**Exploring the Correlation between GRE and First-Year GPA**

**with the Moderating Role of Demographic Factors**

Ziyu Ren

*University of Minnesota*

PSY 8712

**Exploring the Correlation between GRE and First-Year GPA**

**and the Influence of Demographic Variables**

In the current project, I explored the distribution of GRE scores and 1st-Year graduate school GPA, together with their correlations. I showed the varying GRE-GPA correlations across different demographic groups with a shiny app. Machine learning models are built to predict GPA from GRE and demographic variables.

**Hypotheses and Research Questions (RQ)**

I explored the following three research questions.

**H1**: Is there a significant sex difference of GRE Verbal / Quantitative score?

**RQ1**: What are the correlations between GRE Verbal / Quantitative score and GPA across sex?

**RQ2**: How well can I predict 1st-Year GPA with GRE and demographic information using machine learning models?

**Method**

**Open Science Materials**

A binder is an online code repository contains code and contents, together with configuration files to create a project from the scratch. When saving an R project in a binder, it saves the R version being used to create the project to avoid possible confusions in future replications. The binder for the current project can be found online via the link <https://mybinder.org/v2/gh/Lindsey-R/psy8712-final/HEAD?urlpath=rstudio> . Clicking this link will automatically open an R studio session built under R 4.2.3 (the version I use to run the project) with corresponding codes and data.

Similarly, material to create the current project can be accessed through GitHub following the link <https://github.com/Lindsey-R/psy8712-final> . Clicking on this link will lead to the GitHub page will all files, a README file with project descriptions can be accessed by screwing done to the bottom of the page.

**Participants**

The current project used pre-collect publicly available data from OpenICPSR (<https://www.openicpsr.org/openicpsr/project/155721/version/V1/view;jsessionid=CCA87775E2BAE63EE1B4FC92FF3AE409>). Participants are 3538 de-identified graduate students (41.2% female) from IVY League universities.

**Measures**

Below are variables used in the current analysis and their corresponding measures.

**Sex** Male / Female as reported by students.

**Citizenship** US / International Citizens as reported by students.

**GRE Quantitative** GRE Quantitative scores.

**GRE Verbal** GRE Verbal Scores

**GRE Writing** GRE Writing Scores

**GraduateFieldProgram** The student’s major in their college.

**GPA**  GPA as reported.

**Stay** Whether the student stay in the program or not (i.e., dropout).

**Procedure**

Given that the data is obtained from online sources, the procedure to collect data is unknown to me. Likely self-report is used for demographic information, and GRE and GPA are obtained from school profile.

**Analysis**

**Descriptive Statistics and Static Visualizations**

Table 1 shows the descriptive data for overall scores of GRE (Verbal, Quantitative, Writing) and GPA. A detailed table of descriptive data for each demographic group can be found online through the GitHub links.

**Table 1.**

*Descriptive data of GRE and GPA.*

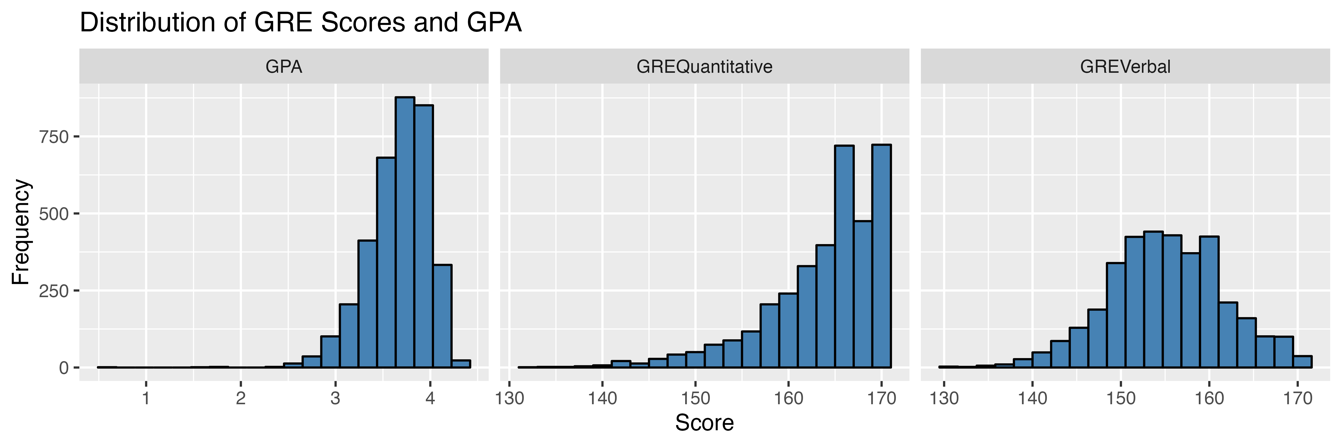
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | ***N*** | **Mean** | **SD** |
| GRE Quantitative | 3538 | 164.11 | 5.98 |
| GRE Verbal | 3538 | 155.07 | 6.55 |
| GRE Writing | 3538 | 3.68 | 0.69 |
| GPA | 3538 | 3.67 | 0.32 |

*Note: N* = total sample size.

Figure 1 shows the distribution of GPA, GRE Quantitative, and GRE Verbal scores. While GPA and GRE Quantitative scores are skewed to the left, GRE Verbal is roughly normal skewed.

**Figure 1.**

*Descriptive plot of GRE and GPA.*



**Interactive Visualization**

An online shiny app (<https://purplefishlovespig.shinyapps.io/greshinyapp/>) was created for data visualization. The app was created as an extension of RQ1, with interacting plots showing correlations based on different demographic groups. Specifically, it allows user to choose from five options: Sex, Citizenship, Major (Graduate Field Program), whether student finished program (Stay), and GRE (GRE Verbal, Quantitative, or Sum score). Based on the user’s choice, the app returns 1) a scatter plot showing the correlations between the selected GRE and GPA, with a linear predicting line and 2) a line of text output the correlation.

**Data Cleaning**

I kept all observations from participants who report GRE and GPA scores. In addition, participants without information about whether they stay with the program or not were removed. Then, I selected all variables of interest: Sex, Citizenship, GRE scores, GPA, Graduate Field Program, and Stay. Finally, I kept majors with more than 300 students for analysis.

**Analysis**

***H1: Is there a significant sex difference on GRE Verbal / Quantitative score?***

To test the current hypothesis, I conducted a two-sample t-test comparing male and female scores on GRE Verbal and GRE Quantitative scores, separately. The null hypothesis is that there is no significant difference, while the alternative hypothesis is that there is significant difference. For both GER Verbal and Quantitative, the null hypothesis was rejected. There was a significant difference of GRE Quantitative scores between male and female, t(2828) = -6.84, p = .00. There was a significant difference of GRE Verbal scores between male and female, t(3239) = -4.08, p = .00. Figure 2 shows a visualization of the mean GRE score differences across sex.

**Figure 2.**

*Boxplot of mean GRE score differences across sex.*

A graph of different colored squares

Description automatically generated with medium confidence

***RQ1: What are the correlations between GRE Verbal / Quantitative score with GPA for male and female, separately?***

Table 2 shows the correlation pattern between GRE scores and GPA for male and female.

**Table 2.**

*Correlation between GRE and GPA.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Sex** | **Cor\_QG** | **Cor\_VG** | **Cor\_QV** |
| Female | 0.18 | 0.19 | 0.13 |
| Male | 0.27 | 0.14 | 0.20 |

*Note: Cor\_QG = correlation between GRE Quantitative and GPA; Cor\_VG = correlation between GRE Verbal and GPA; Cor\_QV = correlation between GRE Quantitative and GRE Verbal.*

As shown from the table, both male and female group showed moderate correlations between GPA and GRE. Male showed higher correlation between GRE Quantitative and GPA, and higher correlation between GRE Quantitative and Verbal scores. Correlations among GRE and GPA separated by different groups are interactively shown in the shiny app (mentioned above).

***RQ2: How well can I predict 1st-Year GPA with GRE and demographic information using machine learning models?***

I predicted GPA using the following variables: Sex, Citizenship, GRE Quantitative, GRE Verbal, GRE Writing, Graduate Field Program, and Stay. Machine learning with 10-fold CV and holdout CV are conducted. Four algorithms are used in the current prediction: OLS model, Elastic Net model, Random Forest model, and xgbLinear model. Table 3 shows the resulting cv\_R2 and ho\_R2 of these models.

**Table 3.**

*Machine Learning predicting results.*

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **cv\_R2** | **ho\_R2** |
| OLS regression | 0.13 | 0.10 |
| Elastic Net | 0.13 | 0.10 |
| Random Forest | 0.71 | 0.11 |
| xgbLinear | 0.74 | 0.07 |

*Note: N* = total sample size.

In general, although Random Forest and xgbLinear performed well in 10-fold CV, all models predicted GPA poorly with very low R2 for hold-out CV.

**Reflection**

One major takeaway I have from data science is the process of learning new coding skills: (possibly) data camp, the manual for package, practices, and get aids from ChatGPT. I will bring this to my future research every time when I need to learn about a new skill / new package. For important change is that I’m now better with cleaning up dataset with tidyverse and work within the tidy format with pipelines. Previously I just use any code that could work to clean up my data. The most valuable thing I gain is that I’m now more confident about myself learning advanced data science skills. Things like ML, NLP, and using MSI used to seem very hard, but I found out they are easy to learn and feel confident to use them for future research.