ANA 605: Week Four Assignment

Due Saturday, October 26th, 2019, at 11:59 PM, Pacific time

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

Assignment Questions

1. For each of the following variables contained in the below table,
    a. Get the frequencies for each categorical variable and means/sd for each quantitative variable.

Variable	If quantitative, Mean If categorical, Frequency group 0	If quantitative, SD If categorical, Frequency group 1
Research Experience	Freq: 181, 45%	Freq: 219, 55%
Undergraduate GPA	8.6	0.6
Letter of Recommendation (LOR)	3.45	0.9
Statement of Purpose (SOP)	3.4	1.01

    b. Get correlations (and p-values) between the outcome, **chance of admit,** and explanatory variable. Put those values in the table below, labeled r (p).
    c. Perform a multiple regression analysis for the outcome variable, **chance of admit,** and fill in the rest of the table below.

Model Equation:

$$Chance_i = b_0 + b_1 * Research_{1i} + b_2 * GPA_{2i} + b_3 * LOR_{i3} + b_4 * SOP_{i4} + e_i$$

PRE =	0.787	F =	364.798
df1 =	4	DF =	395
p	< 0.0001	b0 =	-0.8317

Explanatory Variable	r (p)	b_1 (p)	Lower bound 95% CI	Upper bound 95% CI
Research Experience	N/A	0.0355 (0)	0.02	0.05
Undergraduate GPA	0.8733 (0)	0.1693 (0)	0.15	0.19
Letter of Recommendation (LOR)	0.6699 (0)	0.0218 (0.0001)	0.01	0.03
Statement of Purpose (SOP)	0.6757 (0)	0.0017 (0.7531)	-0.01	0.01

2. Interpretations of the values above.

Model Parameters

Parameter	supernova() Interpretations
PRE	The PRE value of 0.7870 suggests the model explains approximately 78.7% of the variance in the chance of admission. This is a strong indication that the independent variables collectively have a large impact on predicting admission. This model significantly reduces the error in predicting the outcome compared to the null model that only uses the mean for the chance of admission variable.
F	The high F-statistic of 364.8 suggests the model as a whole is statistically significant. So, at least one of the included variables is significantly related to the chance of admission. The high F-value indicates that the variance explained by the model is much greater than the unexplained variance (error), making it highly unlikely that these results are due to random chance.
p-value	The associated p-value is extremely small (2.2e-16), well below the 0.05 threshold. This suggest a strong evidence against the null hypothesis, which assumes that none of the variables impact the chance of admission. Since the p-value is so low, and F high, we can reject the null hypothesis and conclude that the model provides a statistically significant explanation of the variance seen in chance of admission.

Correlations:

Explanatory Variable	Correlation Interpretations, r (p)
Research Experience	On average, applicants with research experience tend to have a higher chance of admission compared to those without research experience.
Undergraduate GPA	Applicants with higher undergraduate GPAs are associated with a higher chance of admission.
Letter of Recommendation (LOR)	Stronger letters of recommendation are generally associated with a higher chance of admission.
Statement of Purpose (SOP)	The strength of the SOP is associated with a minimal increase in the chance of admission.

Multiple Regression:

Explanatory Variable	Regression Weights Interpretations, b_0 and b_1 (p)
Intercept, b_0	The intercept represents the expected chance of admission for an applicant who has no research experience, 0.0 CGPA, no LOR, and no SOP. In this case, the expected chance of admission is negative, -0.83, which doesn't make practical sense, but with a minimum GPA of 6.8 within the data, suggests that without strong qualifications, the chance of admission is very low. As a more practical starting point, a predicted chance of admission for an applicant with a CGPA of 6.4 and 0 for Research, LOR, and SOP is approximately 0.25 (25%).
Research Experience	Applicants with research experience (coded as 1) on average, increase their chance of admission by 3.55% over those without (coded 0), when other variables are held constant. Suggesting that research experience is valued a little in the admission process.
Undergraduate GPA	For each unit increase in CGPA (7.0 to 8.0), the predicted probability of admission increases by 16.93%, when other variables are held constant. This indicates that higher GPAs strongly improve admission chances.
Letter of Recommendation (LOR)	For each increase in the strength of the letter of recommendation, the chance of admission is expected to increase by 2.18%. Stronger letters positively influence admission chances, but not by a large amount, when other variables are held constant.
Statement of Purpose (SOP)	The strength of the statement of purpose has a minimal effect on the chance of admission, with an estimated increase of 0.17% for each unit increase in SOP strength, when other variables are held constant (not statistically significant). With a p-value over the .05 threshold and 0 included within the confidence intervals (-0.01 to 0.01) we would accept the null hypothesis on this variable. This reiterates that SOP is not a major factor in the admissions process.

3. Fit a reduced model after removing explanatory variables that you believe do not contribute to the model. Revise the following model equation and fill in the table for the new model that has only those variables that were used (remove rows/terms as needed).

Model Equation:

$$Chance_i = \quad b_0 \quad + \quad b_1 * GPA_{1i} \quad + \quad e_i$$

PRE =	0.7626	F =	1279
df1 =	1	DF =	398
p	< 0.0001	b0 =	-1.0715

Explanatory Variable	b_1 (p)	Lower bound 95% CI	Upper bound 95% CI
Undergraduate GPA	0.2089	0.1974	0.2203

4. Which model is best: the first model with four explanatory variables, or the reduced model with only those that contribute to the model? Why?

The second model, only including GPA is the best fit: It is simpler, including only the more statistically significant variable. It provides a similar PRE (0.787_4 to 0.7626_1) while only utilizing 1 degree of freedom to produce a much higher F-value (365_4 to 1279_1). Excluded the non-significant SOP variable. Ran the one variable model and then tested results against the two and three models. The model with CGPA and Research maintained predictive accuracy, improving the PRE by about 1%, but reducing F-stat by 592. The analysis with CGPA, LOR and Research again improved the chance of admission by about 1%, but reduced F-stat by 104. In terms of cost of effort it is obvious to focus only on undergraduate GPA. However, in the competitively tight environment of graduate admissions, it would be logical to submit a strong letter of recommendation, and participate in research, those incremental increases in admission chance could make the difference. But focus should be primarily on high grades.

5. How many parameters are in the multiple regression model in Q1?

4 coefficients and 1 intercept; total parameters = 5

6. How many parameters are in the reduced model from Q3?

After removing the non-significant SOP variable, the LOR variable and the Research variable:
1 coefficients and 1 intercept; total parameters = 2

7. How did the **parameter estimates** change as a result of removing explanatory variable(s)? Why do you believe this occurred?

In the four-variable model, the coefficient for CGPA was 0.1693.
In the one-variable model, the coefficient for CGPA increased to 0.2088.
Research, LOR, and SOP were accounting for some portion of the total variance in chance of admit. After removal, the model now attributes all of this variance to CGPA, leading to a higher coefficient.

8. How did your **interpretations** change as a result of removing explanatory variable(s)? Why do you believe this occurred?

The removal of the other variables from the model leads to an increase in the estimated effects of CGPA on the chance of admission, at almost 21%. This occurs because the variance that Research, LOR and SOP previously explained is reevaluated to only one variable, making it appear more influential in the simplified model. This illustrates the tradeoff made between a simple model and a more complex one. A simple model will approximate the answer, but if we really wanted to understand the chance of admission, we would also want to include all the other variables, including those not looked at in this paper (TOEFL, GRE, university). Especially if we wanted to make an accurate prediction of one application.

Extra credit:

EC1. What is the predicted chance of admission for respondent #234, using the multiple regression model from Q1?

$$\text{Chance } i = -0.8317 + (0.0355 \times 0) + (0.1693 \times 8.07) + (0.0218 \times 3.5) + (0.0017 \times 2.5)$$

Chance i = 0.61495, 61.5% is slightly lower than the 0.64 calculated for that respondent, indicating that all the other variables not in Q1(GRE, TOEFL) account for about 0.025 of the outcome.

EC2: What is the residual for respondent #128, using the multiple regression model from Q1?

$$\text{Predicted Chance } i = -0.8317 + (0.0355 \times 1) + (0.1693 \times 8.71) + (0.0218 \times 2.0) + (0.0017 \times 2.5)$$
Predicted Chance I = 0.7258

Residual = Actual Chance of Admit − Predicted Chance of Admit
Residual = 0.78 − 0.7258

Residual = 0.0542, so the actual chance is about 5.4% higher than what the Q1 model predicted.