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

Completed By
Randall C. Crawford 11/19/2024

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	181	219
Undergraduate GPA	8.599	0.596
Letter of Recommendation (LOR)	3.453	0.898
Statement of Purpose (SOP)	3.400	1.007

    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__ df = __4__ p = __0.000__ b0 = __-0.832__

Explanatory Variable	r (p)	b_1 (p)	Lower bound 95% CI	Upper bound 95% CI
Research Experience	N/A (0)	0.036 (0)	0.020	0.051
Undergraduate GPA	0.873 (0)	0.169 (0)	0.152	0.187
Letter of Recommendation (LOR)	0.670 (0)	0.022 (0)	0.011	0.033
Statement of Purpose (SOP)	0.676 (0)	0.002 (0.753)	-0.009	0.012

2. Interpretations of the values above.

When interpreting a correlation between two *quantitative* variables, do not indicate that the coefficient is positive or negative; rather, use a description of the variables involved, like "A greater distance from home is associated with more commuting time."

When interpreting a correlation between a *categorical* and *quantitative* variable, do not indicate that the coefficient is positive or negative; rather, use a description of the variables involved, like "Those who were in a romantic relationship tended to have more absences from school than those who were not in a romantic relationship."

When interpreting b_0 with multiple explanatory variables (categorical and quantitative), describe what is the expected outcome, for a person who have a **predicted score of 0 for all explanatory variables.**

When interpreting b_1 for *quantitative* explanatory variables, describe whether the outcome variable is expected to increase or decrease, and by how much, for each unit change in the explanatory variable, **controlling for other variables in the model.**

When interpreting b_1 for *categorical* explanatory variables, describe whether there is a mean difference (direction and amount) in the outcome variable, between the group coded 0 and the group coded 1 (use names for the groups; you must know the coding!), **controlling for other variables in the model.**

Model Parameters

Parameter	supernova() Interpretations
PRE	The suggested linear model explains 78.7% of the variation in the empty model for the *Chance.to.Admit* outcome variable.
F	The suggested linear model explains almost 365 times the variance of the empty model for the *Chance.to.Admit* outcome variable.
p-value	With the suggested linear model, the probability of getting an estimate as extreme or more extreme than the sample estimate, given the assumption that the empty model for the *Chance.to.Admit* outcome variable is true, is near zero, notably less than the required 0.05 alpha standard. Consequently, the suggested linear model is deemed significant, and the empty model could be rejected.

Correlations:

Explanatory Variable	Correlation Interpretations, r (p)
Research Experience	Possessing research experience is associated with a slightly improved chance for admission.
Undergraduate GPA	A higher GPA is definitely associated with a greater chance for admission.
Letter of Recommendation (LOR)	A stronger letter of recommendation is associated with a modestly improved chance for admission.
Statement of Purpose (SOP)	A stronger statement of purpose does not seem to have much bearing on the chance for admission.

Multiple Regression:

Explanatory Variable	Regression Weights Interpretations, b_0 *and* b_1 (p)
Intercept, b_0	The chance for admission estimation is -83.2%, when an applicant has zero value for all of the explanatory variables.
Research Experience	The mean difference between not having (0) and having (1) research experience is 0.036, respecting other variables in the model. With research experience, a 3.6% chance improvement is added to the chance of admission estimation
Undergraduate GPA	For every GPA point (10 max), a 16.9% chance improvement is added to the chance of admission estimation.
Letter of Recommendation (LOR)	For every letter of recommendation strength point (5 max), a 2.2% chance improvement is added to the chance of admission estimation.
Statement of Purpose (SOP)	For every statement of purpose strength point (5 max), a 0.2% chance improvement is added to the chance of admission estimation.

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:

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

PRE = __0.787__ F = __487.474__ df = __3__ p = __0.000__ b0 = __-0.838__

Explanatory Variable	b_1 (p)	Lower bound 95% CI	Upper bound 95% CI
Research Experience	0.036 (0)	0.020	0.051
Undergraduate GPA	0.170 (0)	0.155	0.186
Letter of Recommendation (LOR)	0.023 (0)	0.013	0.034

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 reduced model is best. The statement of purpose variable was unnecessary, not significant due to a high p-value of 0.7535, and b1 confidence interval (-0.009, 0.012) that included zero. The PRE value remained the same, which was another indicator that the statement of purpose variable was contributing nothing. The reduced linear model explains approximately 487 times the variance of the empty model for the *Chance.to.Admit* outcome variable. This is about 122 times more than the first model.

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

There are five parameters in the Q1 model.

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

There are four parameters in the Q3 model.

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

After removing that statement of purpose variable, thus reducing the degrees of freedom from 4 to 3, the reduced model's F-Ratio increased approximately 122. The reduced linear model explains approximately 487 times the variance of the empty model for the *Chance.to.Admit* outcome variable. The PRE and p-value remained the same for both models.

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

There were only slight changes to b0/b1 coefficients and confidence intervals. This is another clear indicator that the statement of purpose variable was unnecessary.

Extra credit:

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

Respondent #234 Prediction for Q1 Model: 0.615

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

Respondent #128 Residual for Q1 Model: 0.054