DATA 231

HOMEWORK 2

In this homework we will continue working with the US States data set investigated in Homework 1.

**Part A: relationship between a state’s % of College graduates and % of High School graduates**

1. Use R to fit the regression line predicting College using HighSchool.
2. Interpret the value of the slope in context.
3. Produce a scatterplot that includes the least squares line. Does the regression line seem to fit the data well? Are there any obvious outliers or influential points in this plot?
4. What percent of the variation in College graduate rate es explained by the simple linear model with HighSchool as the predictor?
5. Produce 3 residual plots: the residual vs. fits plot, the normal probability plot, and a histogram of the residuals. Discuss what these plots tell you about potential problems (if any) with the regression conditions.

**Part B: a transformation on College**

1. Use R to fit the regression line predicting log(College) using HighSchool.
2. Check the residual plots. Does this model meet the conditions better than the one in Part A?
3. Is the slope of the least squares line predicting log(College) from HighSchool significantly different from 0? Show code to support your answer.
4. Construct and interpret a 90% confidence interval for the slope coefficient in this model.
5. Provide the ANOVA table that partitions the total variability in College graduation rate and interpret the F-test.

**Part C: use a different variable**

1. There are other variables in this data set that might be better predictors of College graduation rate than HighSchool. Find one and explain why you chose it.
2. Find and interpret a 95% confidence interval for the mean College graduation rate at the median of your chosen variable.
3. Find and interpret a 95% prediction interval for a state’s College graduation rate at the median of your chosen variable.

**Article Critique:**

Read the following article:

“Examining the Effects of Student Involvement   
on African American College Student Development”

You can find this article on the Moodle page in the Folder “Articles” (it’s called “student involvement - regression.pdf”).

After reading, briefly **critique** the author’s use of Simple Linear Regression.

You’ll want to address questions like…

Is a linear regression model appropriate in this context? Does the author discuss any diagnostic plots or mention whether the assumptions of the linear regression model are met? Is the sample representative of the larger population about which he makes conclusions? What inference methods did he perform? Are his conclusions valid for the tests that he performed? Do you feel there are any limitations to his study that he didn’t explicitly discuss?

…But you should feel free to talk about any other issues/concerns that you have about his analysis.