DS 64660 Generalized Linear Models Project Instructions

The semester project gives you the chance to apply the knowledge and skills you will learn in the course in a way that mimics the ways you can expect to use logistic regression modeling in a real-world setting.

For the project, you will choose **one data set** from the two listed below and perform a logistic regression analysis. Specifically, you will build a regression model and report on the model statistics and diagnostics. Your final deliverable for the project will be an R Markdown file. The file will contain the analyses detailed in each step with 5-7 written bullet points. Basically, pretend you are presenting to a non-technical audience, and the bullet points would serve as a script outline for your explanation. By non-technical I mean an audience who knows things like “a p-value less than 0.05 mean statistical significance” but cannot really explain the underlying concepts in great detail.

You can choose one of two different data sets to complete the project, either the wine quality data set first analyzed during week 3 or the data set diabetes from the `faraway` package, which we do not cover elsewhere in the course. Here are the details for each. Remember, you just need to choose one, not both.

- wine quality: rather than building a model to discriminate between white and red wine, for the project you may collapse the quality variable into a binary response equal to 1 (high) if quality > 5 and 0 (low) if quality <= 5. Build a logistic regression model to explain what factors are related to a high rating.
- diabetes: the help file for the data states that glycosolated hemoglobin (denoted as glyhbin the data) greater than 7.0 is usually taken as a positive diagnosis of diabetes. Thus, collapse glyhb into a binary response equal to 1 if gly > 7 (positive) and 0 if gly <= 7 (negative). Build a logistic regression model to explain what factors are related to a positive diagnosis.

Project Steps

The project consists of five steps, which you will work on throughout the second half of the semester. The steps you complete each week will accumulate in your R Markdown file, due at the end of the course. You will receive credit for completing drafts of Steps 1-5 according to the schedule below. Each of these steps are described in greater detail in the week they are due.

Step 1 Exploratory Data (Week 5)

- Task: Construct interleaved histograms and scatterplots to explore the relation between the predictors and response. Specifically, choose two predictors and

make an interleaved histogram and scatterplot for each. Thus, you should construct four total graphs.

- Deliverables:
    - Your code in Markdown should produce the plots.
    - Your 5-7 bullet points should explain the graphs including the axes and what each plot means. Remember, interleaved histograms may be easy for you but maybe not so for others.

Step 2 Variable Selection (Week 6)

- Task: Perform variable selection via the AIC using the step() function. Your starting model should include all available predictors. The reduced model should be used as your final model for all subsequent steps. Or, if you disagree with the recommended model, you need to indicate why.
- Deliverables:
    - Your code in Markdown should show how you used the function, produce the results, and indicate your final model
    - Your bullet points should give a brief explanation of the algorithm and comment on the variables retained/removed from the model. Are the results intuitive?

Step 3 Assess Model Fit (Week 7)

- Task: Check that continuous predictors have a linear relation with the logit using loess plots and perform the Hosmer-Lemeshow (HL) goodness of fit test. If the loess plot is nonlinear, then splines should be used to account for the nonlinearity.
- Deliverables:
    - Your code in Markdown should produce the loess plot(s) and the result of the HL test.
    - Your bullet points should explain the axes of the loess plot, the conclusion of the loess plot, the basic concepts of the HL test, and its result. If the HL test reveals poor fit, this should be discussed.

Step 4 Model Inferences (Week 8)

- Task: Report p-values and confidence intervals for significant predictors and check for influential observations. Any influential observations should be removed and the model should be refit. Note any changes in the inferences due to the removal.
- Deliverables:
    - Your code in Markdown should display the inferences and describe the influential observations

    - Your bullet points should give practical interpretation of the inferences and explain the process of checking for influential observations.

Step 5 Assess Predictive Power (Week 9)

- Task: Summarize predictive/discriminatory power of the model with an ROC curve and plots of predicted probabilities.
- Deliverables:
    - Your code in Markdown should produce the ROC curve and the plots.
    - Your bullet points should give a brief explanation of each plot and the interpretation for your model.

Step 6 Finalize Project (Week 10)

- Task: Add a brief Introduction and Conclusion
- Deliverables:
    - The Intro should give basic information on the data--how many observations, how many variables you are considering, and brief definitions/explanations of those variables.
    - The Conclusion should answer the following: Did you feel that you found a model that satisfactorily describes your y variable? Could the model have been improved with additional data and/or predictor variables that were not available for some reason at this time? Are there additional questions that you are curious about?

Final Project

You will upload your R Markdown file. This is due by Thursday, March 26 at 11:59pm.

Grading

The semester project counts for 25% of your course grade. The final project will be scored out of 100 points, with the following breakdown:

- 10 points each for steps 1-5 (50 points total)
- 10 points each for the Introduction and Conclusion
- 15 points for your code
- 15 points for overall writing quality and grammar

Weekly Steps

In addition to the grade, you receive for your final project, completing drafts of Steps 1-5 on time throughout the semester will count for 5% of your total grade for the course.

Only the final project and code you submit at the end of the course will be graded by the instructor. You will be able to revise all previously submitted steps up until that time. You will need to compile all the work you have done thus far into one final slide deck and upload it to Nexus by the due date in order to receive credit for completing the semester project.

FAQs:

*How do I decide whether to use splines or not?
If the loess plot looks nonlinear, then use splines. Otherwise, do not. During live session we may have compared AICs of models with and without splines or something similar, but you are not required to do this for the project.

*My final model summary output has “NAs” in it. Is this a problem?
Yes. Your final model (in any situation) should not have NA in the summary. In general, this usually occurs when one of your predictors is perfectly correlated with another predictor. In the context of this project, this may happen if you made a error specifying the knots for the splines)

*Does my model need to follow the Rule of 5?
Yes.