STAT 5320 - ASSIGNMENT 1

Pham Thi Thai - T00727094

John Joshua Bardelosa - T00728432

April 5, 2024

1 Introduction

In this exercise, we embark on an analysis of Anscombe's Quartet, a set of datasets designed to share identical summary statistics while revealing distinct patterns when visually inspected. Our analysis focuses on conducting simple linear regressions on each dataset and closely examining the resulting residuals by extracting key parameters through the method of maximum likelihood. Additionally, we will explore the coefficient of determination, a crucial metric that quantifies the proportion of the variability in the dependent variable explained by our model.

2 Data Description

Anscombe's quartet comprises four datasets, each comprising 11 data points with corresponding x and y coordinates, demonstrating identical summary statistics such as means ($\bar{x} = 9.00$, $\bar{y} = 7.50$), variances ($\text{var}(x) = 11.02$, $\text{var}(y) = 4.13$), correlations ($\text{corr} = 0.82$), and linear regression coefficients.

For all sets (1 through 4), the coefficient estimates for the intercept (β_0) are close to 3.0, indicating the expected value of y when x is zero. The coefficient estimates for slope (β_1) are consistently around 0.5. This suggests a linear relationship, as an increase of 1 unit in x is associated with an increase of approximately 0.5 units in y. The R-squared values (R^2) are relatively high for all sets, ranging around 0.67. This suggests that a significant portion (67%) of the variability in the response variable y is explained by the linear regression model. In conclusion, the linear regression models demonstrate statistically significant relationships between x and y in each Anscombe set.

3 Literature Review

3.1 Scatter plot and Linear regression line

Anscombe's quartet highlights the limitations of relying solely on summary statistics, emphasizing the risk of overlooking patterns and outliers. It advocates for a holistic approach by integrating data visualization techniques, such as residuals vs. fitted and Q-Q plots, to enhance the depth of statistical exploration.

In the course of this investigation, a comprehensive visual analysis was undertaken to complement the descriptive statistical properties across the four datasets. Distinctive scatter plots, meticulously accompanied by linear regression lines, were methodically generated for each dataset, as illustrated in Figure 1.

- Dataset 1 showcases characteristics akin to many well-behaved datasets, displaying clean and well-fitting linear models.
- Dataset 2, conversely, lacks a discernible linear correlation, leading to the inference that a non-linear relationship between x and y exists.
- In the case of Dataset 3, a near-perfect linear relationship is observed for the majority

of data points, with the exception of one outlier conspicuously deviating from the linear model.

- Dataset 4, in its entirety, defies conformity to any linear model. However, the singular outlier in this dataset prevents triggering an alarm. This instance exemplifies the scenario where the presence of a single high-leverage point is adequate to generate a seemingly high correlation coefficient.

3.2 Residuals analysis

In the analysis of residuals for each dataset within Anscombe's Quartet, noteworthy observations emerge. The mean of residuals across the four datasets is found to be approximately zero, suggesting that, on average, the statistical model provides predictions that are unbiased. This implies that the model neither consistently overestimates nor underestimates the observed values, indicating a fair representation of the underlying relationships in the data.

Furthermore, the variance of residuals across all four datasets is observed to be approximately 1.38. This measure reflects the spread or dispersion of the residuals around their mean. A variance close to 1.38 indicates a moderate level of variability in the residuals, suggesting that the model captures a reasonable amount of the inherent complexity within each dataset.

3.3 Residuals vs fitted values plot

Transitioning to a more sophisticated tool, we delved into the realm of residuals vs fitted values plots (refer to Figure 2).

For Dataset 1, the residuals demonstrate a random distribution, devoid of any discernible pattern. This randomness underscores the model's ability to closely align predictions with observed values, reinforcing the reliability and accuracy of the developed machine learning model for Dataset 1.

In contrast, Dataset 2 presents a residual plot with a U-shaped curve, suggesting an inadequate fit for the linear model. This pattern implies that the linear model may not accurately capture the underlying relationship between variables, advocating for the consideration of a non-linear model.

Dataset 3's residual plot reveals a conspicuous linear decreasing pattern, indicative of a systematic decrease in residuals and suggesting a linear trend. The presence of an outlier further underscores the inadequacy of the model.

Finally, Dataset 4's residual plot forms a vertical line parallel to the vertical axis, featuring an outlier. This pattern signifies a specific structure in the residuals, posing challenges to the model's accuracy. The resemblance of the residual plot's shape to the x vs. y pattern suggests a struggle in capturing the true nature of the relationship between variables. The presence of a vertical line and an outlier accentuates the difficulties in accurately representing the data within this dataset.

3.4 Quantile-Quantile plot

Next, we embark on a more nuanced exploration of the residuals through the lens of Quantile-Quantile (Q-Q) plots for Anscombe's Quartet datasets. Our meticulous examination of the Q-Q plots, elegantly depicted in Figure 3, follows a comprehensive approach to ascertain the normality assumptions of the residuals.

Q-Q plot for Dataset 1 demonstrates a nearly straight line, suggesting that the residuals are approximately normally distributed. This aligns with the assumption of normality, indicating that the errors from the linear model in Dataset 1 follow a normal distribution.

The Q-Q plot for Dataset 2 deviates from a straight line, showing an upward or downward curve. This departure indicates a deviation from normality in the residuals. The non-linearity suggests that the errors may not follow a normal distribution, and this could prompt

consideration for alternative modeling approaches.

The Q-Q plot for Dataset 3 exhibits a noticeable departure from a straight line, indicating potential non-normality in the residuals. The curvature suggests that the errors may deviate from a normal distribution. Additionally, the presence of an outlier may contribute to this departure.

The Q-Q plot for Dataset 4 showcases a pronounced departure from linearity, forming a distinct curve. This departure suggests a non-normal distribution of residuals. The curvature, coupled with potential outliers, implies challenges in assuming normality for the errors in Dataset 4.

4 Appendix

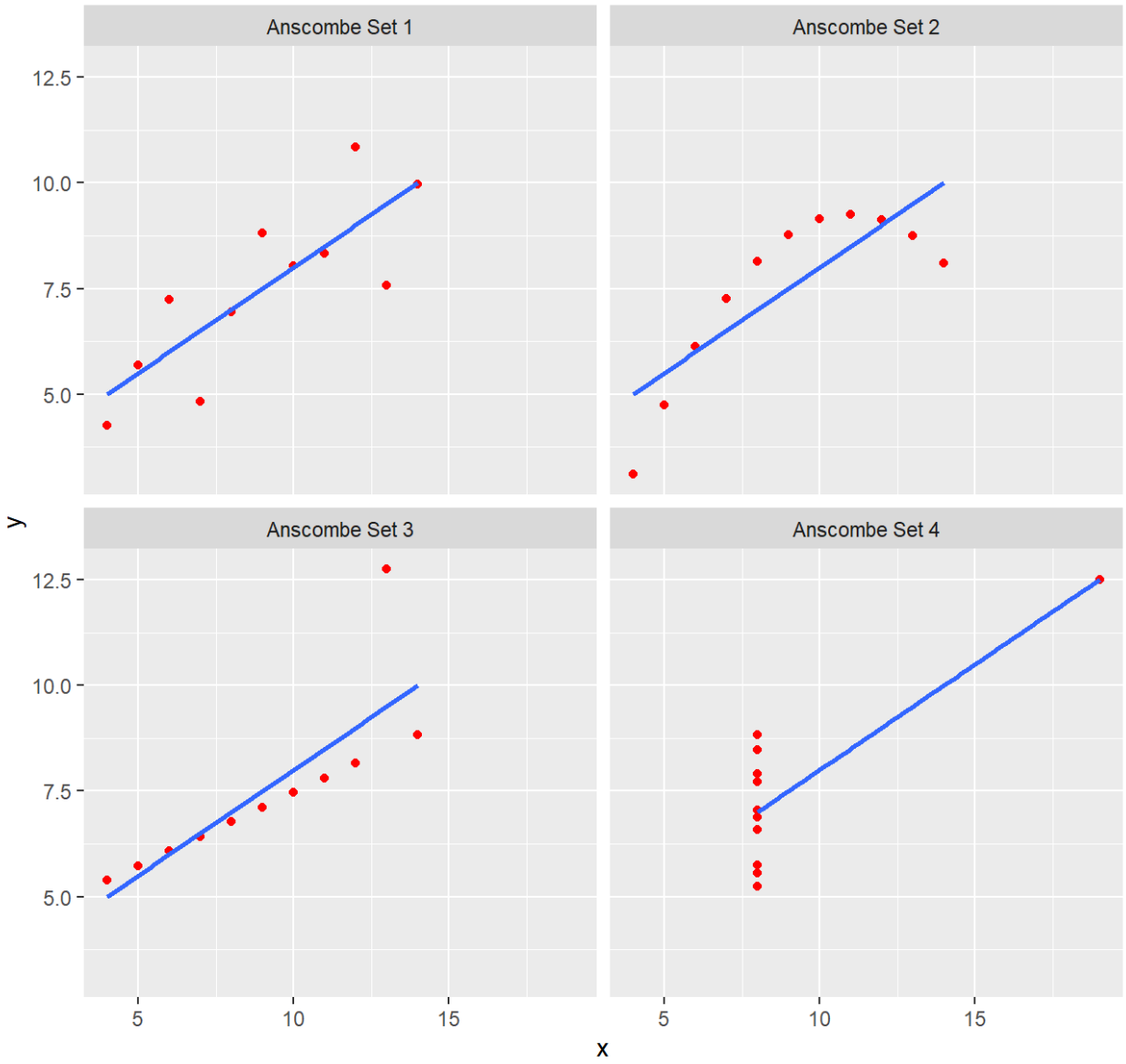
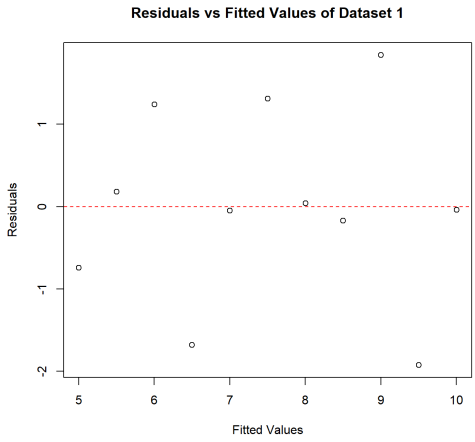
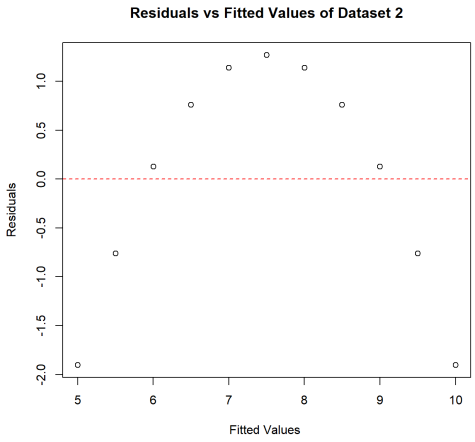


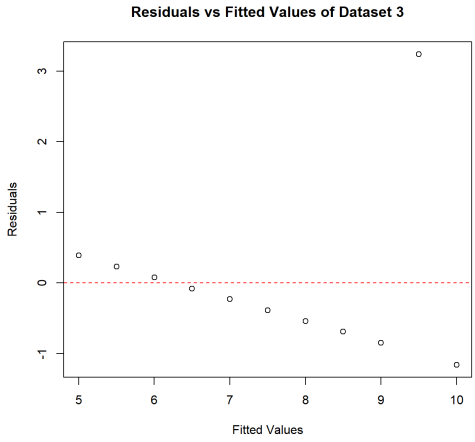
Figure 1: Scatter plot and Regression line for each dataset



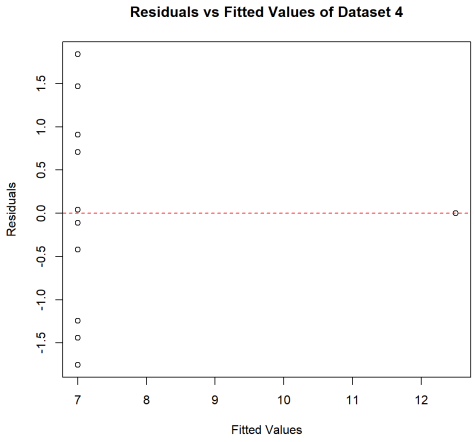
(a) Dataset 1



(b) Dataset 2



(c) Dataset 3



(d) Dataset 4

Figure 2: Residuals vs Fitted values plot

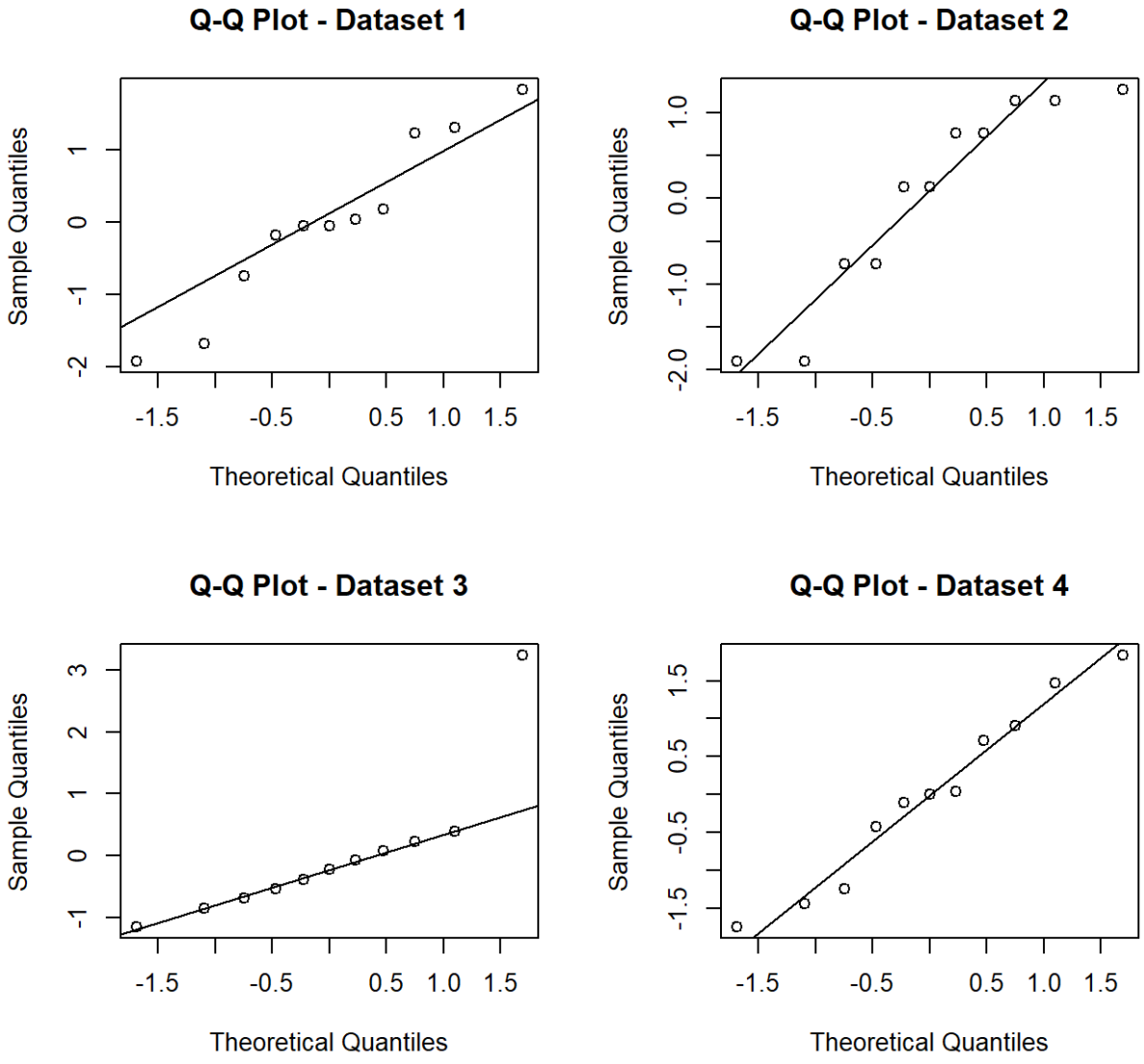


Figure 3: Q-Q plot for each dataset

5 References

- Anscombe's quartet, GeeksforGeeks.org,

  https://www.geeksforgeeks.org/anscombes-quartet/

- Interpreting the residuals vs. fitted values plot for verifying the assumptions of a linear model, stats.stackexchange.com

  https://stats.stackexchange.com/questions/76226/interpreting-the-residuals-vs-fitte

- Anscombe's Quartet of 'Identical' Simple Linear Regressions, rstudio-pubs-static.s3.amazonaws.com

  https://rstudio-pubs-static.s3.amazonaws.com/52381_36ec82827e4b476fb968d9143aec7c4f
  html

- Utkarsh, 2023, Residual Analysis

  https://www.scaler.com/topics/data-science/residual-analysis/

- Faraway, J. J. (2016). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models (2nd ed.). (pp. 6-19).