Homework 5

P8130 Fall 2025

Due: December 10, 2025 at 11:59pm

Guidelines for Submitting Homework

- Your homework must be submitted through Courseworks. No email submissions!
- Only one PDF file should be submitted, including all derivations, graphs, output, and interpretations. When handwriting is allowed (this will be specified), scan the derivations and merge ALL PDF files (http: //www.pdfmerge.com/).
- You are encouraged to use R for calculations, but you must show all mathematical formulas and derivations. Please include the important parts of your R code in the PDF file but also submit your full, commented code as a separate R/RMD file.
- To best follow these guidelines, we suggest using Word (built in equation editor), R Markdown, Latex, or embedding a screenshot or scanned picture to compile your work.

REMINDER: You are encouraged to collaborate on homework, explain things to each other, and test each other's knowledge. But **everyone must complete their own assignment and write their own solutions.**

Problem 1 (20 points)

Accurately predicting clinical outcomes is a central challenge in public health and medical research. For biomarkers like HbA1c—an indicator of long-term blood glucose regulation—building a reliable predictive model can support early identification of individuals at risk for diabetes or metabolic disease. Because health data often include multiple, correlated predictors, selecting an appropriate model is as important as the modeling technique itself.

The data used in this assignment comes from a community health screening initiative conducted through several primary-care clinics in the southeastern United States. As part of routine preventive visits, adults were assessed on demographic, lifestyle, and physiological measures, including age, BMI, diet, and physical activity. The data for this exercise are in the data file "HbA1c.csv".

a) Fit the full linear regression model to the data and examine standard diagnostics. Are there any suggestions of violation of regression assumptions?
b) Apply the LASSO regression procedure to the data using 10-fold cross-validation for choosing the tuning parameter. Fit LASSO models to the data with two different selections of lambda – one using the minimum of the CV function and one using the "1se". Write an expression for the fitted regression model in each case.
c) Use each model to predict HbAic in the data. Plot fits vs. residuals for each model. Based on the plots and the MSPE, which model would be preferred?
d) A new dataset was gathered – it is in the data file "HbA1c_val.csv". Use each model from (c) (fitted only to the original data) to predict HbA1c in the new data. Based on these out-of-sample predictions, which model would be preferred?
e) Considering prediction accuracy and the Principle of Parsimony, comment on the relative merits of the two fitted models.

Problem 2 (20 points)

Asthma is a common chronic respiratory condition and accurately understanding the factors that influence symptom severity is critical for guiding public health interventions and improving patient care. Environmental exposures, such as air pollution, as well as lifestyle and household factors, can affect both the level and variability of symptoms.

Adult participants were surveyed for demographic and lifestyle factors, including physical activity and exposure to household smoking, and nearby PM2.5 measurements were collected. The outcome variable represents a continuous measure of daily asthma symptom severity, with higher values indicating more severe symptoms. The data for this exercise are in the data file "HbA1c.csv".

a) Fit the full regression model to the data. For each of the predictor variables examine a plot
   - Of the residuals vs. the predictor.
   - Of the absolute values of the residuals vs. the predictor.

   Do any of these plots show apparent patterns of non-constant variance?
b) Regress the squared residuals from the model in (a) on PM2.5. Is there any apparent relationship between these two variables?
c) Using the results from (b), calculate appropriate weights and, using these weights, refit the model in (a). Compare the parameter estimates between the unweighted and weighted version. Which do you think provides a more accurate model for these data? Why?
d) If the weights calculated in (b) are appropriate, then a plot of *weighted* residuals from the model in (c) against PM2.5 should not show any pattern. Plot $\sqrt{w_i}e_i$ vs. PM2.5 and comment on the appearance of the plot.