CSCE 587 – Big Data Analytics

Homework Assignment 3: Linear and Logistic Regression

Due: Thursday, 2/27, by 11:59pm

Deliverables: You will submit an R script (a .R file) with commands to answer the questions below, and comments clearly labelling each answer, to the assignment in Blackboard. You will also submit PDFs with the plots. You may submit these as separate files, or you may compress them to a zip file and upload it.

The Data: The csv file attached to the assignment in provides the data to be modeled.

Preparation:

Use wget at the terminal to download the file from my CSE site to your virtual machine:

wget https://cse.sc.edu/~bhipp/587/hw3data.csv --no-check-certificate

Additional Instructions of Note:

Your answer for each question must be clearly labelled in your R script and must include the command(s) necessary to generate the answer. Unlabeled answers and answers that don't contain the R command(s) to justify them will not receive credit.

The R commands must be submitted in an R script (.R) file so that they can easily be run during grading.

Part A – Simple Linear Regression:

All Students

1. Generate a scatter plot with hrsTreatment as the x-axis variable and secResponse as the y-axis variable, and save the plot as a PDF. This PDF must be submitted with your assignment.
2. Based on the scatter plot, does there appear to be a correlation between the two variables? If so, would a simple linear regression model, a logistic regression model, or neither seem appropriate for predicting secResponse based on hrsTreatment? Explain your answer. Answer this question in a comment in your R script file.
3. Use the lm function to create a linear regression model predicting secResponse based on hrsTreatment. Show the summary statistics for the model.
4. Fill in the blanks with the correct values from the model as a comment in your R script file: the model predicts that secResponse will ____ by _____ with each additional hour of treatment.
5. Plot the Residuals vs Fitted, Q-Q Residuals, Scale-Location, and Residuals vs Leverage plots for the linear model and save the plot as a PDF. This PDF must be submitted with your assignment.
6. Explain what the Residuals vs Fitted plot tells us about the validity of the simple linear regression model for this dataset. Answer this question in a comment in your R script file.

Graduate Students Only

7. Create a linear regression model predicting secResponse using both hrsTreatment and minExercise as the two independent variables. Output the summary statistics for this multiple regression model.
8. Compare the multiple R^2 value for this model to the value from the single regression model found in question 3. Based on this value, does it appear that adding the second predictor variable is worthwhile? Answer this question in a comment in your R script file.

Part B – Logistic Regression:

All Students

1. Generate a scatter plot with hrsTreatment as the x-axis variable and hitTarget as the y-axis variable, and save the plot as a PDF. This PDF must be submitted with your assignment.
2. Based on the scatter plot, does there appear to be a correlation between the two variables? If so, would a simple linear regression model, a logistic regression model, or neither seem appropriate for modeling hitTarget based on hrsTreatment? Explain your answer. Answer this question in a comment in your R script file.
3. Using the same methodology used in the logistic regression lab (pages 9-14 of the Logistic Regression pdf in Blackboard) build a dataframe with the table statistics for hrsTreatment and hitTarget. Write the R command to show the summary statistics, which should yield the values shown below:

missedTarget		hitTarget		hrsTreatment	
Min.	:0.00	Min.	:0.000	Min.	: 2.700
1st Qu.	:0.00	1st Qu.	:0.000	1st Qu.	: 5.150
Median	:0.00	Median	:1.000	Median	: 7.600
Mean	:1.04	Mean	:1.267	Mean	: 7.636
3rd Qu.	:2.00	3rd Qu.	:2.000	3rd Qu.	: 9.950
Max.	:5.00	Max.	:5.000	Max.	:12.900

4. Use glm to build a logistic model and visualize the fit of the model using the lines function to add the fitted values of the model to a scatter plot of the probabilities of hitting the target (y-axis) for the values of hrsTreatment. Review the Logistic Regression lab (pages 9-14 of the Logistic Regression pdf in Blackboard) for a similar example. Save the plot as a PDF. This PDF must be submitted with your assignment.
5. View the summary statistics for the model, and write the log odds function generated by the model in a comment in your R script file: ln(odds(hitTarget)) = _____

Graduate Students Only

6. Using the log odds function from question 5 above, what are the predicted odds (not log odds) of hitting the target when hrsTreatment is 7.5? Answer this question in a comment in your R script file.
7. Using the log odds function from question 5 above, what is the predicted probability of hitting the target when hrsTreatment is 8.1? Answer this question in a comment in your R script file.