**F24-OIT-249-01/02 - MSx: Data and Decisions  
Fall 2024**

**Regression Project Assignment**  
Agata Braja, Alexis Brown, Joselyn Moreira, Anan Ogawa

**Internet Usage and Its Determinants**

**Introduction**

This report provides an analysis of the factors influencing internet usage, with insights drawn from socioeconomic and infrastructural predictors. The modeling process involved systematic iterations to improve diagnostic outcomes and refine interpretations, culminating in actionable recommendations for enhancing digital access.

**1. Data Processing and Feature Engineering**

**Key Steps:**

* **Data Cleaning**:
  + Missing values were addressed, and variables were standardized.
  + Variable names were updated for clarity.
* **Feature Engineering**:
  + Log transformations applied to stabilize variances in Internet\_Usage, Access\_Electricity, and Poverty\_Rate.
  + Interaction terms were initially created but excluded during model refinement due to collinearity issues.
  + Poverty rates were categorized into quartiles (Poverty\_Quartile) to enable stratified analysis.

A graph of numbers and text

Description automatically generated with medium confidenceA graph showing a number of squares

Description automatically generatedA graph with purple dots and a red line

Description automatically generatedA graph with a red line and green dots

Description automatically generatedA graph with a red line and blue dots

Description automatically generatedA graph showing a red line and green dots

Description automatically generatedA graph with a red line and a red line

Description automatically generatedA graph of a graph

Description automatically generated

**Files:**

* data\_cleaning.R: Data preprocessing script.
* feature\_engineering.R: Feature engineering and transformation script.
* outputs/refined\_data\_with\_log\_transformations.csv: Cleaned and transformed dataset.

**2. Regression Modeling**

**Iterative Model Development:**

* **Initial Model**:
  + Assessed relationships between predictors (e.g., Access\_Electricity, Poverty\_Rate) and Internet\_Usage.
  + Residual diagnostics highlighted issues with heteroscedasticity and influential points.
* **Refinements**:
  + Log transformations addressed heteroscedasticity.
  + Centered variables (e.g., Mobile\_Cellular\_Subscriptions) reduced multicollinearity.
  + Influential points were removed to enhance robustness.

**Files:**

* regression\_modelling.R: Initial model-building script.
* model\_refinements.R: Refinement and final modeling script.
* outputs/final\_model\_summary.txt: Final model summary.

**3. Diagnostics and Refinements**

**Residual Analysis:**

* **Diagnostics**:
  + Residuals vs. Fitted: Verified linearity.

A graph of a graph showing a number of dots

Description automatically generated with medium confidence

This plot demonstrates the relationship between the residuals (differences between observed and predicted values) and the fitted values from the regression model.

**Observations:**

1. **Non-Random Pattern**:
   * The residuals display a slight curvature, and the variance appears to increase at higher fitted values. This suggests **heteroscedasticity**—a violation of the assumption of constant variance in residuals.
2. **Outliers or Influential Points**:
   * Specific points, such as observations labeled as 568, 1469, and 1969, deviate significantly from the general distribution. These may represent **outliers** or points of high leverage, which could unduly influence the model.

**Implications:**

* **Heteroscedasticity** needs to be addressed to improve the reliability of the model's standard errors and inference.
* **Potential Influential Points** (e.g., Cook's Distance) should be further examined to assess their impact on the model. These points may need to be excluded or adjusted in subsequent iterations.
  + Q-Q Plot: Checked normality.

A graph of a graph

Description automatically generated

This Q-Q plot illustrates the standardized residuals of the regression model against the theoretical quantiles of a standard normal distribution.

**Observations:**

1. **Normality of Residuals**:
   * The residuals mostly follow the 45-degree reference line, suggesting approximate normality.
   * However, deviations are noticeable at both tails, particularly at the **upper end**, indicating the presence of **heavy tails** or extreme values.
2. **Outliers**:
   * Observations labeled as 568, 1469, and 1969 are significant outliers, deviating strongly from the expected distribution.

**Implications:**

* While the normality assumption is reasonably satisfied in the central range, the **deviations at the tails** suggest that certain outliers may need further investigation or removal.
* Addressing the outliers could enhance the model's robustness, particularly for statistical inference.
  + Scale-Location: Evaluated variance homogeneity.

A diagram of a graph

Description automatically generated

This plot visualizes the standardized residuals' spread (√|Standardized Residuals|) against the fitted values of the model.

**Observations:**

1. **Homoscedasticity**:
   * The red trend line slightly decreases, indicating a mild non-constant variance in residuals (heteroscedasticity). The residuals appear to be more spread for higher fitted values.
   * The model may benefit from techniques to stabilize variance, such as transformations.
2. **Outliers**:
   * Points labeled 568, 1469, and 1969 show high residual values and are clear outliers. These observations should be investigated further to determine their influence on the model.

**Implications:**

* The presence of heteroscedasticity could affect the reliability of model coefficient estimates and standard errors.
* Applying variance-stabilizing transformations or robust regression may improve model performance.
* The outliers need to be reviewed for potential removal or adjustments to ensure they do not disproportionately affect the model's accuracy.
  + Leverage Plot: Identified influential observatioA graph with a red line

    Description automatically generated

This plot evaluates the influence of data points by visualizing the standardized residuals against their leverage values. It also overlays Cook's distance thresholds, helping identify influential observations.

**Observations:**

1. **Influential Points**:
   * Points 1042, 1043, and 1409 have high leverage and are potential influential points. These points may disproportionately impact the model's coefficients and predictions.
   * These points are above Cook's distance threshold, suggesting they have a strong influence on the regression fit.
2. **Leverage**:
   * Most data points cluster around low leverage values, which is expected. However, the outliers with high leverage warrant investigation to understand their underlying characteristics.
3. **Residual Distribution**:
   * Residuals appear relatively well-distributed around zero for lower leverage values. However, the red trend line shows some deviation at higher leverage points, indicating potential nonlinearity or other modeling issues.

**Implications:**

* Investigating the influential points (e.g., 1042, 1043, 1409) is critical to determine if they are errors, outliers, or represent unique patterns in the data.
* Sensitivity analysis or robust regression methods may mitigate the influence of such points.
* Adjustments to the model or further data transformations may be required to ensure that these observations do not distort overall model performance.
* **Adjustments**:
  + Removed high-leverage points and re-fitted the model.
  + Verified improvements in residuals using updated diagnostic plots.

**Multicollinearity:**

* Variance inflation factors (VIFs) were computed to ensure all predictors were within acceptable ranges.

**Files:**

* diagnostics\_refinements.R: Diagnostic and refinement processes.
* outputs/final\_model\_diagnostics.png: Updated diagnostic plots.
* outputs/refined\_vif\_values.txt: VIF analysis results.

A group of graphs showing different types of data

Description automatically generated with medium confidence

Final model summary:

A screenshot of a computer

Description automatically generated

A graph with a bar

Description automatically generated

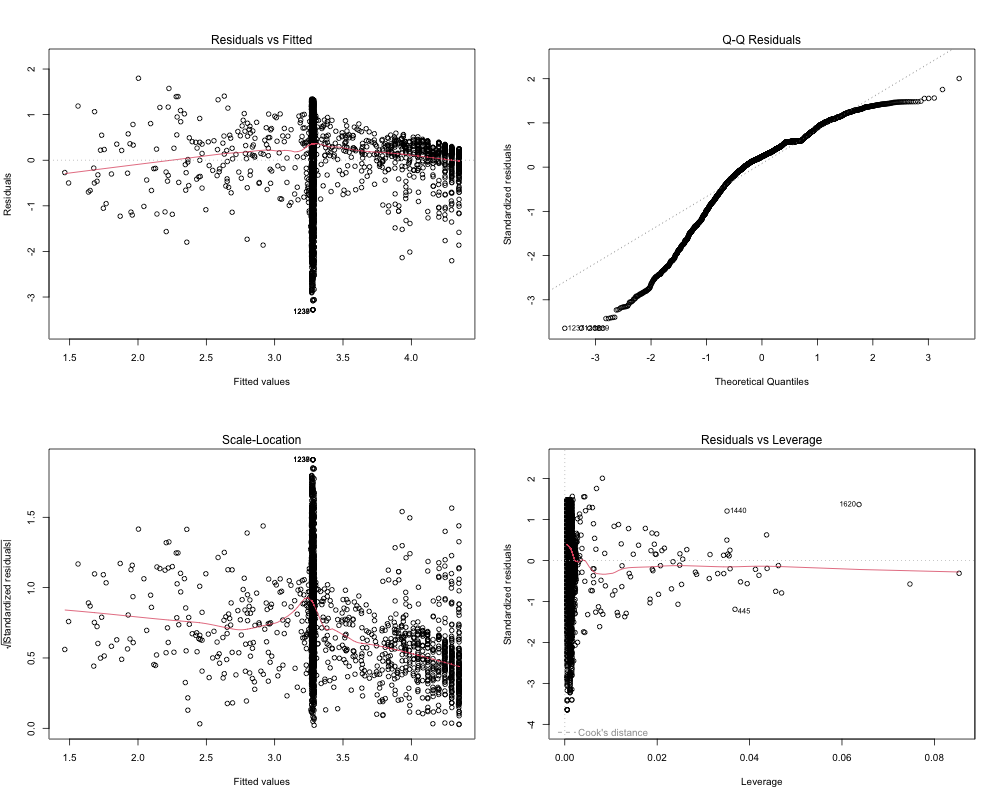
**4. Model Interpretation**

**Key Findings:**

1. **Log\_Poverty\_Rate**:
   * A significant negative association indicates higher poverty rates lead to reduced internet usage.
   * Economic barriers remain a key digital inclusion challenge.
2. **Mobile\_Cellular\_Subscriptions\_Centered**:
   * Insignificant predictor after centering, suggesting marginal contribution when controlling for poverty and infrastructure.
3. **Poverty\_Gap**:
   * Negative but not statistically significant in the final model, warranting further exploration.
4. **Poverty\_Quartile**:
   * Stratified analysis shows disparities in internet usage based on economic strata.

**Files:**

* model\_interpretation.R: Code for interpreting coefficients.
* outputs/refined\_residual\_diagnostics.png: Supporting plots.



**5. Policy Recommendations**

1. **Expand Infrastructure**:
   * Increase investments in electricity and broadband to underserved regions.
2. **Reduce Economic Barriers**:
   * Subsidize internet access for low-income communities.
   * Implement digital literacy initiatives.
3. **Target High-Poverty Regions**:
   * Prioritize interventions in higher Poverty\_Quartiles to reduce the digital divide.

**6. Reflections and Future Work**

**Model Iterations:**

* Initial steps involved variable selection and basic regression diagnostics.
* Subsequent iterations refined predictors, removed influential data points, and validated diagnostics.

A graph with blue lines

Description automatically generated

A diagram of a different model

Description automatically generated with medium confidence

**Recommendations for Future Research:**

* Incorporate additional predictors (e.g., education, employment).
* Investigate regional or longitudinal variations to understand policy impacts.

**Files:**

* Scripts: data\_cleaning.R, feature\_engineering.R, diagnostics\_refinements.R.
* Outputs: final\_model\_summary.txt, final\_model\_diagnostics.png.

**Conclusion**

This analysis underscores the importance of addressing infrastructure gaps and economic barriers to foster digital inclusion. Policymakers should leverage these insights to prioritize targeted investments in connectivity and digital equity programs. The iterative modeling approach ensures reliable findings that can guide actionable interventions.