Dylan Watson

Professor Hasenstab

STAT 200: Introduction to Data Science with R

3 May 2024

Predicting Future Lung Forced Expiratory Volume with Methods of Statistical Modeling

In this study we were provided with a .csv file containing 35 variables of health data for over 5747 patients. These variables include information such as age, gender, race, height, weight, systolic and diastolic blood pressure, lung capacity, expiratory volume, and many other lung related factors. However, 1747 patients did not have data on a variable name FEV1\_phase2 which is expiratory volume at 5 years follow up. It is our goal in this study to use statistical modeling to create accurate predictions for this expiratory volume at follow up given the other patient health metrics.

We began by exploring the distributions of certain numeric variables of our dataset such as our response variable of interest FEV1\_phase2 among others. We concluded that FEV1\_phase2 follows an approximately normal distribution with a mean of 2.12 liters volume and standard deviation of 0.826 liters.

Proceeding, we decided to explore the relationships between FEV1\_phase2 and other categorical variables. We found that forced expiratory volume at follow up was strongly correlated with smoking status and gender and not significantly correlated by race.

Next, we looked into correlations between FEV1\_phase1 and other numeric variables. These included FEV1, FVC, and height. There was found to be strong positive linear correlations between FEV1\_phase2 and FEV1 as well as FVC with R values of 0.8887 and 0.7079 respectively. Many other variables were found to have some correlations but not nearly as significant as those two.

The next step came the phase of modeling. We decided to begin with simple linear regressions. First, we modeled FEV1\_phase2 vs FVC and received an R^2 of 0.7897. This model was a strong start for its simplicity, but we decided to add in some other predictors to try to improve our model’s accuracy. With the addition of FVC and height\_cm in a multiple linear regression model we increased the R^2 to 0.8129. This second model produced a RMSE of 0.357.

For the final stage of modeling, we went with using the step() function which considers the Akaike Information Criterion when adding variables at each step. This function will start with a base regression and test adding each variable to the base to identify which variable improves the accuracy of the regression most without introducing too much complexity. The model will step forward through this process until no variable is left which results in a lower AIC score. In other words, until the complexity added outweighs the improvement in accuracy. In the end of this process, we got a regression which includes the following variables: FEV1, FEV1\_FVC\_ratio, FVC, height\_cm, copd, pct\_gastrapping, exp\_meanatt, smoking\_status, visit\_age, total\_lung\_capacity, gender, hr, race, CigPerDaySmokAvg, and pneumonia. This final model has an R^2 of 0.8753 and RMSE of 0.292 which are solid improvements on the previous model.

Finally, we can use the predict() function to use the model we trained on the 4000 patients with FEV1\_phase2 data to predict values for the remaining 1747 patients.