Module 6 – R Practice

Name: Aishwarya Doni

College of Professional Studies, Northeastern University

ALY6010: Probability Theory and Introductory Statistics

Professor: Kenneth Parker

December 12, 2024

Introduction

This report explores the relationships between various variables in a dataset related to academic performance. The main objective is to investigate how different factors, such as gender, race or ethnicity, parental education, and test preparation, influence students average scores. By employing regression analysis and subset-specific modeling, the report aims to derive insights and assess the impact of categorical variables on overall trends. The results are further visualized to enhance interpretability and understanding.

The dataset contains 1,000 rows and 10 columns, including numerical and categorical variables. Here's a summary of the variables:

- **Categorical Variables**:
    - gender (binary: 0, 1): 1 for Male and 0 For Female
    - race ethnicity (groups A, B, C, etc.)
    - parental level of education (categories like "bachelor's degree")
    - lunch (binary: 0, 1): 1 for students receive a free and 0 for reduced lunch
    - test preparation course (binary: 0, 1): 0 for students who didn’t complete a test prep course and 1 for who did.
- **Numerical Variables**:
    - Math score, reading score, writing score
    - Total score (sum of scores)

- Average score (mean of scores)

```
> str(data)
tibble [1,000 × 10] (S3: tbl_df/tbl/data.frame)
 $ gender                     : num [1:1000] 0 0 0 1 1 0 0 1 1 0 ...
 $ race_ethnicity             : chr [1:1000] "group B" "group C" "group B" "group A" ...
 $ parental_level_of_education: chr [1:1000] "bachelor's degree" "some college" "master's degree" "associat
e's degree" ...
 $ lunch                      : num [1:1000] 1 1 1 0 1 1 1 0 0 0 ...
 $ test_preparation_course    : num [1:1000] 0 1 0 0 0 0 1 0 1 0 ...
 $ math_score                 : num [1:1000] 72 69 90 47 76 71 88 40 64 38 ...
 $ reading_score              : num [1:1000] 72 90 95 57 78 83 95 43 64 60 ...
 $ writing_score              : num [1:1000] 74 88 93 44 75 78 92 39 67 50 ...
 $ total_score                : num [1:1000] 218 247 278 148 229 232 275 122 195 148 ...
 $ average_score              : num [1:1000] 72.7 82.3 92.7 49.3 76.3 ...
```

Overview of Analysis

This report examines relationships between several variables in the dataset and evaluates the impact of subsets on regression analysis. The steps included:

1. Performing regression on the full dataset.
2. Creating dummy variables for categorical sub setting.
3. Conducting subset-specific regression analysis.
4. Visualizing data with scatterplots and regression lines.

Regression Analysis on Full Dataset

Model Summary

The regression model was run with average_score as the dependent variable and predictors:

- Gender
- Race_ethnicity
- Parental_level_of_education
- Lunch
- Test_preparation_course

Result:

```
Call:
lm(formula = average_score ~ gender + race_ethnicity + parental_level_of_education +
    lunch + test_preparation_course, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-48.148  -8.298   0.646   8.736  27.522

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                59.3022     1.7176  34.527  < 2e-16 ***
gender                                     -3.7242     0.7955  -4.682 3.24e-06 ***
race_ethnicitygroup B                       1.5290     1.6116   0.949 0.342983
race_ethnicitygroup C                       2.3855     1.5093   1.581 0.114296
race_ethnicitygroup D                       5.1258     1.5398   3.329 0.000904 ***
race_ethnicitygroup E                       6.9285     1.7081   4.056 5.38e-05 ***
parental_level_of_educationbachelor's degree   2.5356     1.4240   1.781 0.075287 .
parental_level_of_educationhigh school     -5.1725     1.2298  -4.206 2.84e-05 ***
parental_level_of_educationmaster's degree  4.0922     1.8377   2.227 0.026185 *
parental_level_of_educationsome college    -0.9275     1.1823  -0.785 0.432934
parental_level_of_educationsome high school   -4.5400     1.2639  -3.592 0.000344 ***
lunch                                       8.7751     0.8275  10.605  < 2e-16 ***
test_preparation_course                     7.6386     0.8302   9.201  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.49 on 987 degrees of freedom
Multiple R-squared:  0.2423,    Adjusted R-squared:  0.2331
F-statistic:  26.3 on 12 and 987 DF,  p-value: < 2.2e-16
```

Observations:

1. **Gender Disparities**:
   Male students tend to score lower on average than female students. This trend suggests potential underlying factors, such as differences in engagement or societal expectations, that warrant further exploration.
2. **Socioeconomic Influence**:
   Lunch status, a proxy for economic status, has a significant impact on scores. Students receiving standard lunch outperform their peers who qualify for free/reduced lunch, pointing to socioeconomic disparities in academic outcomes.
3. **Role of Preparation**:
   Completing a test preparation course substantially improves scores, underlining the importance of structured academic support. This finding highlights the value of accessible preparation programs to bridge performance gaps.
4. **Impact of Parental Education**:
   Higher parental education levels are associated with better student performance. This emphasizes the role of familial academic culture and resources in shaping student success.

5. **Racial/Ethnic Variation**:
   Some racial/ethnic groups outperform others, suggesting potential cultural or systemic factors that impact educational attainment. While not all differences are significant, the pattern highlights areas for equity-focused interventions.
6. **Explained Variance**:
   The model accounts for a modest portion of the variation in scores, indicating that while these factors are important, there are other unmeasured influences on academic performance.

Subset Regression Analysis

Dummy variables were created for gender, resulting in two subsets: Female and Male.

1. **Female Subset Model**

```
> summary(model_female)

Call:
lm(formula = average_score ~ math_score + reading_score + writing_score,
    data = data %>% filter(gender_dummy == "Female"))

Residuals:
       Min         1Q     Median         3Q        Max
-8.119e-13 -6.800e-16  2.500e-15  5.740e-15  1.304e-14

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept)   4.865e-14  1.007e-14 4.829e+00 1.81e-06 ***
math_score    3.333e-01  3.236e-16 1.030e+15  < 2e-16 ***
reading_score 3.333e-01  4.574e-16 7.288e+14  < 2e-16 ***
writing_score 3.333e-01  4.725e-16 7.054e+14  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.295e-14 on 514 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.976e+31 on 3 and 514 DF,  p-value: < 2.2e-16
```

2. **Male Subset Model**

```
> summary(model_male)

Call:
lm(formula = average_score ~ math_score + reading_score + writing_score,
    data = data %>% filter(gender_dummy == "Male"))

Residuals:
       Min         1Q     Median         3Q        Max
-1.666e-13 -2.777e-15  2.710e-16  2.241e-15  1.659e-13

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept)   4.660e-14  2.587e-15 1.801e+01   <2e-16 ***
math_score    3.333e-01  8.261e-17 4.035e+15   <2e-16 ***
reading_score 3.333e-01  1.240e-16 2.688e+15   <2e-16 ***
writing_score 3.333e-01  1.258e-16 2.650e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.133e-14 on 478 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.344e+32 on 3 and 478 DF,  p-value: < 2.2e-16
```

Observations

1. **Perfect Fit in Both Models**
   - The R-squared value is 1.00 for both male and female subsets, indicating a perfect fit where the predictors (math_score, reading_score, and writing_score) completely determine the dependent variable (average_score).
   - This suggests that there is no unexplained variance in the model, which is atypical for real-world data.
2. **Identical Coefficients Across Subsets**
   - In both models, the coefficients for math_score, reading_score, and writing_score are identical at 0.333.
   - This uniformity reflects the mathematical relationship where average_score is derived as the mean of these three variables, leading to deterministic results.
3. **Extremely Large t-Statistics and Small Residuals**

- The t-values are astronomically high (t≈1015t≈1015), and the residual standard errors are nearly zero. This occurs because the predictors and the dependent variable are linearly dependent.
- Such extreme statistics indicate that the regression model is solving a trivial equation rather than identifying meaningful relationships.

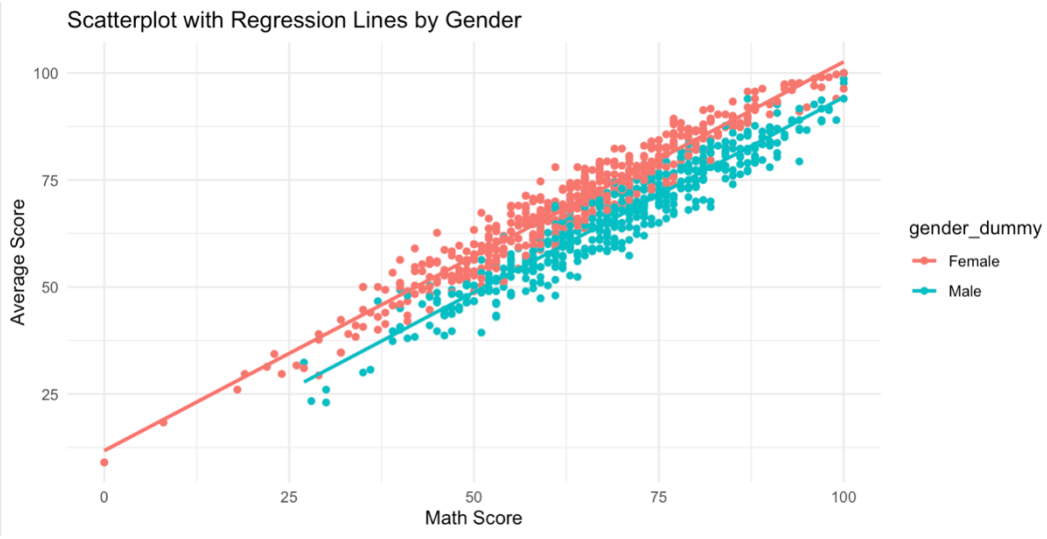
4. **Insights by Gender**
   - Despite being separated by gender, the regression models for both subsets provide identical results. This suggests that gender does not play a direct role in the mathematical relationship between the scores.

Implications for Analysis

1. **Redundancy of Subsetting by Gender:** Both subsets produced identical results with perfect fits (R-squared = 1) and identical coefficients for the predictors. This suggests that gender does not provide additional insights since the dependent variable, average_score, is mathematically derived from the independent variables.
2. **Deterministic Nature of the Model:** The perfect fit and extremely large t-statistics indicate that the model is solving a trivial equation, as average_score is directly calculated from the predictors, leaving no unexplained variance.

Visualizations

A scatterplot with regression lines by gender:



Regression Lines: Two separate lines are fitted to the data, one for each gender, showing the relationship between the math score and the average score.

Interpretation:

- **Positive Relationship:** Both regression lines slope upward, indicating that as math scores increase, average scores also increase for both genders.
- **Deterministic Model:** The graph suggests the model has a nearly perfect fit, meaning the average score is likely calculated directly from predictors like the math score, leaving minimal unexplained variance.
- In conclusion, the graph. Visually compares how math scores predict average scores for males and females.

Conclusion

This analysis explored the impact of gender, race or ethnicity, parental education, lunch status, and test preparation on student performance. Key findings include:

1. **Key Influences on Performance**:

- **Socioeconomic Status**: Students with standard lunch performed better, indicating socioeconomic disparities.
- **Test Preparation**: Completing a test preparation course significantly boosted scores.
- **Parental Education**: Higher parental education was linked to better student performance.
- **Racial/Ethnic Differences**: Variations in performance suggest potential cultural or systemic factors.

2. **Regression Analysis**:
    - The model showed a modest explained variance, indicating other unmeasured influences.
    - Gender subsetting produced identical results, suggesting gender does not significantly impact the relationship between individual scores and the average score.
3. **Implications**:
    - Key factors influencing academic performance include socioeconomic status, parental education, and test preparation.
    - The deterministic nature of the model highlights the need for further analysis to explore additional influences on student performance.

References:

1. Bluman, A. G. (2017). **Elementary statistics: A step-by-step approach** (10th ed.). McGraw-Hill Education.
2. Field, A. (2013). **Discovering Statistics Using IBM SPSS Statistics** (4th ed.). Sage Publications.

3. Dataset: https://www.kaggle.com/datasets/muhammadroshaanriaz/students-performance-dataset-cleaned