ANA 605: Week One Assignment

If you have questions about the following instructions or about your assignment, please send me an email with a description of your question and what you've tried in attempt to answer it. Be sure to include your data file and R script. Please do NOT submit late assignments; they will not be accepted after answers are posted.

Data Description

This data was downloaded from Kaggle.com, a site that houses open source datasets. This specific dataset is titled: "Graduate Admission 2." From the Kaggle website (below):

https://www.kaggle.com/mohansacharya/graduate-admissions

Context
This dataset is created for prediction of Graduate Admissions from an Indian perspective.

Content
The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are : 1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1)

Acknowledgements
This dataset is inspired by the UCLA Graduate Dataset. The test scores and GPA are in the older format. The dataset is owned by Mohan S Acharya.

Inspiration
This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.

Citation
Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

Name: **Jason Noble**

Assignment Questions

1. For each of the following explanatory variables, fit a separate regression and correlation for the outcome variable, **chance of admit**. Fill in the following cell with the appropriate value for each element of the model.

Explanatory Variable	r	b_0	b_1	df	SS Model	SS Total	PRE	F
Research Experience	N/A	0.638	0.158	1	2.483	8.115	0.306	175.514
Undergraduate GPA	0.873	-1.072	0.209	1	6.188	8.115	0.763	1278.734
GRE Scores	0.803	-2.436	0.010	1	5.227	8.115	0.644	720.554
TOEFL Scores	0.792	-1.273	0.019	1	5.085	8.115	0.627	667.941

2. Interpret the values above in the table below.

Explanatory Variable	Term	Interpretation
Research Experience	r	N/A - Cannot calculate correlation for categorical data
	b_0	The expected chance of admission for applicants with no research experience is approximately 0.638
	b_1	Applicants that did have research experience tended to have an additional 0.158 higher chance of admission than those without
	PRE	Research experience explains 30.6% of the variation in admission chances within our dataset
GRE Scores	r	Higher GRE scores are strongly associated with a higher chance of admission
	b_0	The expected chance of admission when the GRE score is zero is approximately -2.436, which lacks practical meaning, but is needed for calculating predictions. (Because the GRE range is 290-340, the b_0 is a projection to 0, and with the slope associated to b_1, b_0 is negative)
	b_1	For each additional point on the GRE, the chance of admission is predicted to increase by 0.010
	PRE	GRE score explains 64.4% of the variation in admission chances within our dataset

3. Which model is the best, according to PRE? Why?

Undergraduate GPA, with a PRE value of 0.763, indicates 76.3% of the variance is explained by this model, which is the highest among the given models. The high PRE value suggests that Undergraduate GPA is a strong predictor of the outcome, reducing the error more than the other variables.

4. Which model is best, according to the F Ratio? Why?

Based on the F-ratio, again, the GPA is the best. The high F-statistic of 1278.734, indicates a strong relationship between GPA and admission chance.

5. Which model explains more variation, per df used? Why?

All the models use 1 degree of freedom, GPA is still the best fit; because it has the highest F-statistic per df.

6. Obtain predictions for the following respondents for each of the explanatory variables below. Based on your analyses, who do you believe has the best chance of getting admitted? Why?

Explanatory Variable	Respondent #40	Respondent #56	Respondent #78	Respondent #93
Research Experience	0.6376796	0.6376796	0.6376796	0.6376796
Undergraduate GPA	0.5366120	0.5366120	0.6452126	0.6055316
GRE Scores	0.6265115	0.7561980	0.5666562	0.5367286
TOEFL Scores	0.7353236	0.6423271	0.5679299	0.5493306

Subject #56 has the best chance of getting admitted. With the highest predicted chance from the GRE model (0.7562; highest scores are highlighted) and a relatively high TOEFL (0.6423), indicating Subject #56 is consistent across multiple indicator models for admissions. Calculated the numbers to verify highest overall prediction score.

7. For the best model that predicts the outcome variable per df used (see Q5), edit the following equation (see GLM notation) with inputted values that you would use to make predictions with.

$$Y_i = -1.072 + 0.209 \times X_i + e_i$$

8. Input the values (those subscripted by i in the model equation) for respondent i = 100 into the equation you edited in Q7.

$$.79 = -1.072 + 0.209 \times 8.88 + e_i$$
$$.79 = 0.783 + e_i$$
$$.79 - .783 = e_i$$
$$e_i = .007$$

A small difference between the actual and predicted values, suggesting the model prediction is quite close to the actual outcome for respondent 100.