1 Question 1

1.1 Part 1: Expected Correlation between Class Size and Test Scores

1.1.1 Hypothesis and Rationale

The hypothesis states that smaller class sizes contribute to enhanced student performance. This proposition is grounded in educational theories that advocate for more personalized teaching approaches, which are more feasible in smaller classes. The presumed mechanisms include improved student-teacher interactions, more targeted feedback, and increased engagement and participation from students. Empirical studies, such as those by Hattie [1] and Blatchford et al. [2], have also suggested potential benefits of smaller class sizes on student learning outcomes, although the evidence is mixed and context-dependent.

1.1.2 Literature Review

A thorough review of the literature reveals that the relationship between class size and academic performance is a subject of ongoing debate. Some studies indicate that reduced class sizes can lead to significant improvements in test scores, particularly for younger and disadvantaged students (Krueger [3]). Other research, however, suggests that the impact of class size is less pronounced when compared to other factors such as teacher quality and curriculum rigor (Hanushek [4]). This contradictory evidence underscores the complexity of isolating the effects of class size from other educational variables.

1.1.3 Anticipated Correlation

In light of the hypothesis and existing literature, we would anticipate a negative correlation between class size and test scores. This implies that, ceteris paribus, smaller classes are associated with higher standardized test scores. It is important to note, however, that the strength of this correlation can be influenced by a variety of confounding factors that also need to be accounted for in a comprehensive analysis.

1.1.4 Statistical Expectations

To empirically test this hypothesis, a statistical analysis on a dataset containing thousands of observations from different schools and classes would be conducted. The Pearson correlation coefficient, r, would serve as the primary statistical measure to evaluate the strength and direction of the linear relationship between class size and test scores. It is expected that, if the hypothesis is correct, r would be significantly less than zero, indicating an inverse relationship. However, it is crucial to recognize that correlation does not reveal the nature of the relationship, and a robust analytical approach would necessitate the application of regression techniques, possibly including hierarchical linear modeling to control for nested data structures (students within classes within schools).

1.1.5 Methodological Considerations

When designing the study, it would be essential to consider potential biases and confounding variables. Selection bias, measurement error, and omitted variable bias could

all lead to incorrect inferences about the correlation between class size and test scores. As such, the study design would involve careful consideration of these factors, perhaps by incorporating propensity score matching or fixed-effects modeling to mitigate the impact of confounding variables. Furthermore, the inclusion of a rich set of covariates that capture student background, teacher characteristics, and school-level variables would be necessary to reduce the likelihood of spurious correlations.

1.2 Part 2: Analysis of Correlation and Causation

1.2.1 Understanding Correlation

While a negative correlation coefficient suggests an inverse relationship, it is purely a measure of the strength and direction of a linear association between two variables. This statistical measure does not convey any information about the potential mechanisms linking the variables nor does it control for the influence of other variables. It is crucial to distinguish between correlation and causation, as correlation alone does not confirm a causal effect. This distinction is fundamental in statistics and is critical for interpreting empirical findings in research.

1.2.2 Causation Criteria

The Bradford Hill criteria for causation could serve as a framework to evaluate whether a causal relationship might exist between class size and student performance. These criteria include considerations such as strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy. Particularly, temporality is essential – the cause must be shown to precede the effect. For example, if smaller class sizes are indeed causing better test scores, changes in class sizes should consistently lead to changes in performance, and not vice versa.

1.2.3 Interpreting Negative Correlation

When interpreting a negative correlation between class size and test scores, one must consider alternative explanations. Confounding variables that are not included in the analysis could be responsible for the observed relationship. For instance, socioeconomic factors or educational policies that are correlated with both class size and test scores may confound the relationship. Moreover, latent variables, such as parental involvement and intrinsic student motivation, might also play a role. Without accounting for these factors, the observed correlation could lead to erroneous conclusions about the effect of class size on student performance.

1.2.4 Methodological Approaches to Establishing Causation

Establishing causation in non-experimental settings requires careful methodological design. One approach is the use of natural experiments where external events or policies cause variation in class size independently of other factors. Another approach is instrumental variable analysis, which requires finding a variable that affects class size but is not directly related to test scores, to serve as an instrument in the analysis. Propensity score matching can also be used to create a quasi-experimental comparison between students in different class sizes who are matched on other observed characteristics. Ultimately, while

these methods can strengthen causal inferences, they have limitations and assumptions that must be carefully considered.

1.3 Conclusion

The distinction between correlation and causation is a cornerstone of empirical research. While a negative correlation between class size and test scores is anticipated based on the hypothesis, establishing a causal relationship requires a much more nuanced approach than simple correlation analysis. Only through the application of rigorous and appropriate econometric methods can we hope to isolate the true effect of class size on student performance. Therefore, while the data may suggest a negative correlation, asserting causality is contingent upon a thorough and methodologically sound analysis that accounts for all other potential influences.

2 Question 2

2.1 Part 1: Descriptive Statistic

- Minimum number of children: The smallest family size in the dataset is zero children. This indicates that there are women with no children.
- Maximum number of children: The largest family size recorded is thirteen children.
- Average number of children: On average, women in the sample have approximately 2.2678 children. This suggests a moderate average family size among the sample.

The following R code was used to calculate the descriptive statistics in RStudio:

Listing 1: Descritive Statistics Code

```
# Load data from dataset
load(fertil2.RData)

# Calculation of minimum number of children
min_children <- min(data$children)
# Calculation of maximum number of children
max_children <- max(data$children)
# Calculation of average number of children
average_children <- mean(data$children)

# Output the results
min_children
max_children
average_children
```

2.2 Part 2: Husband's Education

- The average years of the husband's education is 5.144683, which is approximately 5 years. This reflects a relatively low level of educational attainment.
- Number of observations: 1,956. This count indicates the number of non-missing data entries used to compute the average husband's years of education, out of a

total of 4,361 entries in the dataset. This means that a significant portion of the data, approximately 2,405 entries, which is about 55% of the total, did not have information on the husband's education. Hence, it is important to highlight potential biases or limitations in the dataset due to the substantial amount of missing data, which might reflect disparities in data availability or recording practices.

The R code used in RStudio to compute the above statistics is as follows:

Listing 2: Husband's Education Statistics Code

```
1
2  # Calculate the total number of entries in the dataset
3  total_entries <- nrow(data)
4
5  # Compute the average years of husband's education, excluding missing
       values
6  average_year <- mean(data$heduc, na.rm = TRUE)
7
8  # Calculate the number of non-missing data entries for husband's
       education
9  observations <- sum(!is.na(data$heduc))
10
11 # Output the results
12 total_entries
13 average_year
14 observations
```

Explanation of the R Code:

- `nrow(data)` calculates the total number of rows (entries) in your dataset.
- `mean(data$heduc, na.rm = TRUE)` computes the average years of education for husbands, ignoring any missing values (`na.rm = TRUE` ensures that `NA` values are removed from the calculation).
- `sum(!is.na(data$heduc))` counts the number of entries that do not have missing data for the husband's education (`!is.na(data$heduc)` creates a logical vector where `TRUE` corresponds to non-missing values).

2.3 Part 3: Visual Analysis of Relationship between Children and Husband's Education

The plot reveals a broad distribution of data points across different education levels, suggesting a diverse sample in terms of the husbands' educational background. Notably, there is a concentration of points at the lower education levels, from 0 to around 10 years. Similar to the average found in Part 2, the concentration shows that within the sample, a significant portion of husbands have relatively lower education levels.

As for number of children, many data points are clustered around the lower numbers. The number of children tend to decrease as the education level increases, although this is not a strict pattern. There are a few outliers at higher numbers of children. Meanwhile, they do not show a clear relationship with the husband's education level. These could be due to other socio-cultural or even economical factors not being captured by the husband's

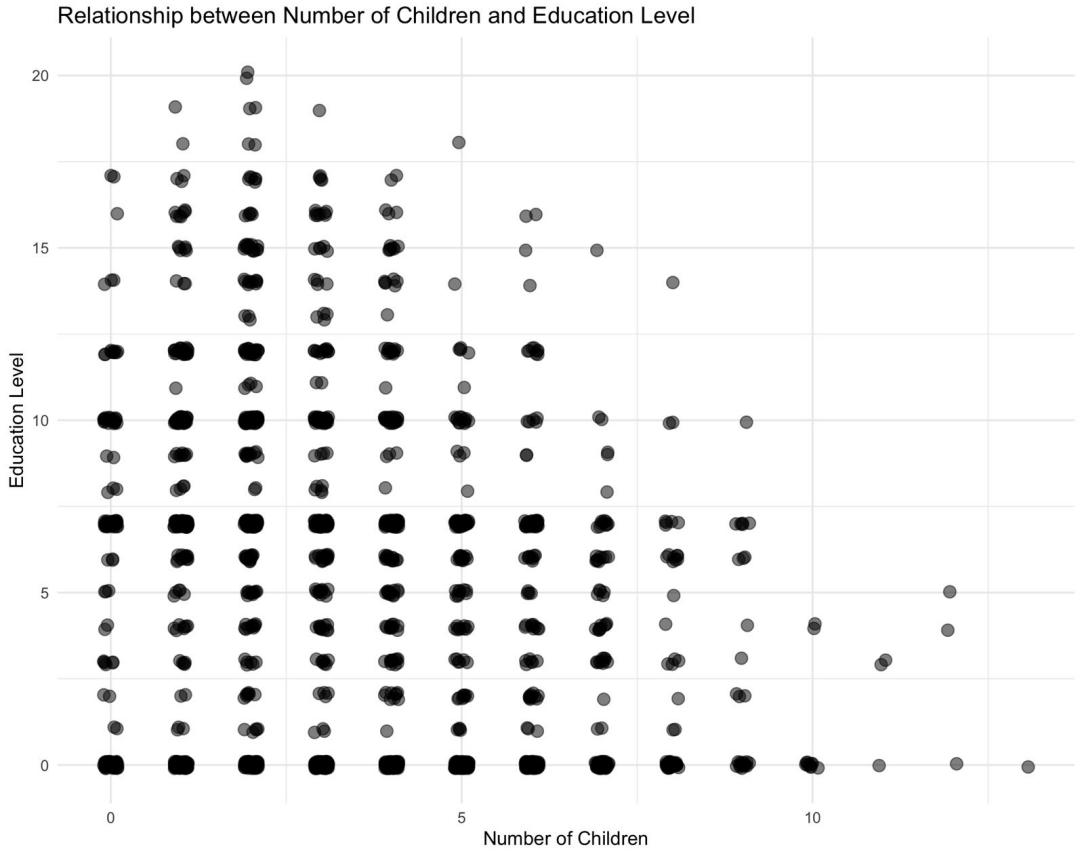

Figure 1: scatterplot

education variable.

Overall, there is no clear linear relationship. The plot does not show a strong or consistent correlation between the husband's education and the number of children. However, there seems to be a slight decrease in the number of children as the husbands' education increases, particularly beyond 10 years of education.

In summary, while the plot suggests there may be some relationship between husband's education and family size, the relationship is not strongly linear and is likely influenced by a variety of other factors. To draw more definitive conclusions, additional analysis, perhaps controlling for other variables would be necessary to understand the dynamics better.

The following R code was executed in RStudio to generate the scatter plot:

Listing 3: Scatterplot Code

```
# Load required package
install.packages("ggplot2")
library(ggplot2)

# Filter out entries with missing husband's education data
clean_data <- data[!is.na(data$heduc), ]

# Generate a scatter plot using ggplot2 with jitter to prevent
```

```
    overplotting
p <- ggplot(clean_data, aes(x = children, y = heduc)) +
  geom_jitter(alpha = 0.5, size = 3, width = 0.1, height = 0.1) +
  labs(title = "Relationship between Number of Children and Education
     Level",
       x = "Number of Children",
       y = "Education Level") +
  theme_minimal()

# Display the plot
print(p)
```

Explanation of the R Code:

- The `ggplot2` package is installed and loaded to create advanced graphical designs.
- Entries with missing husband's education data (`NA` values) are filtered out to prepare for analysis.
- A scatter plot is created with `geom_jitter`, which adds a small amount of random variation to the location of each point to make overlapping points more visible.
- The `alpha` parameter controls the transparency of the points, making it easier to perceive the density of the points.
- The plot is given a title and labels for the x and y axes for clarity.
- `theme_minimal()` is used to apply a minimalistic theme to the plot.

2.4 Part 4: Availability of Electricity

The percentage of home with electricity is approximately 14.02%. This low percentage highlights the developmental challenges the country faced during that period. Notably, the electrical infrastructure in 1988 was not widespread, indicating a gap that the country has been working to close in its path to development and modernization. As referenced in chapter 6 of Africa World Press Inc. [6], Botswana has been facing challenges in its infrastructure which posed hurdles in bridging the gap towards modernization. Limited access to electricity inhibits the adoption of modern technologies and affects the overall quality of life for its citizens. This lack of infrastructure has also an impact on other sectors such as education where reliable infrastructure is crucial. Addressing these issues is essential for Botswana to fully realize its potential for development and be able to enhance its competitiveness in the global economy.

The following R code was executed in RStudio to find the percentage of women with electricity in the home:

Listing 4: Availability of Electricity Code

```
percentage_with_electricity <- mean(data$electric, na.rm = TRUE) * 100
```

2.5 Part 5: Comparison of Family Size by Electricity Availability

- The average number of children in homes without electricity is about 2.33. This number is slightly higher, suggesting that families without access to electricity tend

to be larger. Home without electricity might be more common in rural areas where traditional values favors larger families. In these areas, economic activities might be more focused on labor due to the lack of electricity. Hence, larger families where children can contribute to the household income or farming activities might be beneficial. Moreover, lack of electricity could potentially affect education and family planning method, resulting to larger families. We have seen such pattern in Part 3, where husband with lower education level tend to have more children.

- The average number of children in home with electricity is about 1.90. Slightly lower on average, which might suggest that access to electricity is associated with smaller family sizes, potentially due to better access to information and resources that influence family planning decisions. Notably, access to electricity is often linked to better overall living conditions, including access to urbanized areas, health services and family planning. It also allows for better access to information such as media, which can influence personal choices about family size and access to educational content regarding family planning.

Overall, the difference in average number of children between households with and without electricity is not significant. The difference might not solely be due to the availability of electricity. It could also be indicative of several associated factors mentioned above, such as economic status, education levels, and access to information. Notably, when considering these factors, it's also crucial to recognize that correlation does not imply causation. While the data shows an association between electricity and family size, it does not prove that the presence of electricity directly causes smaller family sizes. Moreover, the averages found can be a reflection of the development stage of Botswana's economy in 1988, where infrastructure was still being developed.

The following R code was executed in RStudio to compute the average of children for those without electricity and do the same for those with electricity:

Listing 5: Analyzing Electricity Impact on Family Size Code

```
# Calculate the average number of children for women without
    electricity
average_children_no_electricity <- mean(data$children[data$electric ==
    0], na.rm = TRUE)

# Calculate the average number of children for women with electricity
average_children_with_electricity <- mean(data$children[data$electric
    == 1], na.rm = TRUE)
```

2.6 Part 6: Comparison of Family Size by Electricity Availability

The regression model explores the influence of electricity on family size in Botswana. With an estimated coefficient of -0.4292 for the electric variable, the model suggests a significant negative association between the presence of electricity in a home and the number of children. This result indicates that homes with electricity tend to have, on average, approximately 0.429 fewer children compared to homes without electricity.

On the other hand, the intercept of the regression, 2.3273, gives the estimated number of children in homes without electricity. This indicates the baseline number of children for the comparison group in the absence of the independent variable's effect.

Moreover, the strength of the association is underlined by the t-statistic of -4.437 and a p-value that is highly significant, which indicates that the likelihood of this association being due to chance is extremely low. The significance levels, marked by asterisks, confirm the robustness of these estimates.

However, the R-squared value at 0.004499 indicates that only about 0.45% of the variation in family size across households is accounted for by electricity availability. This low explanatory power suggests that there are other factors that could affect the number of children, which are not included in this model. Notably, factors such as access to education and information, economic conditions, and cultural practices could also play significant role.

The residual standard error (RSE) of 2.217 reflects the typical distance that the data points fall from the regression line. In the context of this model, it means that the actual number of children in a household can typically deviate by about 2.217 from the number predicted by the model based on electricity availability alone.

Given the model's low R-squared value and relatively high residual standard error, the results should be interpreted cautiously. While statistically significant, the practical relevance of electricity availability as a sole predictor of family size is limited. A comprehensive approach that includes a broader set of variables might provide a more nuanced understanding of the factors influencing family size.

	Dependent variable:
	children
electric	-0.429***
	(0.097)
Constant	2.328***
	(0.036)
Observations	4,358
R2	0.004
Adjusted R2	0.004
Residual Std. Error	2.217 (df = 4356)
F Statistic	19.686*** (df = 1; 4356)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 2: Regression Output

The following R code was executed in RStudio to compute the linear regression using the Ordinary Least Squares (OLS) method:

Listing 6: Regression Code

```
# Estimate the linear regression model
model <- lm(children ~ electric, data = data)

#We use the library that was shown in the live-session
install.packages("stargazer")
library(stargazer)
stargazer(model,type="text")
```

2.7 Part 7: Causal Implications

The simple regression reported does not necessarily capture a causal relationship for several reasons:

- Omitted Variable Bias: There may be other variables that are correlated with both electricity and the number of children that are not included in the regression. If such variables are left out, the estimated coefficient on electric might be capturing those effects instead of the true effect of electricity on the number of children.
- Reverse Causality: The direction of causality might be questionable. It is possible that having more children leads to different living conditions, potentially affecting whether a household has electricity.
- Measurement Error: If there is any error in how electric is measured, it could bias the results.
- Selection Bias: The sample could be non-random. For instance, the dataset might overrepresent certain types of households, such as urban over rural, which might have different access to electricity and different fertility patterns.
- Unobserved Heterogeneity: There may be unobserved factors specific to each household that affect both the likelihood of having electricity and the number of children, such as preferences, which the model does not account for.

For these reasons, while the regression can provide evidence of an association between electricity presence and the number of children, it does not prove causation. Establishing causality would require a more sophisticated analysis, possibly involving instrumental variables, natural experiments, or randomized controlled trials that can address these issues

3 Question 3

Ordinary Least Squares (OLS) regression is a staple method in statistical modeling for estimating the parameters of a linear relationship between variables. One of the method's

key assumptions is that the residuals, which are the differences between observed and estimated values, should not be correlated with the explanatory variables. Here, we further explore the implications of this assumption and the potential benefits of introducing polynomial terms such as x^2 into the regression model.

3.1 OLS Residuals and Correlation with Explanatory Variables

The OLS estimation procedure ensures that the residuals $\hat{u}$ are orthogonal to the space spanned by the explanatory variables. Formally, this is represented as:

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \tag{1}$$

This mathematical condition implies that there is no linear association between the residuals and the included explanatory variables. However, it does not preclude the existence of a nonlinear association.

3.2 The Role of Polynomial Terms in Modeling

The consideration of polynomial terms such as x^2 is often warranted when there is theoretical or empirical evidence suggesting a nonlinear relationship between the dependent and independent variables. The addition of x^2 allows for the modeling of curvature in the data that a simple linear model would miss.

3.2.1 Enhancing Model Flexibility

By including x^2, we allow the model to fit a wider range of data shapes. This flexibility can lead to a more accurate representation of the underlying data-generating process, which could be parabolic or more complex.

3.2.2 Addressing Model Misspecification

If a model excluding x^2 exhibits systematic patterns in the residuals, such as a funnel shape or a parabolic trend, it is a clear indication of model misspecification. Adding x^2 can help correct this issue, leading to homoscedastic and normally distributed residuals, which are assumptions of the OLS method.

3.3 Goodness-of-Fit and Model Selection

To evaluate the appropriateness of including x^2, goodness-of-fit tests such as the F-test for nested models are essential. This comparison can inform whether the additional complexity introduced by x^2 significantly improves the model's explanatory power. As discussed in chapter 2 of Wooldridge [5], such tests are crucial for assessing the model's overall fit.

3.3.1 F-Test for Nested Models

The F-test assesses whether the restricted model (without x^2) is significantly different from the extended model (with x^2). A significant F-statistic suggests that the extended model provides a better fit to the data.

3.3.2 Assessing Explanatory Power

Inclusion of x^2 should lead to an increase in the coefficient of determination, R^2, indicating a higher proportion of variance explained by the model. This, along with an improved adjusted R^2, which accounts for the number of predictors, reinforces the argument for the extended model.

3.4 Conclusion

The assertion that the inclusion of x^2 in an OLS regression model does not improve the fit is an oversimplification. The zero correlation between the linear term and the residuals does not negate the value of additional polynomial terms. Instead, the introduction of such terms should be guided by theoretical justifications, empirical evidence, and statistical validation through goodness-of-fit tests. When properly implemented, the inclusion of x^2 can uncover more complex relationships, thereby enhancing the model's explanatory power and predictive accuracy. In conclusion, careful deliberation and rigourous empirical evaluation should guide the incorporation of polynomial terms to ensure a more refined understanding of the underlying data generation mechanisms.

References

[1] Hattie, J. (2009). *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement*

[2] Blatchford, P., Bassett, P., & Brown, P. (2011). *Examining the effect of class size on classroom engagement and teacher-pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools*

[3] Krueger, A. B. (2002). *Economic considerations and class size.* The Economic Policy Institute.

[4] Hanushek, E. A. (1998). *The evidence on class size.* In S. Ehrenberg (Ed.), *What's the right class size for America's schools?* National Education Association.

[5] Wooldridge, J. M. (2018). *Introductory Econometrics: A Modern Approach* (7th ed.). Boston: Cengage.

[6] Africa World Press Inc.; First American Edition (2009). *Self-Determination and National Unity.*