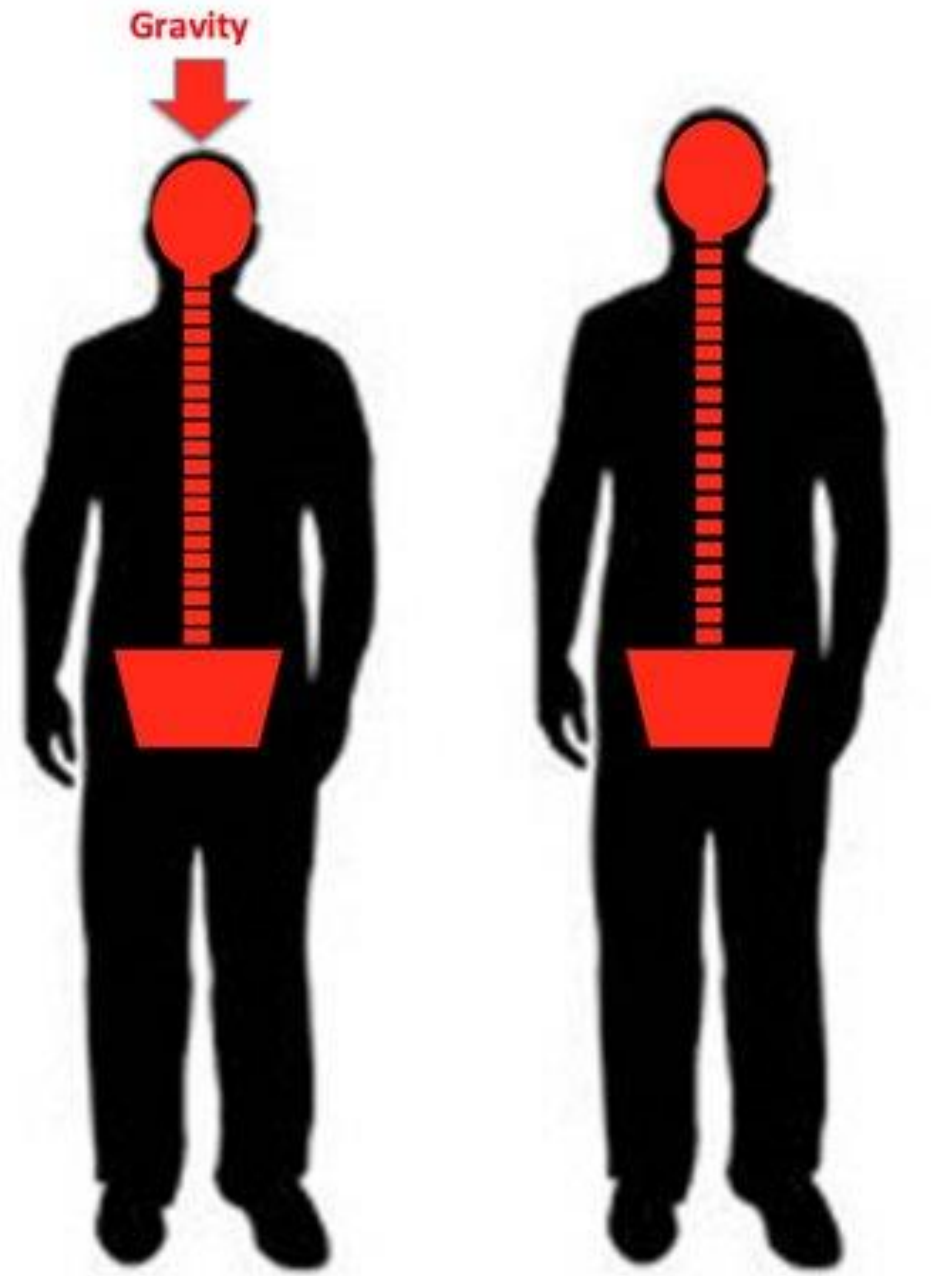
Understanding Height Dynamics Through Regression

JACKIE HARMON



Introduction

Background:

- It is a well-known phenomenon that our height decreases throughout the day due to gravitational compression. Height measured in the evening is often less than in the morning.
- This project investigates this effect using AM and PM height measurements from students at a boarding school in India.

Objective:

- To explore the relationship between morning (AM) and evening (PM) heights using regression analysis to quantify the impact of gravity on daily height variation.

Dataset Overview

Source:
- Data collected from students at a boarding school in India.

Variables:
- AM height measurements (in mm).
- PM height measurements (in mm).

Sample Size:
- Number of observations: 150 students (as per the data available)

Methodology

Tools Used:

In R

- **ggplot2** for visualization
- **lm()** for linear regression
- **gvlma** for global validation of linear model assumptions
- **predictmeans** for calculating predicted means and diagnosing residuals
- **car** for diagnostic tests (e.g., detecting outliers, testing homoscedasticity)
- **caret** for model training, tuning, and cross-validation

In Python

- **Pandas** for data manipulation
- **NumPy** for numerical operations
- **SciPy** for statistical testing
- **Matplotlib** and **Seaborn** for visual analysis
- **statsmodels** for regression modeling and diagnostics
- **pylab** for plotting and visualization utilities

Methodology

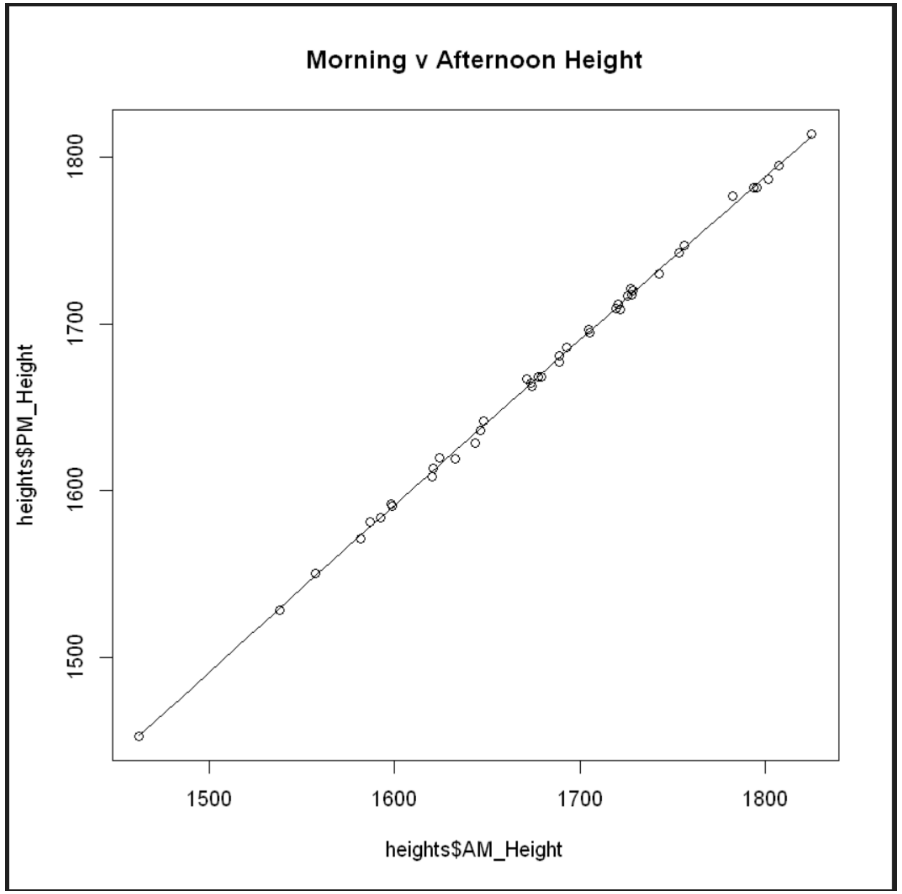
Analysis Steps

- **Data Cleaning and Preparation**: Ensured dataset quality and readiness for analysis.
- **Linear Regression**: Modeled the relationship between AM and PM heights.
- **Assumptions Testing**:
    - Linearity
    - Normality
    - Homoscedasticity
    - Outlier Detection
    - Leverage and Influence

Assumptions Testing - Linearity

In R

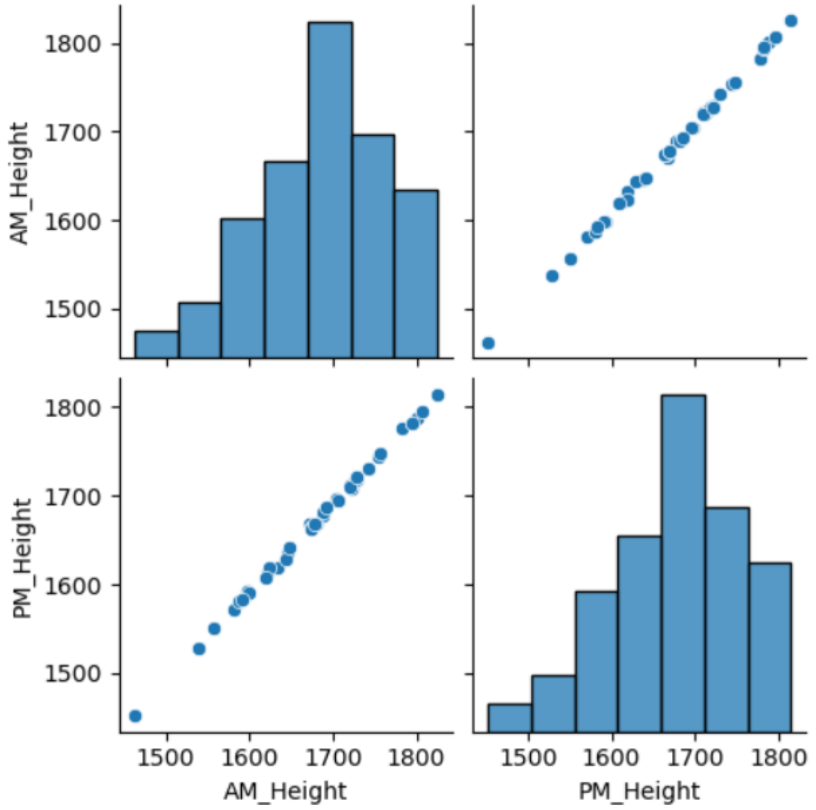
- A scatter plot was used to visually assess the linearity between AM and PM heights.
- Visual inspection indicated a linear relationship, which justified using a linear regression model.



Assumptions Testing - Linearity

In Python

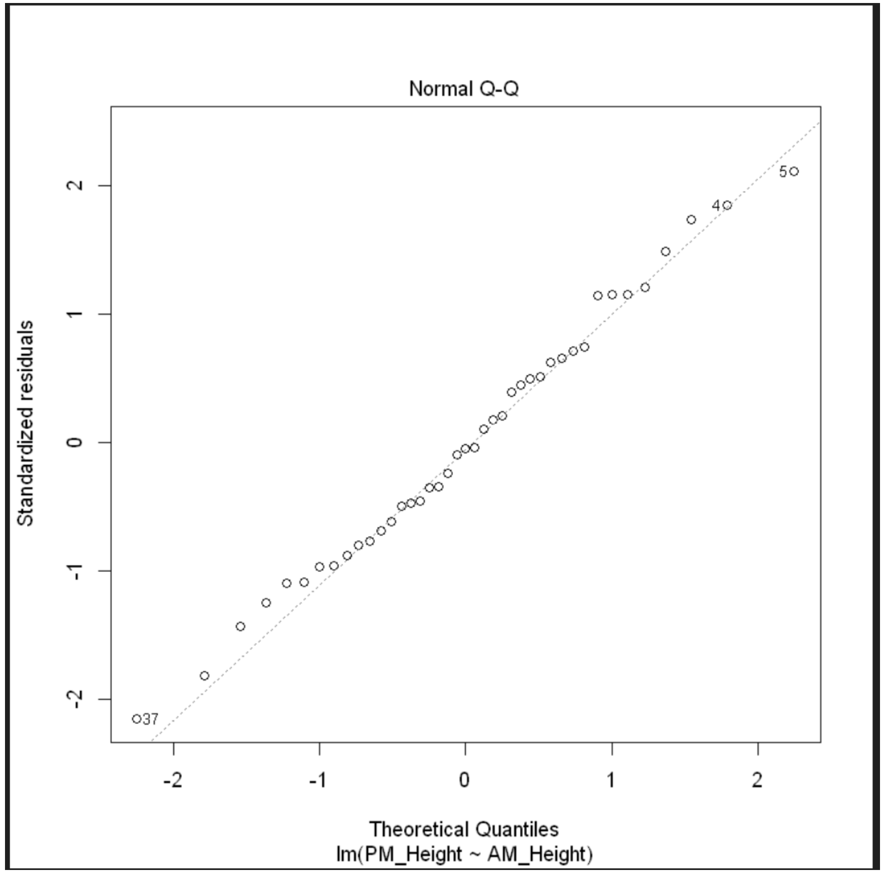
- A **scatter plot** was also utilized to check for linearity.
- The linear relationship was confirmed by observing the plot, validating the use of linear regression.
- A **Harvey Collier test** was conducted with a p-value larger than 0.05, suggesting the linear model is appropriate



Assumptions Testing - Normality

In R

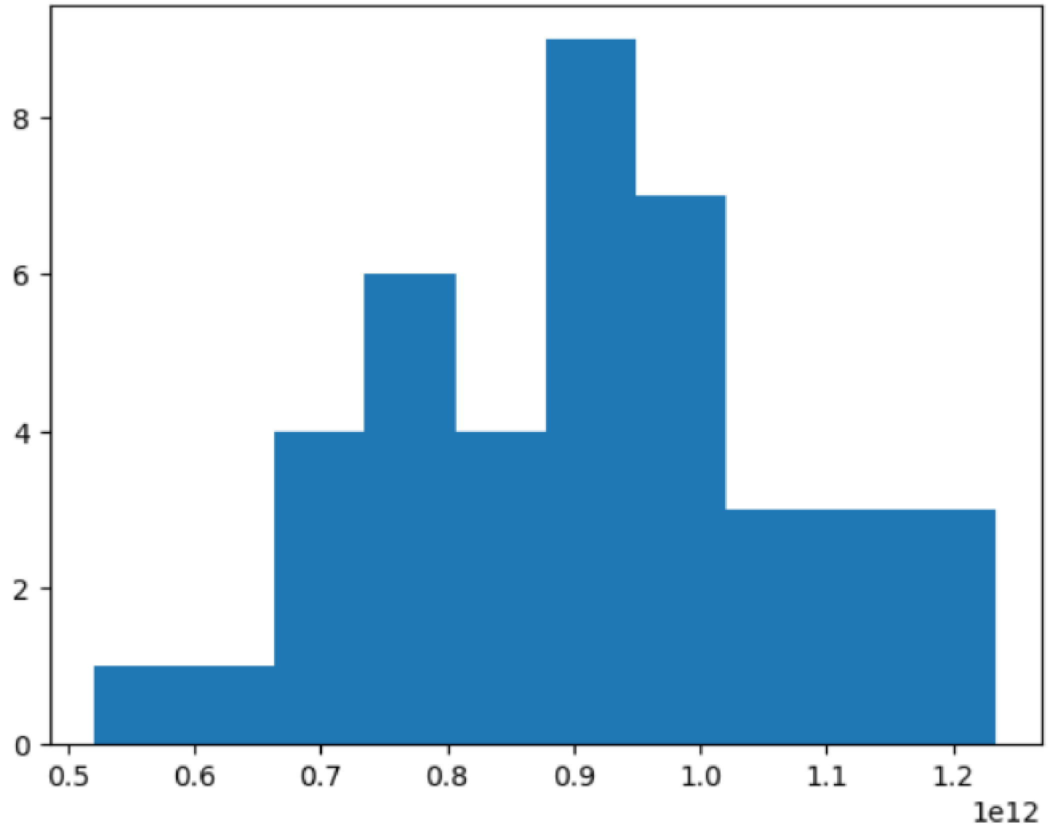
- **QQ plot** was employed to visually evaluate the normality of residuals.



Assumptions Testing - Normality

In Python

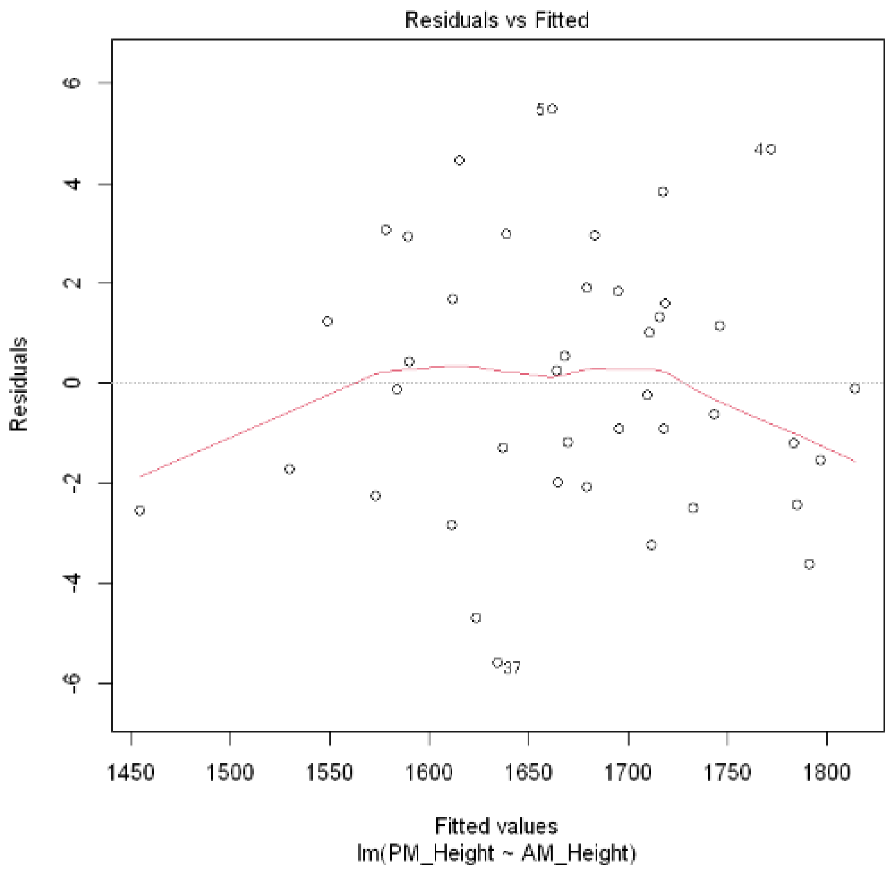
- **Box-Cox Transformation**: The histogram of the transformed data shows a more normal distribution compared to the original data, indicating successful variance stabilization.



Assumptions Testing - Homoscedasticity

In R

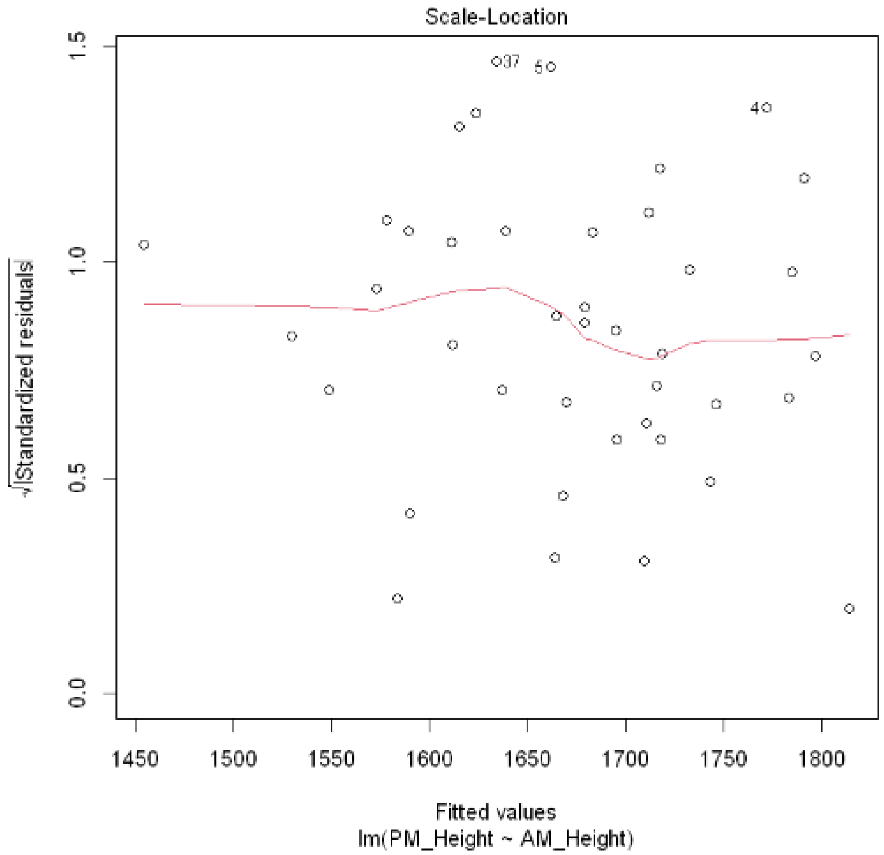
- **Residual vs. Fitted Plot**: The residuals exhibited constant variance across fitted values, indicating homoscedasticity.
- **Breusch-Pagan Test**: Confirmed homoscedasticity with no significant heteroscedasticity detected.
- **Non-Constant Variance Test**: Confirmed homoscedasticity with no significant heteroscedasticity detected.



Assumptions Testing - Homoscedasticity

In R

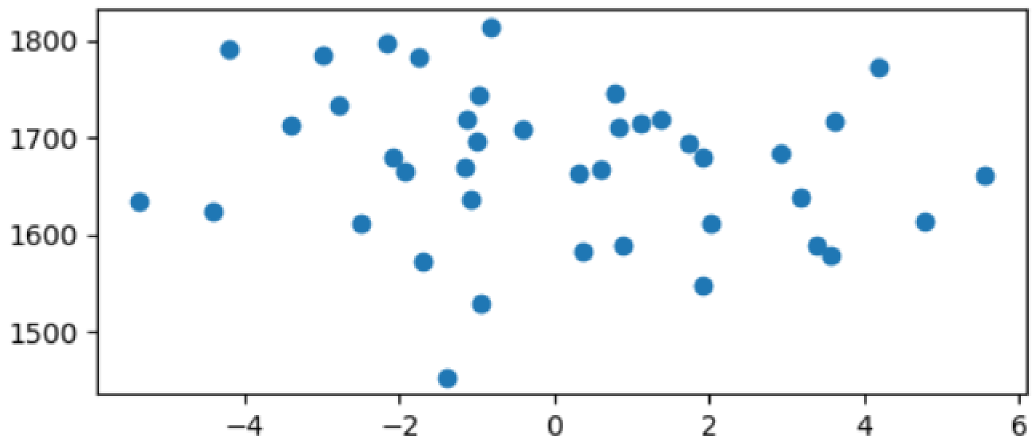
- **Scale-Location Plot**: The residuals exhibited no clear pattern, with red line roughly horizontal, indicating homoscedasticity.



Assumptions Testing - Homoscedasticity

In Python

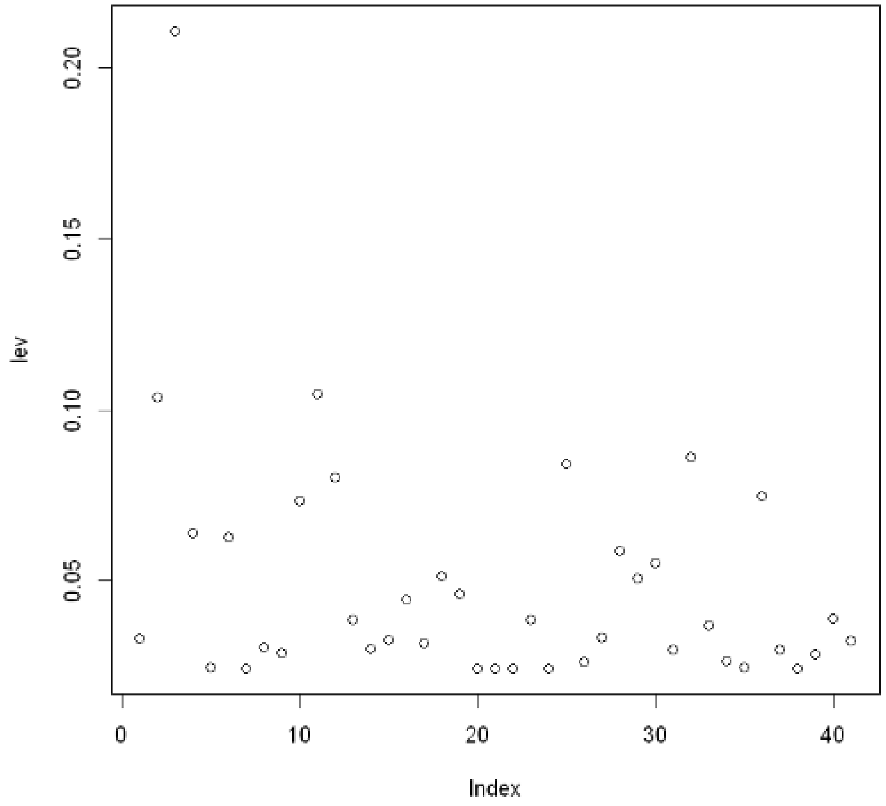
- **Homoscedasticity**:
  - Residual plot: residuals showed no clear pattern, confirming homoscedasticity.
  - Breusch-Pagan test with significant p-value



Outlier Detection

In R

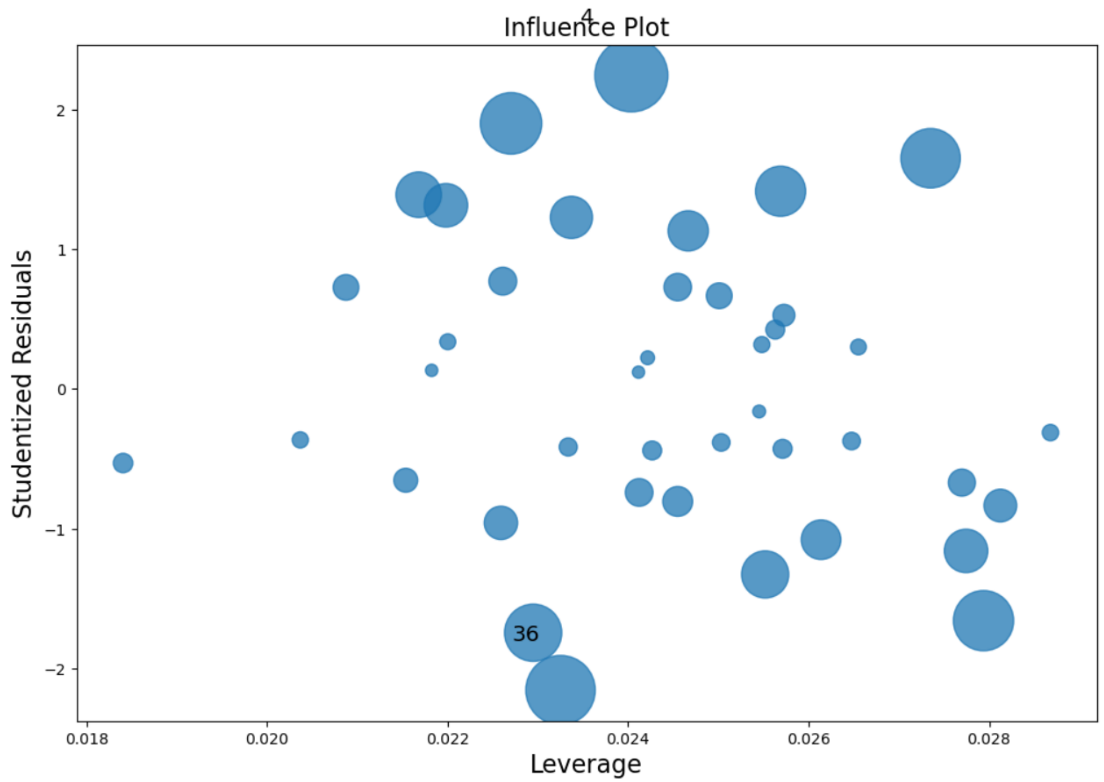
- **GVLMA Test**: Overall model diagnostics confirmed assumptions and model integrity.
- **Screening for Outliers in X Space**:
  - **Cook's Distance**: Identified and addressed influential outliers, ensuring they did not unduly affect the model.
  - **Leverage Values**: Detected high-leverage points, with appropriate adjustments made.
- **Screening for Outliers in Y Space**: Identified outliers in the response variable.



Outlier Detection

In Python

- **Influence Plot:** Observations with high leverage and large studentized residuals, particularly those with a large Cook's distance, flagged as influential



Regression Analysis in R

Model Summary:

- R-squared value: 0.9989, indicating 99.89% of the variance in PM heights is explained by AM heights.
- The residuals appear to be symmetrically distributed around zero, indicating a good fit.
- The residual standard error of 2.627 is relatively small, suggesting that the model predictions are close to the actual values.
- Significant slope coefficient (p-value < 0.05) confirms the effect of gravitational compression.

Assumptions Validation:

- Tests confirmed assumptions were met or adjusted for.
- Transformation applied to improve model fit.

Regression Analysis in Python

Model Summary:

- Similar R-squared value and statistical significance observed as in R analysis.
- Model confirmed findings from R analysis, reinforcing the conclusions.

Assumptions Validation:

- Consistent validation of assumptions across both platforms.

Key Findings

Daily Height Reduction:

- Height measurements confirm a decrease from morning to evening, attributed to gravitational effects.

Statistical Significance:

- Both R and Python analyses indicate a significant correlation between AM and PM heights.

Implications:

- Highlights the importance of considering time of day in height-related measurements for research and health assessments.

Conclusions

Summary:

- The project successfully demonstrated the measurable impact of gravity on daily height variations.
- Showcased proficiency in using R and Python for data analysis, modeling, and assumption testing.

Future Work:

- Investigating other factors influencing height variation, such as age or posture, and expanding the study to different populations.