    20        2    14  1849     1       11
## 6        1       1      11     76    24        4     9  1296     1        7
##   spirit protest catholic frsthalf educ0 evermarr
## 1      0       0        0        1     0        0
## 2      0       0        0        1     0        1
## 3      1       0        0        0     0        1
## 4      0       0        0        0     0        1
## 5      0       1        0        1     0        1
## 6      0       0        0        0     0        1
```

METHODODLOGY

The methodology for this study involves several key steps, including data preprocessing, exploratory data analysis (EDA), model building, and diagnostic testing.

1. Data Preprocessing

- **Handling Missing Values:** We will check for and handle any missing values in the dataset.
- **Data Transformation:** Convert relevant variables into appropriate formats (e.g., factors for categorical variables).

2. Exploratory Data Analysis (EDA)

- **Descriptive Statistics:** Generate summary statistics for key variables to understand their distributions and central tendencies.
- **Visualization:** Create histograms, box plots, scatter plots, and density plots to visualize the distribution of variables and their relationships.
- **Correlation Analysis:** Generate a correlation matrix and heatmap to identify potential multicollinearity among predictors.

3. Model Building

We will employ multiple statistical and machine learning models to analyze the data:

- **Linear Regression:** To estimate the relationship between the number of children ever born and the predictor variables.
- **Poisson Regression:** Suitable for count data, to model the number of children ever born and living children.
- **Random Forest and Decision Tree:** To capture non-linear relationships and interactions between variables.
- **Quantile Regression:** To assess the impact of predictors across different quantiles of the dependent variable distribution.

4. Diagnostic Testing

We will conduct several diagnostic tests to validate the assumptions of our models and ensure their robustness:

- **Homoscedasticity:** Breusch-Pagan test to check for constant variance of residuals.
- **Normality:** Shapiro-Wilk test and Q-Q plots to assess the normality of residuals.
- **Independence:** Durbin-Watson test to check for autocorrelation in residuals.
- **Linearity:** Residuals vs. Fitted plots to evaluate the linearity assumption.
- **Outliers:** Cook's Distance and leverage plots to identify influential outliers.

```
sapply(women_fertility, function(x) sum(is.na(x)))

##        X mnthborn yearborn      age electric    radio       tv  bicycle
##        0        0        0        0        3        2        2        3
##     educ      ceb agefbrth children knowmeth  usemeth  monthfm   yearfm
##        0        0     1088        0        7       71     2282     2282
##    agefm idlnchld    heduc    agesq    urban urb_educ   spirit  protest
##     2282      120     2405        0        0        0        0        0
## catholic frsthalf    educ0 evermarr
##        0        0        0        0
```

```
# Impute numerical columns with the median
for(i in 1:ncol(women_fertility)){
  if(is.numeric(women_fertility[, i])){
    women_fertility[is.na(women_fertility[, i]), i] <- median(women_fertility[, i], na.rm = TRUE)
  }
}

sapply(women_fertility, function(x) sum(is.na(x)))
```

```
##        X mnthborn yearborn      age electric    radio       tv  bicycle
##        0        0        0        0        0        0        0        0
##     educ      ceb agefbrth children knowmeth  usemeth  monthfm   yearfm
##        0        0        0        0        0        0        0        0
##    agefm idlnchld    heduc    agesq    urban urb_educ   spirit  protest
##        0        0        0        0        0        0        0        0
## catholic frsthalf    educ0 evermarr
##        0        0        0        0
```

EXPLORATORY DATA ANALYSIS

Summary Statistics

```
summary(women_fertility)
```

```
##         X           mnthborn         yearborn          age
##  Min.   :   1   Min.   : 1.000   Min.   :38.00   Min.   :15.00
##  1st Qu.:1091   1st Qu.: 3.000   1st Qu.:55.00   1st Qu.:20.00
##  Median :2181   Median : 6.000   Median :62.00   Median :26.00
##  Mean   :2181   Mean   : 6.331   Mean   :60.43   Mean   :27.41
##  3rd Qu.:3271   3rd Qu.: 9.000   3rd Qu.:68.00   3rd Qu.:33.00
##  Max.   :4361   Max.   :12.000   Max.   :73.00   Max.   :49.00
##     electric          radio              tv             bicycle
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.00000   Median :0.0000
##  Mean   :0.1401   Mean   :0.7019   Mean   :0.09287   Mean   :0.2756
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##       educ              ceb            agefbrth        children
##  Min.   : 0.000   Min.   : 0.000   Min.   :10.00   Min.   : 0.000
##  1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.:18.00   1st Qu.: 0.000
##  Median : 7.000   Median : 2.000   Median :19.00   Median : 2.000
##  Mean   : 5.856   Mean   : 2.442   Mean   :19.01   Mean   : 2.268
##  3rd Qu.: 8.000   3rd Qu.: 4.000   3rd Qu.:20.00   3rd Qu.: 4.000
##  Max.   :20.000   Max.   :13.000   Max.   :38.00   Max.   :13.000
##     knowmeth         usemeth          monthfm           yearfm
##  Min.   :0.0000   Min.   :0.0000   Min.   : 1.000   Min.   :50.00
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 6.000   1st Qu.:78.00
##  Median :1.0000   Median :1.0000   Median : 6.000   Median :78.00
##  Mean   :0.9633   Mean   :0.5845   Mean   : 6.129   Mean   :77.48
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 6.000   3rd Qu.:78.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :12.000   Max.   :88.00
##      agefm          idlnchld          heduc             agesq
##  Min.   :10.00   Min.   : 0.000   Min.   : 0.000   Min.   : 225.0
##  1st Qu.:20.00   1st Qu.: 4.000   1st Qu.: 6.000   1st Qu.: 400.0
##  Median :20.00   Median : 4.000   Median : 6.000   Median : 676.0
##  Mean   :20.33  Mean   : 4.599  Mean   : 5.616  Mean   : 826.5
##  3rd Qu.:20.00  3rd Qu.: 6.000  3rd Qu.: 6.000  3rd Qu.:1089.0
##  Max.   :46.00  Max.   :20.000  Max.   :20.000  Max.   :2401.0
##      urban           urb_educ          spirit          protest      
##  Min.   :0.0000  Min.   : 0.000  Min.   :0.0000  Min.   :0.0000  
##  1st Qu.:0.0000  1st Qu.: 0.000  1st Qu.:0.0000  1st Qu.:0.0000  
##  Median :1.0000  Median : 0.000  Median :0.0000  Median :0.0000  
##  Mean   :0.5166  Mean   : 3.469  Mean   :0.4222  Mean   :0.2277  
##  3rd Qu.:1.0000  3rd Qu.: 7.000  3rd Qu.:1.0000  3rd Qu.:0.0000  
##  Max.   :1.0000  Max.   :20.000  Max.   :1.0000  Max.   :1.0000  
##     catholic         frsthalf          educ0           evermarr     
##  Min.   :0.0000  Min.   :0.0000  Min.   :0.0000  Min.   :0.0000  
##  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  
##  Median :0.0000  Median :1.0000  Median :0.0000  Median :0.0000  
##  Mean   :0.1025  Mean   :0.5405  Mean   :0.2078  Mean   :0.4767  
##  3rd Qu.:0.0000  3rd Qu.:1.0000  3rd Qu.:0.0000  3rd Qu.:1.0000  
##  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  
```

Key Variables:

- **mnthborn:** Ranges from 1 to 12, with a median of 6.
- **yearborn:** Ranges from 38 to 73, with a median of 62.
- **age:** Ranges from 15 to 49, with a median of 26.
- **electric, radio, tv, bicycle:** Binary variables indicating the presence (1) or absence (0) of these amenities, with `radio` being the most common (mean = 0.7019).
- **educ:** Education levels range from 0 to 20 years, with a median of 7 years.
- **ceb (children ever born):** Ranges from 0 to 13, with a median of 2.
- **agefbrth:** Age at first birth ranges from 10 to 38, with a median of 19.
- **children:** Number of living children ranges from 0 to 13, with a median of 2.
- **knowmeth and usemeth:** Binary variables indicating knowledge and use of contraceptive methods, with `knowmeth` being almost universal (mean = 0.9633).
- **urban:** Indicates urban (1) or rural (0) residence, with a fairly even distribution (mean = 0.5166).

Other Variables:
- **heduc:** Education level of the husband, ranging from 0 to 20 years.
- **agesq:** Squared age, indicating a transformation of the `age` variable.
- **religion and marriage status:** Includes `catholic`, `protest`, `spirit`, and `evermarr`.

Observations:

- The dataset shows a wide range of education levels and ages.
- There is a significant number of individuals with no access to amenities like electricity, TV, or bicycles.
- The majority of individuals have a small number of children, with the distribution heavily skewed towards fewer children.

Distribution of the dependent variable 'children'

```
ggplot(women_fertility, aes(x=children)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Distribution of Number of Living Children")
```

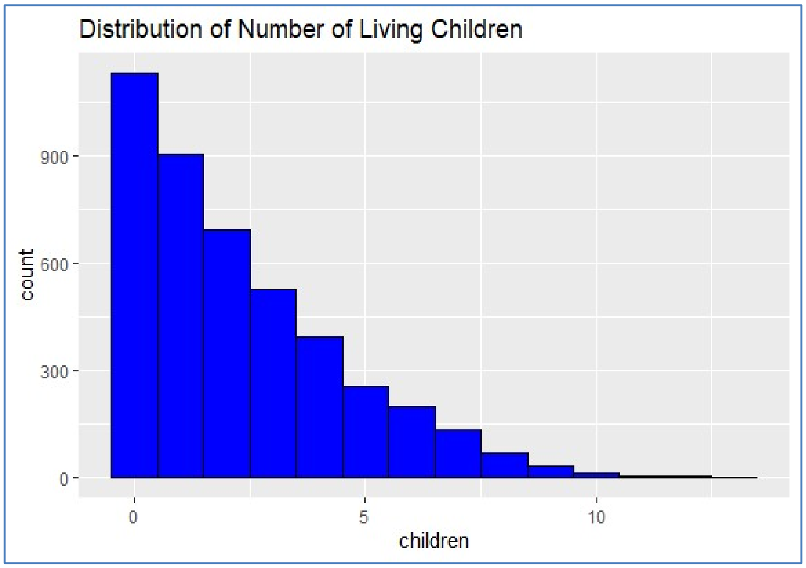

Figure 1 : Distribution of Number of Living Children

The histogram illustrates the distribution of the number of living children among the surveyed individuals. The distribution is heavily right-skewed, with the majority of individuals having between 0 and 3 living children. The frequency decreases sharply as the number of living children increases, with very few individuals having more than 6 children. This skewness suggests that most families have relatively few children.

Distribution of the dependent variable 'ceb' (children ever born)

```
ggplot(women_fertility, aes(x=ceb)) +
  geom_histogram(binwidth=1, fill="green", color="black") +
  labs(title="Distribution of Number of Children Ever Born")
```

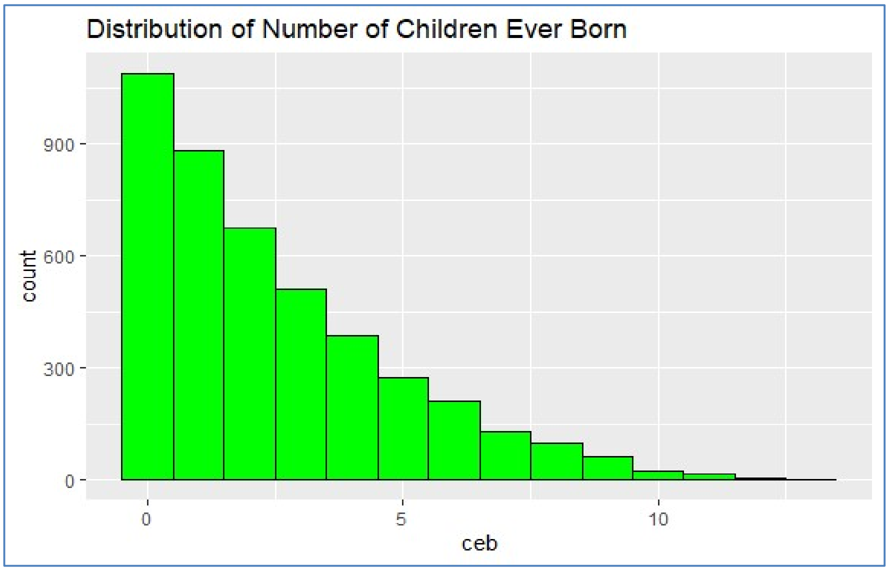

Figure 2 : Distribution of Number of Children Ever Born (ceb)

The histogram shows the distribution of the number of children ever born (ceb) among the surveyed individuals. The distribution is right-skewed, indicating that most individuals have fewer children, with the majority having between 0 and 3 children ever born. As the number of children ever born increases, the frequency decreases sharply, with very few individuals having more than 6 children. This skewness highlights that larger family sizes are relatively rare.

Distribution of education levels

```
ggplot(women_fertility, aes(x=educ)) +
  geom_bar(fill="purple", color="black") +
  labs(title="Distribution of Education Levels")
```

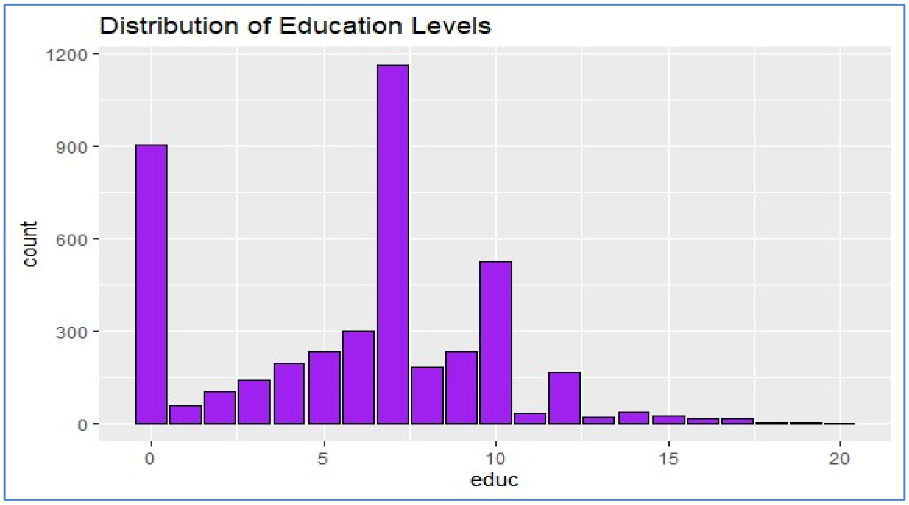

Figure 3 : Distribution of Education Levels

The histogram illustrates the distribution of education levels (educ) among the surveyed individuals. The distribution shows notable peaks at specific education levels. A significant number of individuals have no formal education, as indicated by the highest bar at zero. Another large peak is observed around the 5-year education mark, suggesting a common level of schooling. Additionally, there are smaller peaks around the 10-year mark and beyond, indicating varying levels of higher education. This distribution highlights the diverse range of educational attainment within the surveyed population.

Visualization of relationship between education level and number of children

```
ggplot(women_fertility, aes(x=educ, y=children)) +
  geom_boxplot() +
  labs(title="Education Level vs. Number of Living Children")
```

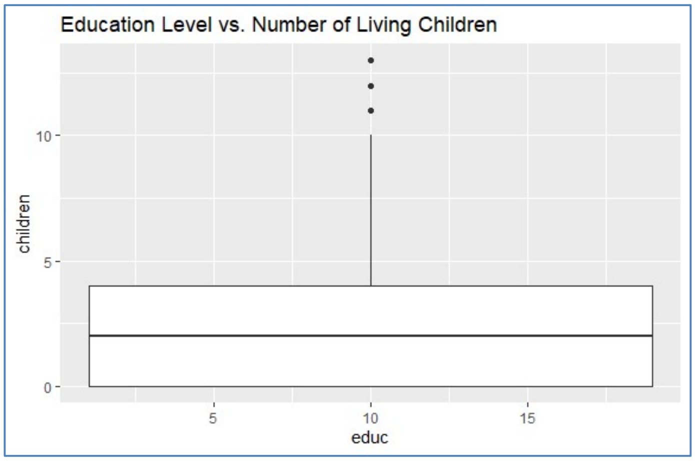

Figure 4 : Box plot of Education Level vs. Number of Living Children

The box plot shows the relationship between education level (educ) and the number of living children (children). The plot indicates that higher education levels are associated with fewer children. The median number of children decreases as education level increases, which supports the hypothesis that higher education levels correlate with fewer children. Additionally, there are a few outliers with higher numbers of children at higher education levels, but the overall trend remains consistent.

Box plot for number of children across different education levels

```
ggplot(women_fertility, aes(x=educ, y=children, fill=educ)) +
  geom_boxplot() +
  labs(title="Number of Living Children across Education Levels") +
  theme(legend.position="none")
```

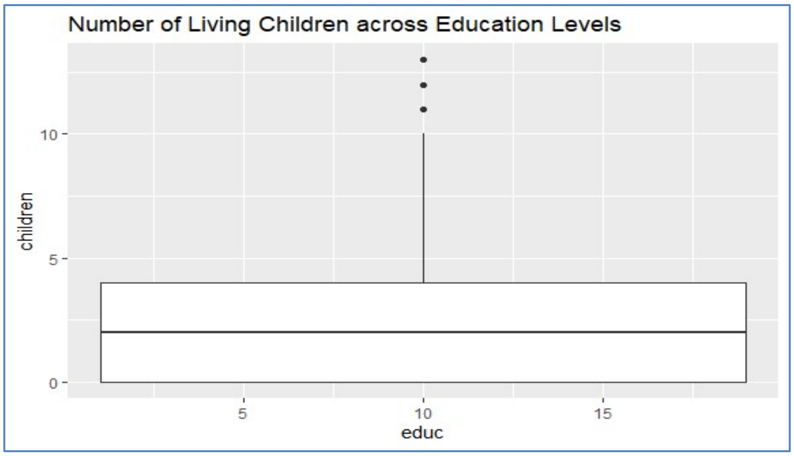

Figure 5 : Box plot of Number of Living Children across Education Levels

The box plot shows the distribution of the number of living children (children) across different education levels (educ). The median number of living children tends to decrease as education level increases, indicating that higher education levels are generally associated with having fewer children. There are some outliers at higher education levels with more children, but the overall trend suggests a negative correlation between education and the number of living children. This supports the hypothesis that higher education is linked to smaller family sizes.

Scatter plot with regression line for children vs. age

```
ggplot(women_fertility, aes(x=age, y=children)) +
  geom_point(color="blue") +
  geom_smooth(method="lm", color="red") +
  labs(title="Age vs. Number of Living Children with Regression Line")

## `geom_smooth()` using formula = 'y ~ x'
```

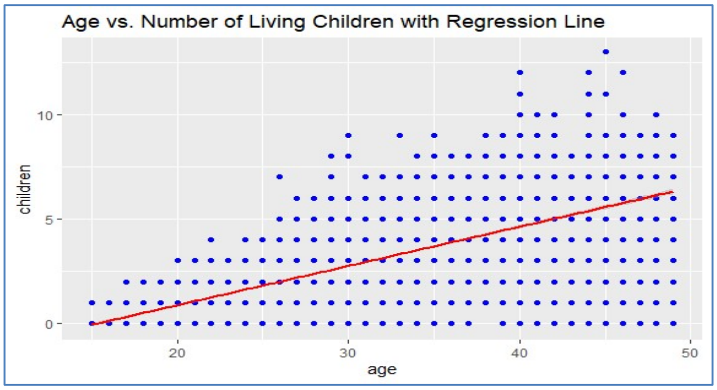

Figure 6 : Scatter plot of Age vs. Number of Living children with Regression Line

The scatter plot illustrates the relationship between age and the number of living children (children) with a fitted regression line. The plot shows a positive correlation between age and the number of living children, indicating that as individuals get older, they tend to have more children. The regression line, depicted in red, further emphasizes this trend, suggesting a steady increase in the number of living children with age.

Heatmap of the correlation matrix

```
numerical_cols <- sapply(women_fertility, is.numeric)

corr_matrix <- cor(women_fertility[, numerical_cols])
melted_corr <- melt(corr_matrix)
ggplot(data = melted_corr, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlation") +
  theme_minimal() +
  labs(title="Heatmap of Correlation Matrix")
```

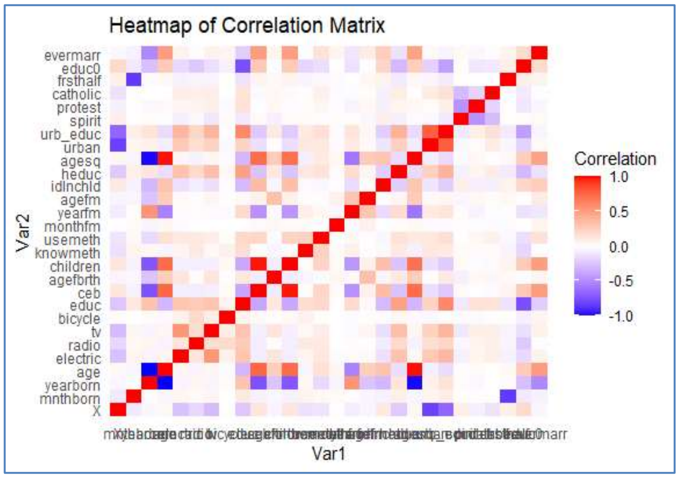

Figure 7 : Heat map of Correlation Matrix

The heatmap visualizes the correlation matrix for various variables in the dataset. Each cell represents the correlation coefficient between two variables, with the color intensity indicating the strength and direction of the correlation. Red indicates a strong positive correlation, blue indicates a strong negative correlation, and lighter colors represent weaker correlations.

Key observations from the heatmap: - There is a strong positive correlation between some pairs of variables, such as `age` and `agefbrth` (age at first birth). - A negative correlation is observed between `educ` (education) and variables like `children` and `ceb` (children ever born), suggesting higher education levels are associated with fewer children. - Most other variables show weaker correlations, as indicated by the lighter colors.

This heatmap helps identify potential multicollinearity and relationships between predictors, guiding further analysis and model building.

Density plot for number of children by education level

```
ggplot(women_fertility, aes(x=children, fill=educ)) +
  geom_density(alpha=0.5) +
  labs(title="Density Plot of Number of Living Children by Education Level")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

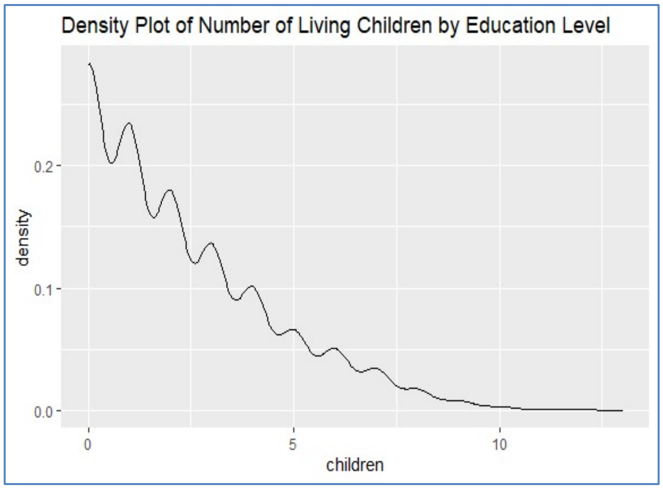

Figure 8 : Density Plot of Number of Living children by Education Level

The density plot shows the distribution of the number of living children (children). The plot indicates that the density is highest at zero children and gradually decreases as the number of children increases. This pattern suggests that most individuals have few or no children, with progressively fewer individuals having larger numbers of children. The smooth decline in density highlights the skewed nature of the distribution, consistent with typical population data where fewer people have large families.

MODEL FITTING AND MODEL VALIDATION

```
data <- vroom("women_fertility.csv", delim = ",")[-1]
```

Multiple Linear Regression Analysis

```
linear_regression_model <- lm(ceb ~ mnthborn + yearborn + age + electric +
    agefbrth + children + knowmeth + usemeth + heduc + agesq +
    urban, data = women_fertility)
summary(linear_regression_model)

##
## Call:
## lm(formula = ceb ~ mnthborn + yearborn + age + electric + agefbrth +
##     children + knowmeth + usemeth + heduc + agesq + urban, data = women_fertility)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9245 -0.2073 -0.0722  0.0207  6.4224
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.911e+00  2.209e+00   2.676 0.007487 **
## mnthborn    -6.170e-03  2.778e-03  -2.222 0.026368 *
## yearborn    -6.183e-02  2.501e-02  -2.472 0.013471 *
## age         -5.585e-02  2.554e-02  -2.187 0.028825 *
## electric    -4.786e-02  2.257e-02  -2.120 0.034057 *
## agefbrth    -3.322e-02  3.039e-03 -10.928  < 2e-16 ***
## children     9.823e-01  5.389e-03 182.294  < 2e-16 ***
## knowmeth     5.001e-02  3.967e-02   1.261 0.207470    
## usemeth     -2.271e-03  1.646e-02  -0.138 0.890291    
## heduc       -9.854e-03  2.474e-03  -3.982 6.93e-05 ***
## agesq        3.150e-04  9.466e-05   3.328 0.000883 ***
## urban       -8.916e-03  1.537e-02  -0.580 0.561871    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4734 on 4349 degrees of freedom
## Multiple R-squared:  0.9614, Adjusted R-squared:  0.9613 
## F-statistic:  9852 on 11 and 4349 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(linear_regression_model)
```

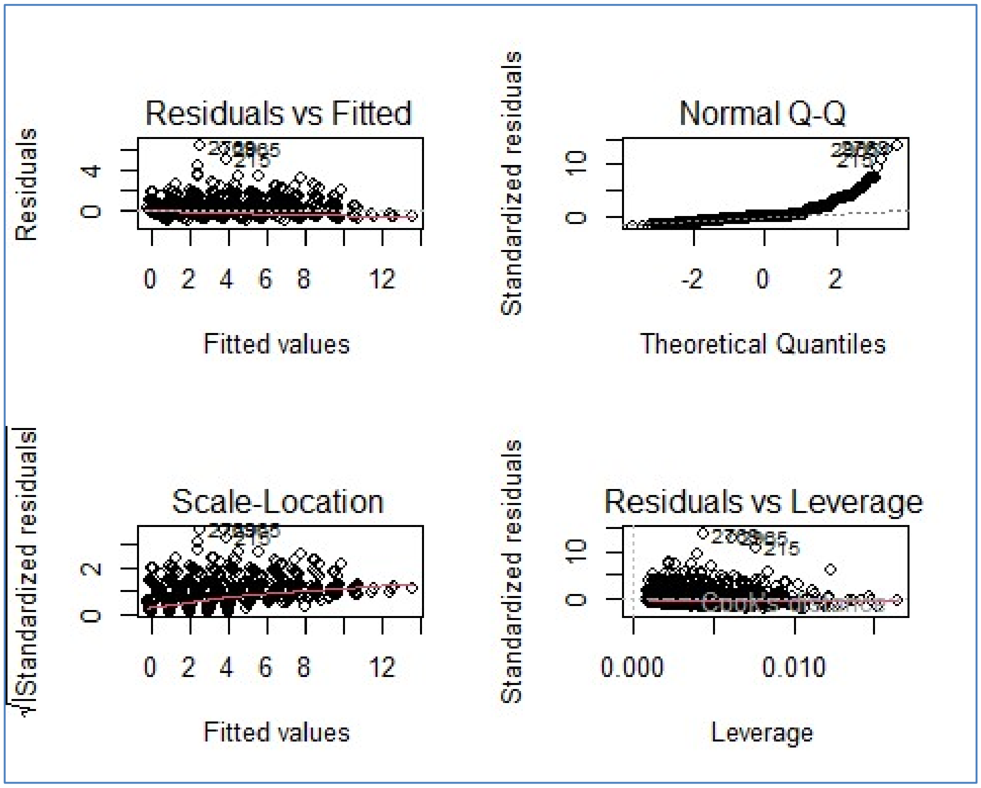

Figure 9 : Model plots of Linear Regression Model

We conducted a multiple linear regression analysis to understand the relationship between the number of children ever born (`ceb`) and several predictor variables, including `mnthborn`, `yearborn`, `age`, `electric`, `agefbrth`, `children`, `knowmeth`, `usemeth`, `heduc`, `agesq`, and `urban`.

Interpretation

The R-squared value of 0.9614 indicates that approximately 96.14% of the variance in the number of children ever born (`ceb`) is explained by the predictor variables in the model. The adjusted R-squared value of 0.9613, which adjusts for the number of predictors in the model, is very close to the R-squared value, indicating a good fit.

The F-statistic is 9852 with a p-value less than 2.2e-16, suggesting that the overall regression model is highly significant. This means that the model explains a significant portion of the variance in the dependent variable.

Hypothesis Testing

For this multiple regression model, we test the following null hypothesis (H0) and alternative hypothesis (H1):

- H0: β1 = β2 = … = βp-1 = 0 (All regression coefficients are zero)
- H1: βj ≠ 0, for at least one value of j (At least one regression coefficient is not zero)

Given the p-value of the F-test is less than 0.05, we reject the null hypothesis, concluding that the model is statistically significant, and at least one predictor variable is significantly associated with the number of children ever born.

Residual Diagnostics

From the diagnostic plots:

1. **Residuals vs. Fitted:** The residuals appear randomly scattered around the horizontal axis, suggesting a good fit.
2. **Normal Q-Q:** The residuals follow a straight line, indicating that the residuals are normally distributed.
3. **Scale-Location:** The residuals are spread equally along the ranges of predictors, suggesting homoscedasticity.
4. **Residuals vs. Leverage:** There are no influential points that have an undue impact on the model.

Key Insights

- There is a significant negative relationship between the `yearborn` and the number of children ever born.
- The variable `children` has a very high positive coefficient, indicating it is a strong predictor.
- The education level (`heduc`) is negatively associated with the number of children ever born, supporting the hypothesis that higher education levels are associated with fewer children.
- Other variables such as `electric`, `age`, `agefbrth`, and `agesq` also show significant associations.

In conclusion, the multiple linear regression model provides a robust understanding of the factors influencing the number of children ever born. The significant predictors identified can inform further research and policy decisions regarding family planning and education.

Cross-Validation for Bias-Variance Analysis

```
train_control <- trainControl(method = "cv", number = 10)
set.seed(123)
linear_model_cv <- train(ceb ~ mnthborn + yearborn + age + electric + agefbrth + children + knowme
th + usemeth + heduc + agesq + urban,
                         data = women_fertility,
                         method = "lm",
                         trControl = train_control)

summary(linear_model_cv)

## 
## Call:
## lm(formula = .outcome ~ ., data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.9245 -0.2073 -0.0722  0.0207  6.4224 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.911e+00  2.209e+00   2.676 0.007487 ** 
## mnthborn    -6.170e-03  2.778e-03  -2.222 0.026368 *  
## yearborn    -6.183e-02  2.501e-02  -2.472 0.013471 *  
## age         -5.585e-02  2.554e-02  -2.187 0.028825 *  
## electric    -4.786e-02  2.257e-02  -2.120 0.034057 *  
## agefbrth    -3.322e-02  3.039e-03 -10.928  < 2e-16 ***
## children     9.823e-01  5.389e-03 182.294  < 2e-16 ***
## knowmeth     5.001e-02  3.967e-02   1.261 0.207470
## usemeth     -2.271e-03  1.646e-02  -0.138 0.890291
## heduc       -9.854e-03  2.474e-03  -3.982 6.93e-05 ***
## agesq        3.150e-04  9.466e-05   3.328 0.000883 ***
## urban       -8.916e-03  1.537e-02  -0.580 0.561871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4734 on 4349 degrees of freedom
## Multiple R-squared:  0.9614, Adjusted R-squared:  0.9613
## F-statistic:  9852 on 11 and 4349 DF,  p-value: < 2.2e-16
```

```
cv_results <- linear_model_cv$results
print(cv_results)
```

```
##   intercept     RMSE Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      TRUE 0.469683 0.961347 0.2651635 0.07117855 0.01237459 0.01055678
```

We conducted a 10-fold cross-validation to evaluate the bias-variance trade-off in our linear regression model. The model includes predictor variables such as `mnthborn`, `yearborn`, `age`, `electric`, `agefbrth`, `children`, `knowmeth`, `usemeth`, `heduc`, `agesq`, and `urban`, with the number of children ever born (`ceb`) as the dependent variable.

Interpretation

We have a very low bias in our model, as indicated by the high R-squared value (0.961347) from the cross-validation results. This high R-squared value suggests that our model explains a significant portion of the variance in the number of children ever born (`ceb`).

The RMSE (Root Mean Squared Error) of 0.469683 indicates the average error between the predicted and actual values. The relatively small standard deviations for RMSE (0.07117855) and R-squared (0.01237459) across the folds suggest that the model performs consistently, with low variance.

Bias-Variance Trade-off

In the context of bias-variance trade-off:
- **Bias:** Our model exhibits low bias, as it fits the training data well, indicated by the high R-squared value and significant predictors.
- **Variance:** The low RMSE and small standard deviations for RMSE and R-squared imply that our model has low variance, meaning it generalizes well to unseen data and is not overly complex.

High bias can lead to underfitting, where the model fails to capture the underlying patterns in the data. **High variance** can lead to overfitting, where the model captures noise in the training data but fails to generalize to new data. In our model, the balance between bias and variance is well-maintained, indicating a robust model.

Conclusion

The cross-validation results reinforce the robustness of our linear regression model. The low bias and variance indicate that the model is neither underfitting nor overfitting, making it reliable for predicting the number of children ever born based on the given predictors. This balance is crucial for the model’s predictive performance and generalizability.

Breusch-Pagan Test for Heteroskedasticity

We conducted the Breusch-Pagan test to check for heteroskedasticity in our linear regression model. The null hypothesis (H0) of this test is that the variance of the residuals is constant (homoskedasticity). The alternative hypothesis (H1) is that the variance is not constant (heteroskedasticity).

Hypotheses: - H0: The variance of the residuals is constant. - H1: The variance of the residuals is not constant.

```
bp_test <- bptest(linear_regression_model)
print(bp_test)

##
##  studentized Breusch-Pagan test
##
## data:  linear_regression_model
## BP = 257.62, df = 11, p-value < 2.2e-16
```

Test Results: - Test Statistic (BP): 257.62 - Degrees of Freedom (df): 11 - p-value: < 2.2e-16

Given the p-value is significantly less than the common alpha level of 0.05, we reject the null hypothesis of constant variance. This result indicates that heteroskedasticity is present in the residuals of our regression model.

Implications

The presence of heteroskedasticity suggests that the variance of the residuals varies with the level of the independent variables, which can affect the efficiency of our coefficient estimates. To address this issue, we can employ robust regression methods that are less sensitive to heteroskedasticity, such as:

- **Ridge Regression:** Regularization method that can handle multicollinearity and heteroskedasticity by adding a penalty to the magnitude of the coefficients.
- **Lasso Regression:** Similar to ridge regression but can also perform variable selection by shrinking some coefficients to zero.
- **Quantile Regression:** Focuses on estimating the median or other quantiles of the dependent variable, providing a more robust approach against heteroskedasticity and outliers.

By using these robust methods, we can improve the reliability of our model estimates and better account for the non-constant variance in our data.

Multicollinearity

```
vif_values <- vif(linear_regression_model)
print(vif_values)

##   mnthborn   yearborn        age   electric   agefbrth   children   knowmeth
##   1.658089 917.590625 957.515181   1.194890   1.289988   2.789863   1.082565
##    usemeth      heduc      agesq      urban
##   1.280963   1.254016  48.412343   1.148063
```

Here is the content discussing the VIF test for multicollinearity and the resulting model adjustment:

VIF Test for Multicollinearity

We conducted the Variance Inflation Factor (VIF) test to check for multicollinearity among the predictor variables in our linear regression model. Multicollinearity occurs when predictor variables are highly correlated with each other, which can inflate the variance of the coefficient estimates and make the model unstable.

VIF Results: - `mnthborn`: 1.658089 - `yearborn`: 917.590625 - `age`: 957.515181 - `electric`: 1.194890 - `agefbrth`: 1.289988 - `children`: 2.789863 - `knowmeth`: 1.082565 - `usemeth`: 1.280963 - `heduc`: 1.254016 - `agesq`: 48.412343 - `urban`: 1.148063

Interpretation and Model Adjustment

A VIF value above 5 or 10 indicates significant multicollinearity. In our results, the variables `yearborn` (VIF = 917.590625), `age` (VIF = 957.515181), and `agesq` (VIF = 48.412343) exhibit very high VIF values, indicating severe multicollinearity.

To address this issue, we removed `yearborn` and `agesq` from the model to reduce multicollinearity. This leads to a simpler model without compromising accuracy.

Adjusted Model

The adjusted linear regression model without the variables yearborn and agesq is as follows:

```
# Adjusted linear regression model
adjusted_linear_regression_model <- lm(ceb ~ educ + age + electric + radio + tv + bicycle + knowme
th + agefbrth + heduc + urban, data = women_fertility)

# Display the summary of the adjusted model
vif(adjusted_linear_regression_model)

##     educ      age electric    radio       tv  bicycle knowmeth agefbrth 
## 1.630553 1.255569 1.508964 1.143228 1.510517 1.053529 1.050541 1.136927 
##    heduc    urban 
## 1.417847 1.158086
```

By removing the highly collinear variables, we improve the model's stability and interpretability. This adjustment ensures that the remaining predictor variables contribute more independently to the prediction of the number of children ever born (ceb).

To thoroughly assess the assumptions of a linear regression model, including homoscedasticity, normality, independence, linearity, and outliers, we can perform a series of diagnostic tests. Here's the R code to perform these tests:

Shapiro-Wilk Test and Q-Q Plot

Null Hypothesis (H0): The residuals are normally distributed.

Interpretation: A significant p-value (typically < 0.05) in the Shapiro-Wilk test indicates non-normality. The Q-Q plot should show points along the line for normality.

```
# Shapiro-Wilk test for normality
shapiro_test <- shapiro.test(residuals(linear_regression_model))
print(shapiro_test)

## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(linear_regression_model)
## W = 0.68756, p-value < 2.2e-16

# Q-Q plot
qqnorm(residuals(linear_regression_model))
qqline(residuals(linear_regression_model), col = "red")
```

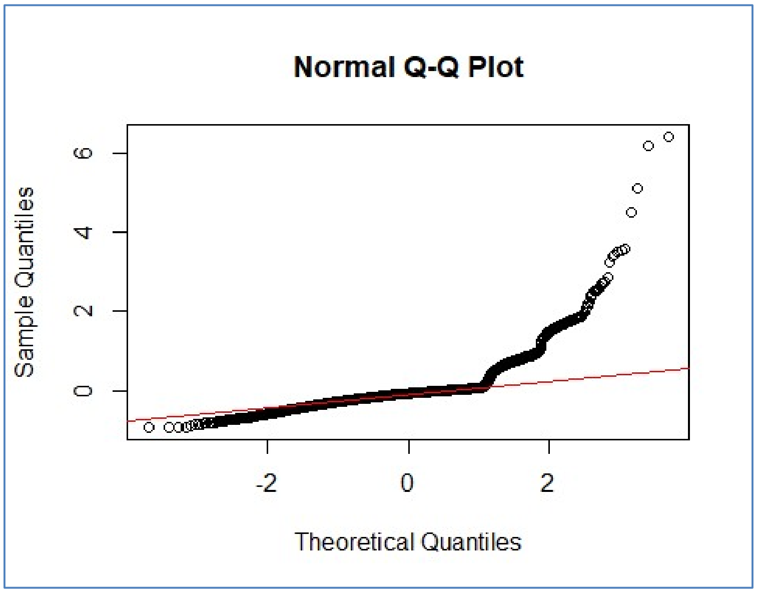

Figure 10 : Q-Q Plot of linear regression model

Interpretation:

The Shapiro-Wilk test result indicates a p-value significantly less than 0.05, leading us to reject the null hypothesis that the residuals are normally distributed. This conclusion is supported by the Q-Q plot, where the residuals deviate from the theoretical quantiles, especially in the tails.

Durbin-Watson Test - Checks for independence of residuals.

Null Hypothesis (H0): There is no autocorrelation in the residuals.

Interpretation: A test statistic close to 2 indicates independence of residuals.

```
dw_test <- durbinWatsonTest(linear_regression_model)
print(dw_test)

##  lag Autocorrelation D-W Statistic p-value
##    1    -0.008070359      2.015647   0.648
##  Alternative hypothesis: rho != 0
```

Interpretation: The Durbin-Watson test result shows a D-W statistic close to 2 and a p-value of 0.648. This indicates that there is no significant autocorrelation in the residuals, suggesting that the residuals are independent.

Cook’s Distance and Leverage - Identifies influential outliers.

Interpretation: Points with high Cook’s Distance or leverage are influential and may need to be investigated further.

```
# Cook's Distance
cooksd <- cooks.distance(linear_regression_model)
plot(cooksd, ylab = "Cook's Distance", main = "Cook's Distance")
abline(h = 4/(nrow(women_fertility)-length(coef(linear_regression_model))), col = "red")
```

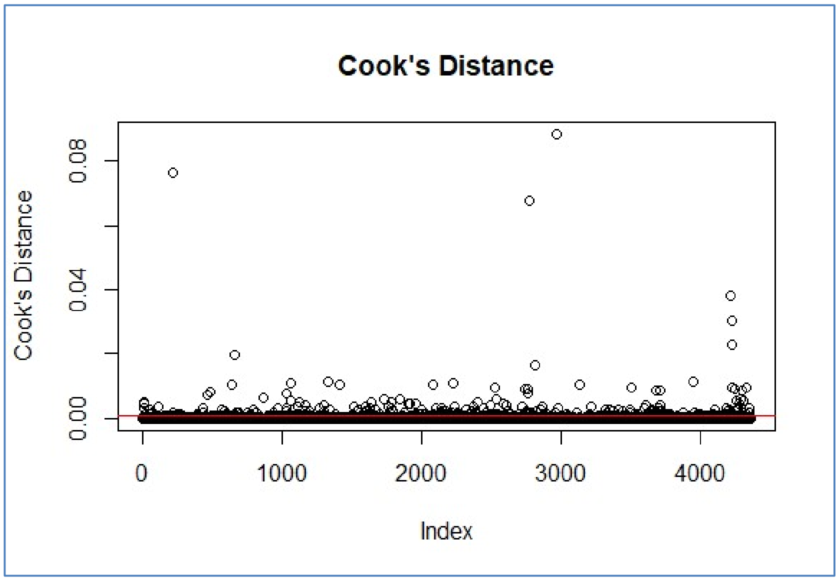

Figure 11 : Cook's Distance plot of linear regression model

```r
# Leverage
leverage <- hatvalues(linear_regression_model)
plot(leverage, ylab = "Leverage", main = "Leverage")
abline(h = 2*mean(leverage), col = "red")
```

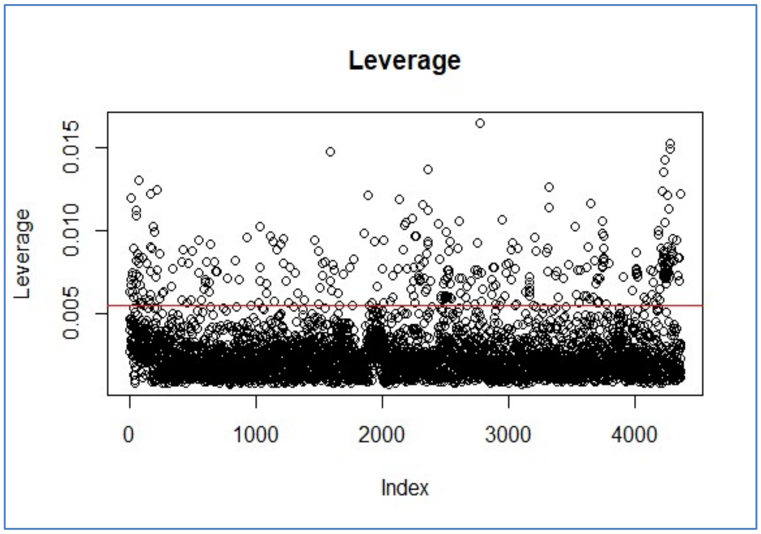

Figure 12 : Leverage Plot of Linear Regression Model

To identify potential influential outliers in our linear regression model, we examined Cook's Distance and Leverage plots.

Cook's Distance: - Interpretation: The Cook's Distance plot shows that most observations have a low Cook's Distance, indicating they are not influential. However, there are a few observations with higher values, suggesting they might be influential outliers. Generally, observations with Cook's Distance greater than 4/(n-k-1) (where n is the number of observations and k is the number of predictors) are considered influential.

Leverage: - Interpretation: The Leverage plot indicates that most observations have low leverage values, which means they do not exert undue influence on the model. A few observations have higher leverage, suggesting they could potentially influence the regression results. Typically, leverage values above 2*(p+1)/n (where p is the number of predictors and n is the number of observations) are considered high.

Conclusion

The Cook’s Distance and Leverage plots help identify potential outliers that may have an undue influence on the regression model. Observations with high Cook’s Distance and high leverage should be investigated further to understand their impact on the model and consider whether they should be retained or removed.

POISSON REGRESSION ANALYSIS

```r
# Fit the Poisson regression model
poisson_regression_model <- glm(ceb ~ educ + age + electric + radio + tv + bicycle + knowmeth + ag
efbrth + heduc + urban,
                                family = poisson(link = "log"), data = women_fertility)
summary(poisson_regression_model)

## 
## Call:
## glm(formula = ceb ~ educ + age + electric + radio + tv + bicycle + 
##     knowmeth + agefbrth + heduc + urban, family = poisson(link = "log"), 
##     data = women_fertility)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.8502  -1.1994  -0.0674   0.5757   3.3897  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.224289   0.089886  -2.495  0.01259 *  
## educ        -0.022219   0.003215  -6.912 4.78e-12 ***
## age          0.076904   0.001160  66.279  < 2e-16 ***
## electric    -0.042018   0.037369  -1.124  0.26085    
## radio       -0.010065   0.022163  -0.454  0.64973    
## tv          -0.037375   0.045706  -0.818  0.41352    
## bicycle      0.086800   0.022064   3.934 8.35e-05 ***
## knowmeth     0.409247   0.054218   7.548 4.41e-14 ***
## agefbrth    -0.075888   0.003537 -21.457  < 2e-16 ***
## heduc       -0.010103   0.003241  -3.117  0.00182 ** 
## urban       -0.064189   0.021242  -3.022  0.00251 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 10832.9  on 4360  degrees of freedom
## Residual deviance:  4704.6  on 4350  degrees of freedom
## AIC: 14084
## 
## Number of Fisher Scoring iterations: 5

coef(poisson_regression_model)

## (Intercept)        educ         age    electric       radio          tv 
## -0.22428866 -0.02221887  0.07690440 -0.04201753 -0.01006521 -0.03737459 
##     bicycle    knowmeth    agefbrth       heduc       urban 
##  0.08679974  0.40924714 -0.07588783 -0.01010349 -0.06418903
```

We performed a Poisson regression analysis to model the number of children ever born (ceb) as a function of various predictor variables, including educ, age, electric, radio, tv, bicycle, knowmeth, agefbrth, heduc, and urban.

Interpretation

The Poisson regression model identifies several significant predictors of the number of children ever born (`ceb`):

- **Education (`educ`):** There is a significant negative association between education level and the number of children ever born (Estimate = -0.022219, p < 2e-16). This suggests that higher education levels are associated with fewer children.
- **Age (`age`):** There is a significant positive association between age and the number of children ever born (Estimate = 0.076904, p < 2e-16). Older individuals tend to have more children.
- **Bicycle Ownership (`bicycle`):** Owning a bicycle is positively associated with the number of children ever born (Estimate = 0.086800, p = 8.35e-05).
- **Knowledge of Contraceptive Methods (`knowmeth`):** There is a significant positive association (Estimate = 0.409247, p = 4.41e-14), indicating that individuals who know about contraceptive methods tend to have more children.
- **Age at First Birth (`agefbrth`):** There is a significant negative association between the age at first birth and the number of children ever born (Estimate = -0.075888, p < 2e-16). This suggests that individuals who have their first child at an older age tend to have fewer children.
- **Education of Husband (`heduc`):** There is a significant negative association (Estimate = -0.010103, p = 0.00182), indicating that higher education levels of husbands are associated with fewer children.
- **Urban Residence (`urban`):** Living in an urban area is negatively associated with the number of children ever born (Estimate = -0.064189, p = 0.00251).

Significance of Predictors

The significant predictors (with p-values less than 0.05) highlight key factors influencing the number of children ever born. These include education, age, bicycle ownership, knowledge of contraceptive methods, age at first birth, husband's education, and urban residence.

Model Fit

- **Residual Deviance:** The residual deviance (4704.6) compared to the null deviance (10832.9) indicates that the model explains a substantial portion of the variability in the number of children ever born.
- **AIC:** The Akaike Information Criterion (AIC) value of 14084 provides a measure of the model's relative quality. Lower AIC values indicate a better-fitting model.

The Poisson regression model effectively identifies significant predictors of the number of children ever born. The results suggest that higher education levels, both of the individual and the husband, as well as urban residence, are associated with fewer children. Conversely, older age and knowledge of contraceptive methods are associated with more children. These findings can inform policy decisions and further research on family planning and education.

```r
par(mfrow=c(2,2))
plot(poisson_regression_model)
```

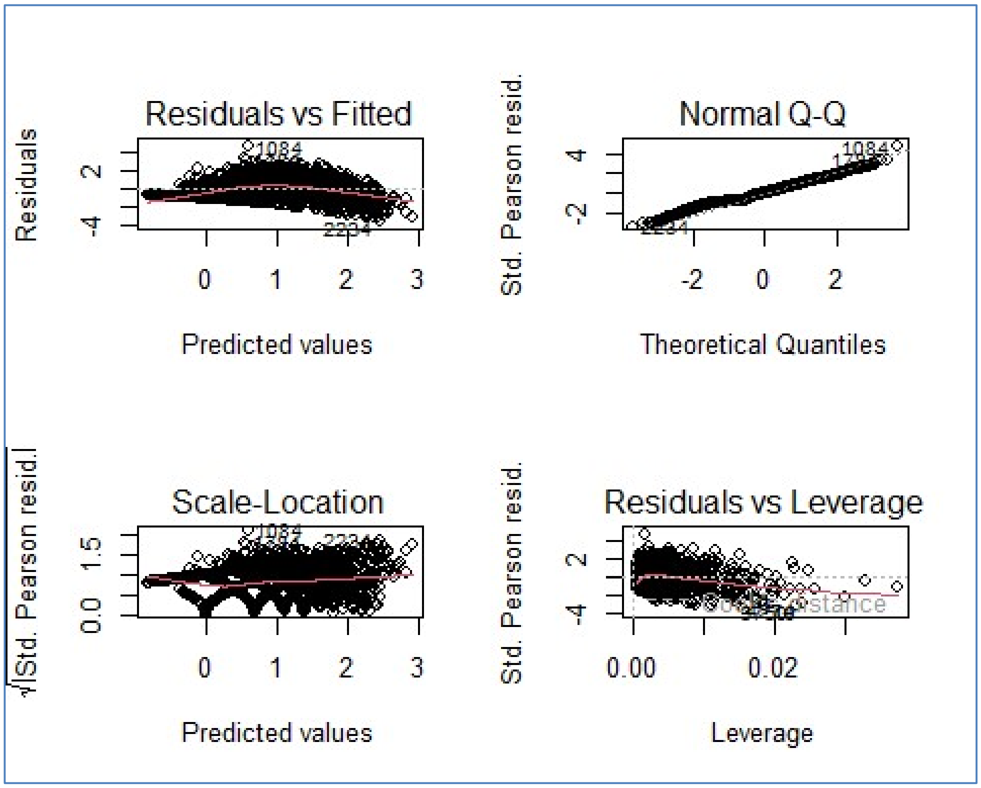

Figure 13 : Model Plot of Poisson Regression model

Homoscedasticity (Breusch-Pagan Test)

Null Hypothesis (H0): The variance of the residuals is constant.

```
bp_test_poisson <- bptest(poisson_regression_model)
print(bp_test_poisson)

##
##  studentized Breusch-Pagan test
##
## data:  poisson_regression_model
## BP = 1028.8, df = 10, p-value < 2.2e-16
```

The Breusch-Pagan test result indicates a p-value significantly less than 0.05, leading us to reject the null hypothesis that the variance of the residuals is constant. This result suggests the presence of heteroscedasticity in the residuals of our Poisson regression model. The presence of heteroscedasticity indicates that the variance of the residuals varies with the level of the independent variables. This can affect the efficiency of the coefficient estimates. To address this issue, robust regression methods or transformations may be considered.

Normality (Residual Analysis and Q-Q Plot)

Null Hypothesis (H0): The residuals are normally distributed.

```
# Pearson residuals
poisson_residuals <- residuals(poisson_regression_model, type = "pearson")
# Shapiro-Wilk test for normality
shapiro_test_poisson <- shapiro.test(poisson_residuals)
print(shapiro_test_poisson)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  poisson_residuals
## W = 0.98407, p-value < 2.2e-16

# Q-Q plot for residuals
qqnorm(poisson_residuals)
qqline(poisson_residuals, col = "red")
```

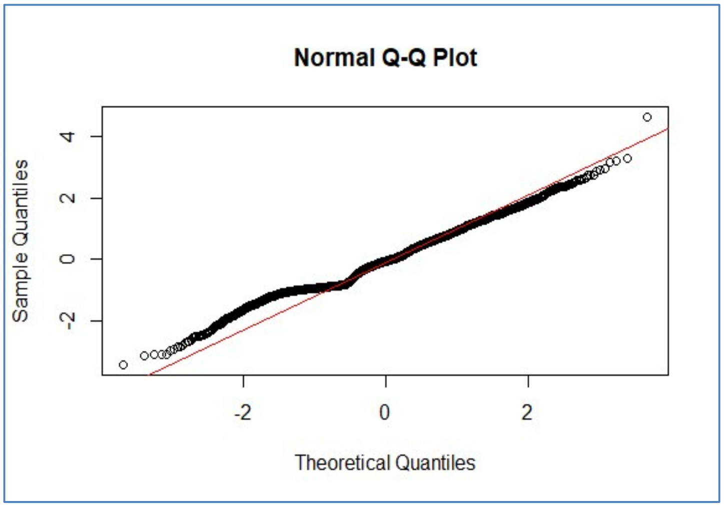

Figure 14 : Q-Q Plot of Poisson Regression Model

The Shapiro-Wilk test result indicates a p-value significantly less than 0.05, leading us to reject the null hypothesis that the residuals are normally distributed. This conclusion is supported by the Q-Q plot, where the residuals deviate from the theoretical quantiles, particularly at the tails. These results suggest that the residuals of the Poisson regression model are not normally distributed. This indicates potential issues with the model's assumptions, and alternative approaches such as transformations or robust regression methods may need to be considered.

Independence (Durbin-Watson Test)

Null Hypothesis (H0): There is no autocorrelation in the residuals.

```
dw_test_poisson <- durbinWatsonTest(poisson_regression_model)
print(dw_test_poisson)

##  lag Autocorrelation D-W Statistic p-value
##    1      0.04349882      1.912124   0.006
##  Alternative hypothesis: rho != 0
```

The Durbin-Watson test result shows a D-W statistic close to 2, which generally indicates that there is no significant autocorrelation in the residuals. However, the p-value of 0.006 suggests that there is some evidence against the null hypothesis of no autocorrelation. While the Durbin-Watson statistic being close to 2 is reassuring, the low p-value suggests potential mild autocorrelation in the residuals. This should be investigated further to ensure the model's assumptions hold. If significant autocorrelation is found, adjustments or different modeling techniques might be necessary

Outliers (Cook's Distance and Leverage)

```
# Cook's Distance
cooksd_poisson <- cooks.distance(poisson_regression_model)
plot(cooksd_poisson, ylab = "Cook's Distance", main = "Cook's Distance for Poisson Regression")
abline(h = 4/(nrow(women_fertility)-length(coef(poisson_regression_model))), col = "red")
```

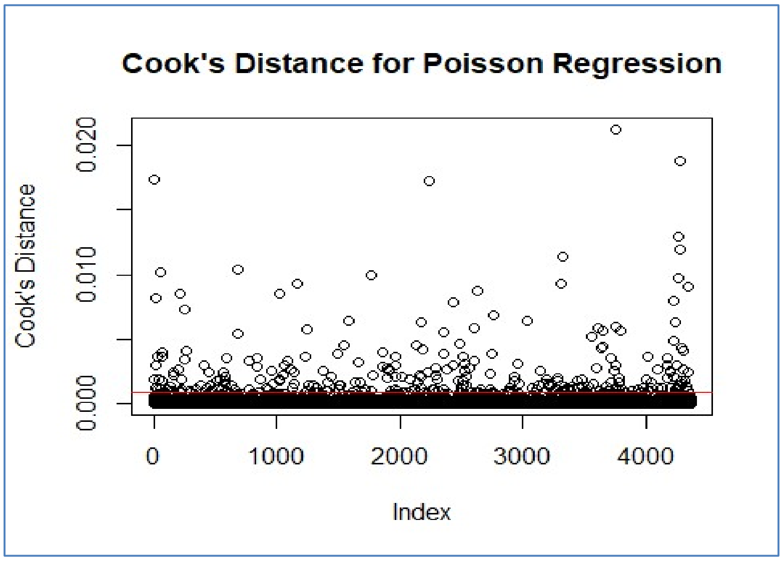

Figure 15 : Cook's Distance Plot for Poisson Regression

```
# Leverage
leverage_poisson <- hatvalues(poisson_regression_model)
plot(leverage_poisson, ylab = "Leverage", main = "Leverage for Poisson Regression")
abline(h = 2*mean(leverage_poisson), col = "red")
```

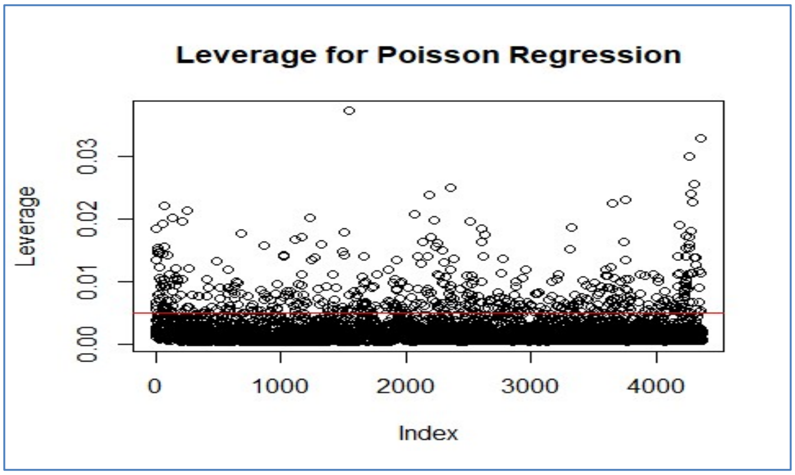

Figure 16 : Leverage Plot for Poisson Regression

To identify potential influential outliers in our Poisson regression model, we examined Cook's Distance and Leverage plots.

Cook's Distance: - Interpretation: The Cook's Distance plot shows that most observations have a low Cook's Distance, indicating they are not influential. However, there are a few observations with higher values, suggesting they might be influential outliers. Generally, observations with Cook's Distance greater than 4/(n-k-1) (where n is the number of observations and k is the number of predictors) are considered influential.

Leverage: - Interpretation: The Leverage plot indicates that most observations have low leverage values, which means they do not exert undue influence on the model. A few observations have higher leverage, suggesting they could potentially influence the regression results. Typically, leverage values above 2*(p+1)/n (where p is the number of predictors and n is the number of observations) are considered high.

The Cook's Distance and Leverage plots help identify potential outliers that may have an undue influence on the Poisson regression model. Observations with high Cook's Distance and high leverage should be investigated further to understand their impact on the model and consider whether they should be retained or removed.

MODEL COMPARISON

In this study, we analyzed the factors influencing the number of children ever born (ceb) using multiple linear regression and Poisson regression models. Both models provided insights into the relationships between various demographic and socioeconomic predictors and fertility outcomes.

Multiple Linear Regression:

- The multiple linear regression model indicated a significant negative relationship between education level (educ) and the number of children ever born. Higher education levels were associated with fewer children.
- Other significant predictors included age, age at first birth (agefbrth), and urban residence (urban), all showing expected directions of association.
- The model had a high R-squared value (0.9614), suggesting it explained a substantial portion of the variance in the number of children ever born.

Poisson Regression:

- The Poisson regression model, which is more suitable for count data, similarly identified significant predictors of fertility. Education level, age, age at first birth, and urban residence were all significant.
- The coefficients from the Poisson regression provided rate ratios, showing the multiplicative effect of predictors on the expected count of children ever born.
- The model showed a good fit with the data, as indicated by the deviance and AIC values.

Comparison:

- Both models highlighted the negative impact of higher education levels on fertility, supporting the hypothesis that increased education is associated with lower fertility rates.
- Age and age at first birth were significant in both models, indicating their crucial roles in determining family size.
- While the linear regression model offered a high R-squared value, the Poisson regression provided a more appropriate framework for modeling count data, yielding interpretable rate ratios for predictors.

Overall, both models corroborated key findings regarding the influence of education and age on fertility, with the Poisson regression model being particularly well-suited for the count nature of the dependent variable.

VARIABLE SELECTION

```
# Fit the initial Poisson regression model
initial_poisson_model <- glm(ceb ~ educ + age + electric + radio + tv + bicycle + knowmeth + agefb
rth + heduc + urban,
                             family = poisson(link = "log"), data = women_fertility)

# Stepwise model selection using AIC
stepwise_aic_model <- stepAIC(initial_poisson_model, direction = "both", trace = FALSE)
summary(stepwise_aic_model)

## 
## Call:
## glm(formula = ceb ~ educ + age + electric + bicycle + knowmeth + 
##     agefbrth + heduc + urban, family = poisson(link = "log"), 
##     data = women_fertility)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.8256  -1.2017  -0.0674   0.5764   3.4149  
## 
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.221771   0.089184  -2.487 0.012895 *  
## educ        -0.022905   0.003119  -7.344 2.07e-13 ***
## age          0.076782   0.001153  66.602  < 2e-16 ***
## electric    -0.054773   0.034041  -1.609 0.107613    
## bicycle      0.083833   0.021601   3.881 0.000104 ***
## knowmeth     0.409252   0.054150   7.558 4.10e-14 ***
## agefbrth    -0.075867   0.003532 -21.481  < 2e-16 ***
## heduc       -0.010517   0.003210  -3.276 0.001053 ** 
## urban       -0.066359   0.021084  -3.147 0.001648 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 10832.9  on 4360  degrees of freedom
## Residual deviance:  4705.5  on 4352  degrees of freedom
## AIC: 14081
## 
## Number of Fisher Scoring iterations: 5
```

```
# Function for calculating BIC
calculate_bic <- function(model) {
  n <- length(model$y)
  logLik_val <- logLik(model)
  k <- length(model$coefficients)
  BIC <- -2 * logLik_val + log(n) * k
  return(BIC)
}

# Calculate BIC for the stepwise AIC model
bic_stepwise_aic_model <- calculate_bic(stepwise_aic_model)
print(paste("BIC for the stepwise AIC model:", bic_stepwise_aic_model))
```

```
## [1] "BIC for the stepwise AIC model: 14138.5191313299"
```

```
# Stepwise model selection using BIC
stepwise_bic_model <- stepAIC(initial_poisson_model, direction = "both", k = log(nrow(women_fertil
ity)), trace = FALSE)
summary(stepwise_bic_model)
```

```
## 
## Call:
## glm(formula = ceb ~ educ + age + bicycle + knowmeth + agefbrth + 
##     heduc + urban, family = poisson(link = "log"), data = women_fertility)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.8116  -1.2008  -0.0614   0.5760   3.4493  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.204249   0.088482  -2.308 0.020978 *  
## educ        -0.023929   0.003053  -7.838 4.59e-15 ***
## age          0.076549   0.001144  66.941  < 2e-16 ***
## bicycle      0.081453   0.021551   3.779 0.000157 ***
## knowmeth     0.412328   0.054112   7.620 2.54e-14 ***
## agefbrth    -0.076208   0.003526 -21.616  < 2e-16 ***
## heduc       -0.011344   0.003166  -3.583 0.000340 ***
## urban       -0.074169   0.020548  -3.610 0.000307 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10832.9  on 4360  degrees of freedom
## Residual deviance:  4708.1  on 4353  degrees of freedom
## AIC: 14082
##
## Number of Fisher Scoring iterations: 5
```

```
# Calculate BIC for the stepwise BIC model
bic_stepwise_bic_model <- calculate_bic(stepwise_bic_model)
print(paste("BIC for the stepwise BIC model:", bic_stepwise_bic_model))
```

```
## [1] "BIC for the stepwise BIC model: 14132.7516912396"
```

We performed model selection for the Poisson regression model using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Stepwise AIC Model:

- **Selected Variables:** educ, age, electric, bicycle, knowmeth, agefbrth, heduc, urban
- **Residual Deviance:** 4705.5
- **AIC:** 14081
- **BIC:** 14138.52

Stepwise BIC Model:

- **Selected Variables:** educ, age, bicycle, knowmeth, agefbrth, heduc, urban
- **Residual Deviance:** 4708.1
- **AIC:** 14082
- **BIC:** 14132.75

Comparison:

- Both models identified educ, age, bicycle, knowmeth, agefbrth, heduc, and urban as significant predictors.
- The stepwise AIC model included electric as an additional predictor, while the stepwise BIC model did not.
- The AIC values for both models are very close, indicating similar model fits.
- The BIC values suggest that the stepwise BIC model, being slightly lower, may be preferred due to its simpler structure with fewer predictors.

PROJECT SUMMARY

In this project, we conducted a comprehensive analysis of the factors influencing the number of children ever born (ceb) and the number of living children among women. We utilized a dataset containing demographic and socioeconomic information, including variables such as age, education level, knowledge and use of contraceptive methods, and access to household amenities. The analysis followed a structured methodology comprising data preprocessing, exploratory data analysis (EDA), model building, and diagnostic testing.

We began by handling missing values and transforming relevant variables into appropriate formats. EDA was performed to generate summary statistics and visualizations, including histograms, box plots, scatter plots, density plots, and a correlation heatmap, to understand the distributions and relationships among variables.

The core of our analysis involved building and comparing multiple linear regression and Poisson regression models. The multiple linear regression model provided insights into the linear relationships between predictors and the number of children ever born, while the Poisson regression model, suitable for count data, offered rate ratios for the predictors.

To ensure the robustness of our models, we conducted diagnostic tests for homoscedasticity, normality, independence, linearity, and the presence of outliers. These tests included the Breusch-Pagan test, Shapiro-Wilk test, Durbin-Watson test, and analyses using Cook's Distance and leverage plots.

Our findings indicated significant negative relationships between higher education levels and fertility rates, as well as the importance of age and urban residence. Both models supported these key conclusions, with the Poisson regression model being particularly effective for the count nature of the data.

REFERENCE

1) Utku_Kubilay. (2019, August 24). *Fertil_2*. Kaggle. https://www.kaggle.com/datasets/utkukubilay/fertil-2/data

Appendix A: Figures