Modeling Assignment 2: Fitting and Interpreting Simple Linear Regression Models

Assignment Overview

Every dataset has a “story” to tell. It just doesn’t have the voice to speak the story. In a sense, it is your job as the data analyst to “tell” the story that the data has to offer. To do this, you have a collection of tools, like the descriptive statistics and graphical methods of EDA, at your disposal. Now, you have added correlation and simple linear regression to your collection. In this assignment, you are to use these tools to start to uncover the "story" the US State data set has to tell. The US State data set (USStates.xlsx) is a 12 variable dataset with n=50 records. The data, calculated from census data, consists of state-wide average or proportion scores. As such, higher scores for these variables translate into having more of that quality. There is no other information available about this data.

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign3_YourLastName.pdf.

Assignment Tasks

1. Given the variables in this dataset, which variables can be considered explanatory (X) and which considered response (Y)? Can any variables take on both roles? Make a table that summarizes your conclusions.

2. What is the population of interest for this problem (yes – this is a trick question!)? Be sure your answer is clear and complete.

3. For the duration of this assignment, let's have HOUSEHOLDINCOME be the response variable (Y). Also, consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain all possible scatterplots relating the non-demographic explanatory variables (Xs) to the response variable (Y).

4. Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with the response variable Y and report the correlations in a table. Given the scatterplots from step 3) and these correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or why not?

5. Fit a simple linear regression model to predict Y using the COLLEGE explanatory variable. Use the base STAT lm(Y~X) function. Why would you want to start with this explanatory variable? Call this Model 1. Report the prediction equation for Model 1 and interpret each coefficient of the model in the context of this problem. In addition, report and interpret the R-squared statistic for Model 1.

6. From your Model 1 results for task 5) – Specify the null and alternative hypothesis separately for each of the two parameters in the model. Report and interpret the results of the T-tests for these hypotheses. In addition, state the null and alternative hypotheses for the omnibus (i.e. overall) model. Report the ANOVA table and interpret the results of the F-test.

7. For Model 1, write R-code to create a variable of predicted values based on your Model 1 prediction equation from task 5. Use the predicted values and the original response variable Y to create a variable of residuals (i.e. residual = Y – Y_hat = observed minus predicted) for Model 1. Using the original Y variable, the predicted, and/or residual variables, write R-code to:
    - Square each of the residuals and then add them up. This is called sum of squared residuals, or sums of squared errors.
    - Deviate the mean of the Y's from the value of Y for each record (i.e. Y – Y_bar). Square each of the deviations and then add them up. This is called sum of squares total.
    - Deviate the mean of the Y's from the value of predicted (Y_hat) for each record (i.e. Y_hat – Y_bar). Square each of these deviations and then add them up. This is called the sum of squares due to regression.
    - Calculate a statistic that is: (Sum of Squares due to Regression) / (Sum of squares Total)

Verify and note the accuracy of the ANOVA table and R-squared values from the regression printout from part 4), relative to your computations here. Report your R-code for these computations.

8. From task 7 you created a variable of residuals for Model 1. Write R-code to standardize the residuals. Do not use residuals from the lm(). Plot the standardized residuals using a histogram. Also, plot the standardized residuals in a scatterplot with the predicted values. Discuss what you see in these two graphs.

9. Select a different explanatory variable and use that variable in a Simple Linear Regression model to predict Y, HOUSEHOLDINCOME. Call this Model 2. Report and interpret the results of Model 2. Which is the better model, Model 1 or Model 2? Give evidence to justify your answer.

10. For this last task, you are welcome to fit any Simple Linear Regression model that you wish on the US States data. You'll need to decide on the response variable as well as the explanatory variable. Call this Model 3. Report and interpret the results of Model 3.

.