Assignment 3
Transformations, Prediction, and Model Selection

Due: March 17 at 11:59 pm on Canvas

Concepts: transformations, interactions, test-train splits, model comparison, stepwise selection

Purpose: This assignment will guide you through the model selection process on a particular data set. We will continue with the kidney data from Assignment 2. We will go through the steps of a potential investigator analyzing the data is interested in developing a model to accurately predict GFR from a set of baseline characteristics in order to identify individuals at risk for kidney disease. This is an example of a prediction question of interest. This assignment will guide you through the model selection process and ask you to summarize the fit and performance of the final model for the investigator.

Task:

1. **Test-Train Split and Initial Predictor Selection** Based on our initial understanding of GFR, the predictors in the data, and the potential use of this model, the investigator decides to first drop the variable black from the data set. If you are interested in reading more about removing race variables from medical calculations that affect treatment decisions, we recommend the following two articles this first of which inspired this assignment.

   - https://www.nejm.org/doi/full/10.1056/NEJMoa2102953
   - https://www.nytimes.com/2020/06/17/health/many-medical-decision-tools-disadvantage-black-patients.html

   Next, we create a train and test set from the data using the code below, where kidney_df is the name of data frame containing the full data.

   ```
   set.seed(1)
   kidney_df$id <- 1:nrow(kidney_df)
   kidney_train <- kidney_df %>% dplyr::sample_frac(0.75)
   kidney_test  <- dplyr::anti_join(kidney_df, kidney_train, by = ’id’)
   ```

2. **Initial Transformations** The first step in the model selection process that the investigator considers is transformations to the continuous variables. To replicate this step, you should use a log transformation of GFR and use all available variables in the model (except for black). First, plot log(GFR) vs baseu, bascre, log(baseu), and log(bascre) (for 4 plots total). Comment on what you observe.

   Then, determine whether you should log transform either bascre and/or baseu. To do so, fit each of the four models below and report the Adjusted R Squared and AIC for each model. Explain why the investigator would use these two measures to compare these models, comment on the results, and explain why model 4 would be chosen using these criteria.

$$
\begin{aligned}
\text{mod}_1 &\leftarrow \text{log(gfr)} \sim \text{baseu} + \text{bascre} + \text{sbase} + \text{dbase} + \text{age} + \text{male}\\
\text{mod}_1 &\leftarrow \text{log(gfr)} \sim \text{log(baseu)} + \text{bascre} + \text{sbase} + \text{dbase} + \text{age} + \text{male}\\
\text{mod}_3 &\leftarrow \text{log(gfr)} \sim \text{baseu} + \text{log(bascre)} + \text{sbase} + \text{dbase} + \text{age} + \text{male}\\
\text{mod}_4 &\leftarrow \text{log(gfr)} \sim \text{log(baseu)} + \text{log(bascre)} + \text{sbase} + \text{dbase} + \text{age} + \text{male}
\end{aligned}
$$

3. **Variable Selection** Next, the investigator considers variable selection. The investigator decides to drop both sbase and dbase from the model by finding the corresponding AIC value when dropping each variable. Additionally, the investigator uses a nested hypothesis to test this decision. Explain what the null hypothesis and alternative hypothesis is for this test and comment on the results.

   Then, interpret the coefficients for this model using the discussion in class on how to interpret coefficients when a log transformation is used for the outcome and/or predictors.

4. **Adding a Polynomial Transformation** Finally, the investigator considers a polynomial transformation for serum creatinine. First, plot the residuals for your model vs log serum creatinine and comment on what you observe. Then, consider a polynomial transformation for log serum creatinine. Be sure to explain how you are choosing the degree for your polynomial transformation.

5. **Evaluating the Final Model** Evaluate your final model using MAE and RMSE on both the training set and the withheld test set and compare the results, relating what you observe to the bias-variance trade-off discussed in class. Report on the diagnostic plots for your resulting model (you may limit yourself to the four default plots in R in this discussion). Last, relate these results to the motivating question — overall how useful do you think your model would be for predicting GFR in practice?

Criteria: Your report will be graded on the following criteria.

- Transformations (10 points)
- Model Comparisons (10 points)
- Variable Selection (10 points)
- Final Model Evaluation (10 points)
- Report and Exposition (10 points)

See the attached rubric on Canvas for more details.