Assignment 08

Polychotomous Categorical Predictors

This goal of this assignment is to give you more experience fitting and interpreting regression models with categorical predictors. Submit your responses to each of the questions below in a printed document. All graphics should be resized so that they do not take up more room than necessary and also should have an appropriate caption. This assignment is worth 15 points. (Each question is worth 1 point unless otherwise noted.)

For this assignment, you will examine whether there are differences in average in-state tuition between Minnesota and its bordering states. To do so, you will use the file *colleges-bordering-mn.csv* (see the data codebook).

Preparation: Create Dummy Variables

Create five dummy variables, one for each state, for the analysis.

Description

1. Create a table of pairwise correlations between average in-state tuition, each of the five state dummy variables, sector (public), 75th-percentile ACT score, and percentage of students on Pell grants.
2. Interpret (i) the correlation between the Minnesota state dummy variable and average in-state tuition, and (ii) the correlation between the Minnesota state dummy variable and the sector (public) variable.

Unadjusted Group Differences Model: ANOVA

Fit the regression model that uses the dummy predictors for state to predict variation in average in-state tuition. In this model, use Minnesota as the reference group.

3. Write the fitted regression equation.
4. Which states, if any, differ from Minnesota in the average in-state tuition (more than we expect because of sampling variation)? Explain
5. Report and interpret the R^2 value for this model.
6. Which state comparisons are not represented in this fitted model that are reflected in the omnibus null hypothesis?

Adjusted Group Differences Model: ANCOVA

Again, fit the regression model that uses the dummy predictors for state to predict variation in average in-state tuition, but this time control for differences in (1) sector (public vs. private), (2) ACT scores, and (3) percentage of students on Pell grants. Again, use Minnesota as the reference group.

7. Write the fitted regression equation.
8. Which states, if any, differ from Minnesota in the average in-state tuition (more than we expect because of sampling variation) after controlling for these other predictors? Explain
9. Report and interpret the R^2 value for this model.

Assumptions

10. Create the density plot of the marginal distribution of the standardized residuals from the ANCOVA model. Add the confidence envelope for the normal distribution. Explain whether or not this plot suggests problems about meeting the normality assumption.
11. Create the scatterplot of the standardized residuals versus the fitted values from the ANCOVA model. In the plot identify observation with extreme residuals (≤ -3 or ≥ 3) by indicating the row number of that observation in the plot.
12. Explain whether or not this plot suggests problems about meeting the linearity and homogeneity of variance assumptions.

Pairwise Differences

13. Create a table (suitable for publication) that presents each of the possible pairwise contrasts (null hypotheses) of interest, the unadjusted p-values, and the Benjamini–Hochberg adjusted p-values for the controlled differences. (Note: To obtain all of these, you may need to fit additional models.)
14. Based on the Benjamini–Hochberg adjusted p-values, which states differ in their average in-state tuition (more than we expect because of sampling error) after controlling for the other predictors in the model?
15. Create a heatmap of the information that you reported in the table in Question 8. You can see an example of a heatmap for correlations here. We want to create a heatmap that shows the p-value for each mean comparison instead of the correlation between variables. So for example, the heatmap we want to create should be a 5x5 grid (the rows and columns would represent states) and the intersecting cells would include the p-value for the comparison (rather than the correlation coefficient). Color will be used to indicate the magnitude of the p-values. You may want to include different levels of color depending on the degree of significance ($< .05$, $< .01$, etc. Feel free to use any software tool you want to create this heatmap.