BMI 715 Final Project
INSTRUCTIONS

For the BMI 715 Final Project, you will conduct a regression analysis addressing a question of your choice, and using the regression methods, variable selection, evaluation metrics, and hypothesis tests that we have discussed.

Data: All students will use the curated subset of NHANES available on Canvas. Please see the accompanying guide linked from the Final Project assignment on Canvas.

Research question/hypothesis: Choose a dependent/outcome variable and ~10 independent/predictor variables from NHANES that you hypothesize may be relevant. Your research question and choice of predictors should be motivated by existing literature—they may be based on a scientific paper, a policy, or even the news.

It is also important that you choose (and state) a goal for your model: Are you trying to predict a value? Are you trying to assess whether a variable is associated with an outcome? Are you trying to find an interaction between variables? There are many possible goals for a regression analysis, but your goal should align with the methods you choose.

Methods: Your analysis should include the following components (outlined further in the rubric on the next page.

- Exploratory analysis of the variables
- Selecting a type of regression
- Variable selection or regularization
- Evaluation of model fit
- Follow-up analysis of results based on your goal

Important notes:

- There are many ways to do a regression. We will be paying attention to your justification for your decisions, so please clearly state those in the write up or RMD file. As long as the justification is reasonable, we will accept many different analytic choices. Without justification, we may not be able to award full credit.
- Where interpretations are required, the interpretation should be a full sentence that is specifically appropriate for the analysis or metric that you are interpreting. (E.g., generic statements like “The model fits well” or “The coefficient is significant” are insufficient.)
- Prioritize doing your analysis correctly over doing the most complex analysis possible. Choose an analysis appropriate for your research question and your personal comfort level with regression analysis.
- You may encounter obstacles that we didn’t cover in class as you’re conducting this analysis (e.g., functions that don’t automatically output p values). Use online resources such as StackOverflow to troubleshoot, and post on the Canvas Discussion board or ask the teaching staff if you have questions.

__Submission__: The final submission will include three files:
(1) a 1-1.5 page summary of your analysis, justifications for each step, and results (including reasonably sized figures), (2) R Markdown of all code used in the analysis (3) knitted PDF/HTML with code and figures from the R Markdown

All parts of the project should be submitted through Canvas. Properly cite any sources that you use. Submissions will be evaluated according to proper execution of the following project components (***all bullet points below will be evaluated***, unless noted as optional):

Criteria	**Points**
Statement of research question • Choose 1 dependent/outcome variable • Choose ~10 independent/predictor variables • Clearly state the goal of your model (e.g., assessing association between a predictor and an outcome, building a predictive model)	5
Motivation for this topic and these variables (e.g., scientific literature, news, policy) with citation	5
Exploratory analysis and QC of the data • Include 1+ plot of the relationship between variables • Check correlations/collinearity between predictors • Scale predictors that have a different range • Code categorical variables appropriately • (optional) Transform data if needed • (optional) Assess/remove/impute missing values • Justify decisions	10
Choose a regression model • E.g., linear, logistic, other generalized linear model • (optional) Decide whether to include interaction terms • Evaluate assumptions for chosen regression (see footnotes [1, 2, 3]) • Justify decisions	10
Conduct variable selection or regularization • Depending on your model's goal, this may be an iterative selection method (forward, backward, stepwise), or regularization (LASSO, Ridge, ElasticNet) • Justify decision	15
Evaluate model fit • Plot residuals (see footnote [4]) • Choose and calculate a metric • Appropriately interpret the metric that you calculate	15
Compare your model to an alternative model • You can choose any alternative model that you think would be interesting • Calculate a metric to compare the models • Appropriately interpret the metric that you calculate	10

Conduct one follow-up analysis • Choose an analysis that fits with the overall goal of your model • E.g., hypothesis testing for coefficient(s) (see footnote [5]) • E.g., accuracy of prediction in held-out data (see footnote [6]) • Appropriately interpret the results	15
Write-up 1-1.5 pages • Describe all data processing and analyses • Explain rationale for analysis decisions • Interpret metrics and tests • Summarize the conclusion of your regression and follow-up analysis in relation to your initial goal	10
All files submitted through Canvas • Comments in code to explain analysis steps (example: https://bookdown.org/ndphillips/YaRrr/a-brief-style-guide-commenting-and-spacing.html)	5

[1] This may not all happen at the beginning of your analysis, e.g., you may need to fit the model first to assess residuals
[2] For multiple regression, every variable should be assessed for linearity with the outcome — but if some variables don't meet that assumption, it's ok to keep them in the regression as long as you make a note of the violation.
[3] If you are using models other than a linear regression, the two assumptions to evaluate are (1) independence, and (2) linear relationship between linkfunction(y) and each x.
[4] For models fit with glmnet, you will have to manually calculate residuals from the real and fitted values (calculated with `predict` — see example in Lesson 11)
[5] Regularized regressions do not allow you to calculate a p value/CI, so you will not be able to choose this follow-up analysis with those methods.
[6] We didn't do this in class, but ask the teaching staff if you would like more guidance.

If you have any questions, post on the Canvas discussion forum or email Aparna Nathan at aparna_nathan@hms.harvard.edu with more individually relevant questions.