

|  |
| --- |
| Assignment #1: OLS Linear Regression  *MSHA 410* |

***Modeling the US States Data***

**Data:** The data for this assignment is the US State data set: USStates.CSV. It is a 12 variable dataset with n=50 records. The data, calculated from census data, consists of state-wide average or proportion scores for the non-demographic variables. As such, higher scores for the composite variables translate into having more of that quality. There is no other information available about this data.

**Objective:** 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 analyst to “tell” the story that the data has to offer. That is your objective here: To uncover the story this dataset has to tell.

**Tasks:** To achieve the objective please complete the following tasks enumerated below. You are to use R to obtain any graphs or statistics requested.

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? What is the population of interest for this problem (yes – this is a trick question!)?

2. For the duration of this assignment, let’s have HOUSEHOLDINCOME be the response variable (Y). Also, please 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 to the response variable (Y).

3. Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with Y and report the correlations in a table. Given the scatterplots from step 2) and the correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or why not?

4. 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 results of Model 1 in equation form and interpret each coefficient of the model in the context of this problem. Report the ANOVA table and model fit statistic, R-squared. Use the summary statistics from steps 2) and 3) to verify, by hand computation, the estimates for the slope and intercept.

5. Write R-code to calculate and create a variable of predicted values based on Model 1. Use the predicted values and the original response variable Y to calculate and 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.

6. Fit a multiple linear regression model to predict Y using COLLEGE and INSURED as the explanatory variables. Use the base lm(Y~X) function. Call this Model 2. Report the results of Model 2 in equation form, interpret each coefficient of the model in the context of this problem, and report the model fit statistic, R-squared. How have the coefficients and their interpretations changed? Calculate the change in R-squared from Model 1 to Model 2 and interpret this value. For this specific problem, is it OK to use the hypothesis testing results to determine if the additional explanatory variable should be retained or not? Think statistically using first principals. Discuss. NOTE: The topic of hypothesis testing in regression is the focus of Module 2 – you should NOT need to read anything about hypothesis testing to answer this.

7. In a sequential fashion, continue to add in the non-demographic variables into the prediction model, one variable at a time. Make a table summarizing the change in R-squared that is associated with each variable added. Based on this information, what variables should be retained for a “best” predictive model? What criteria seems appropriate to you?

During this problem, practice interpreting coefficients for each model. Do any of the interpretations become counter intuitive as you fit more and more complex models? What does, or would, this mean for the model being developed? You do not need to report all of the coefficient interpretations, but this is a general question to contemplate and skill to use in model determination. Please write a short summary of your conclusions here.

8. Now that you have a sense of which explanatory variables contribute to explaining HOUSEHOLDINCOME, refit a model using only the set of variables you consider to be appropriate to model Y. Report this model, interpret the coefficients, and interpret R-squared in the context of this problem. Discuss why is it necessary to refit this model.

9. You are welcome to conduct any other analyses you wish to embellish your understanding of this dataset.

10. Given what you’ve learned from this modeling endeavor, what overall conclusions do you draw? What is the “Story” contained in this data? What have you learned? What are your Prescriptive Recommendations for action based on this evidence? Finally, feel free to reflect on what you’ve learned from a modeling perspective.

**Assignment Document:**

Results should be presented and discussed in the numerical order of the questions given. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document MUST be submitted in pdf format. Please use the naming convention: Assign1\_YourLastName.pdf.