MODELING GPA DETERMINANTS USING MULTIPLE LINEAR REGRESSION IN R FOR A U.S. COMMUNITY COLLEGE

1. Background

A community college in Illinois observed persistent disparities in academic performance across its student body. Despite equal access to faculty and course materials, GPA outcomes varied by demographics, course formats, and personal circumstances.

To inform policy decisions on support services and academic planning, the Office of Institutional Research commissioned a statistical study. We used R to model how study time, part-time work hours, digital access, and classroom environment influence GPA, using multiple linear regression.

2. Objective

- To identify the key quantitative predictors of GPA among community college students
- To test the effects of socioeconomic, behavioral, and institutional factors using regression modeling in R
- To provide actionable recommendations for program design, advising, and equity-based intervention

3. Data Used

Source: Student survey merged with academic records for Fall 2022

Dataset Details:

- 1,326 anonymized undergraduate students
- Variables used:
    - Dependent: GPA (scale: 0.0–4.0)
    - Predictors:
        - Study_Hours_Per_Week
        - Job_Hours_Per_Week
        - Class_Size

- Internet_Access_Quality (Likert: 1–5)
- Course_Format (Dummy: 1=Online, 0=Offline)
- First_Generation_Student (Dummy)
- Age
- Gender (Male/Female/Other as dummies)

4. Methodology

4.1 Data Cleaning

- Processed with dplyr, imputed missing values using median (for continuous) and mode (for categorical)
- Scaled quantitative variables using scale()
- Created dummy variables for Gender, Course_Format, and First_Generation_Student

4.2 Regression Modeling

- Built multiple linear regression model using lm()
- Performed diagnostic checks using:
  - **Variance Inflation Factor (VIF)** with car
  - **Residual plots**
  - **Normality of residuals** using Q-Q plot
  - **Homoscedasticity** via Breusch-Pagan test from lmtest

5. Model Output and Diagnostics

Predictor Variable	Coefficient (β)	p-value	Significance
Study_Hours_Per_Week	+0.032	< 0.001	Yes
Job_Hours_Per_Week	–0.021	0.015	Yes
Class_Size	–0.008	0.091	No
Internet_Access_Quality	+0.125	< 0.001	Yes
Course_Format (Online=1)	–0.143	0.002	Yes

First_Generation_Student	–0.087	0.019	Yes
Age	+0.006	0.110	No
Gender (Female vs. Male)	+0.102	0.008	Yes

Adjusted R^2 = 0.36 **VIF** scores all < 2.0 **Residuals**: Normal, no major outliers, homoscedasticity confirmed

6. Interpretation and Recommendations

- **Study time** strongly predicts GPA; recommended academic support initiatives emphasize time management
- **Job hours** negatively affect GPA; suggested policy for reducing course load for working students
- **Internet quality** and **online format** show digital inequality effects; recommended enhanced broadband support for online learners
- **First-generation students** showed GPA disadvantages; advising and mentoring should be prioritized for this group
- **Gender gap** favored females; further qualitative study recommended to explore academic environment perceptions

7. Reporting Output

- **R Markdown Report (PDF, 22 pages)**:
    - Executive summary
    - lm() model summary
    - Visuals: scatterplots, coefficient plot, residual diagnostics
    - Recommendations linked to findings
- **Excel Sheet**:
    - Clean dataset with dummy-coded predictors
    - Predicted GPA values per student
    - Coefficients with CI and p-values
- **Slide Deck**:

- 10-slide presentation for institutional leadership
- Focus on data-to-policy narrative

8. Institutional Impact

- Findings incorporated into the **Student Success Initiative** (2023–24)
- Led to the creation of a **Study Hours Tracker** app pilot for time management coaching
- Broadband support partnership signed with local ISP
- Academic advising model revised to **flag high-risk profiles based on regression outputs**