Project 4 – Regression

*Joseph Park - x172917 Date: 4/10/20*

Objective

The assignment is an illustration of numeric prediction based on the multilinear regression (and then nonlinear) model using insurance data from Packt Publishing to predict expenses (labeled as “charges”).

Random forest model

Step 1: Download the data set (insurance.csv) from Github,

Step 2: Look for missing data.

Step 3: Look at summary, size and shape.

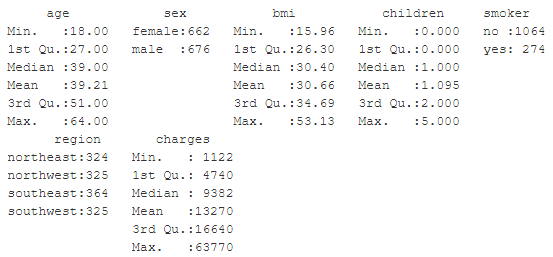
Step 4: Look at pairwise scatter plots.

Step 5: Fit the data to a multilinear model with all but charges as the predictor features in a linear form.

Step 6: Evaluate performance with metrics given by the lm function.

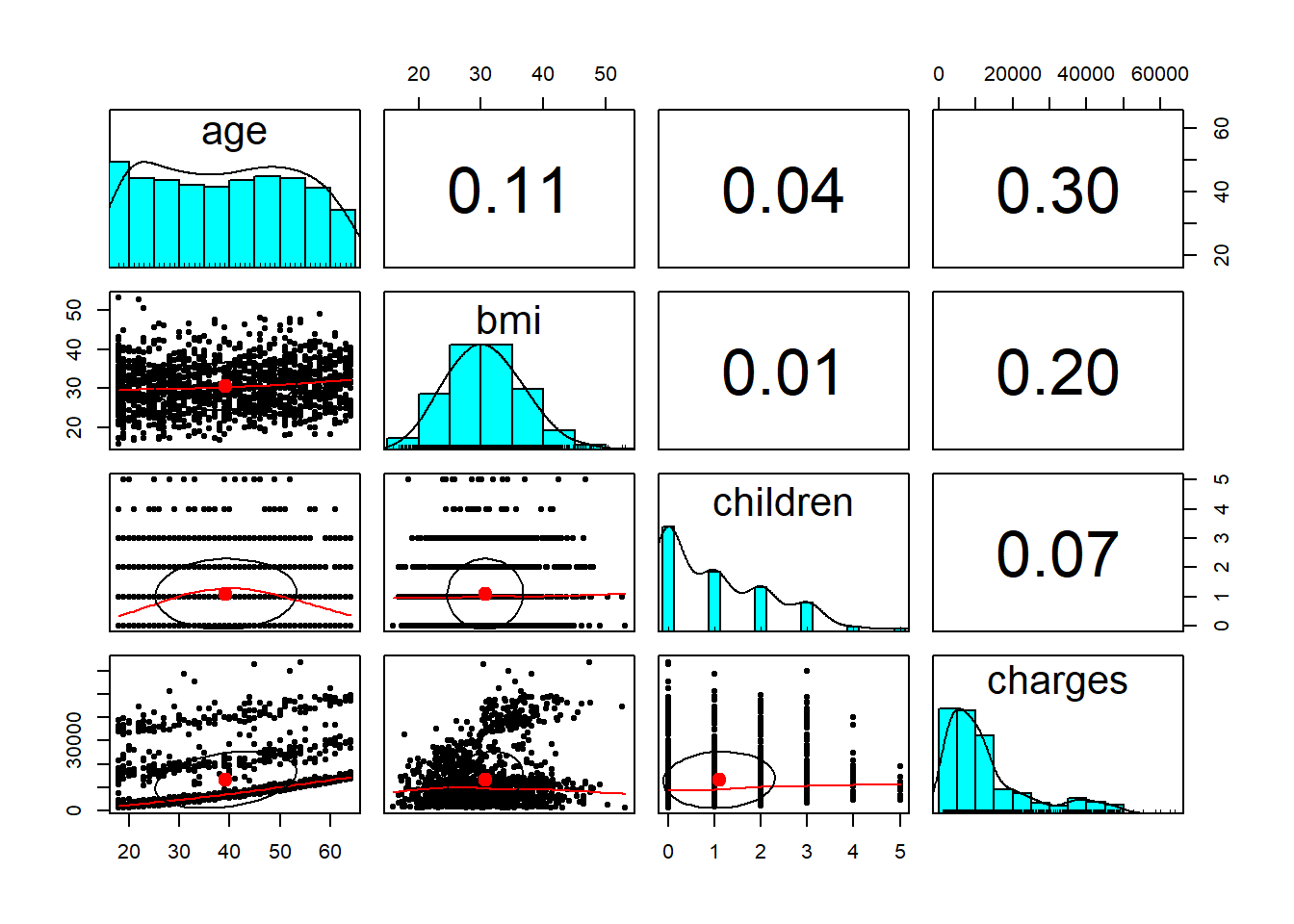
Step 7: Improve performance by adding nonlinear and interaction terms to the linear form we used above.

Step 8: Evaluate performance and compare with the prior results



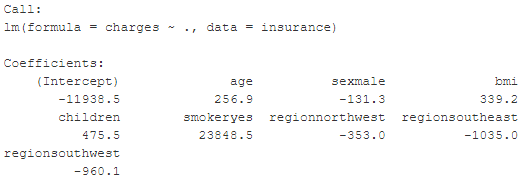
Here are the standard deviations of the numeric variables in order of above left to right then top down.



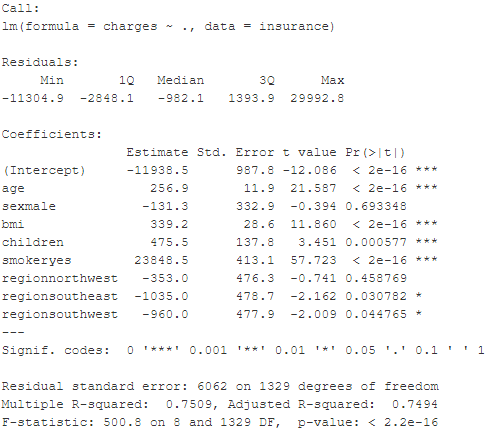
Here are the corresponding histograms (along the diagonal) and pairwise scatterplots. 

Results

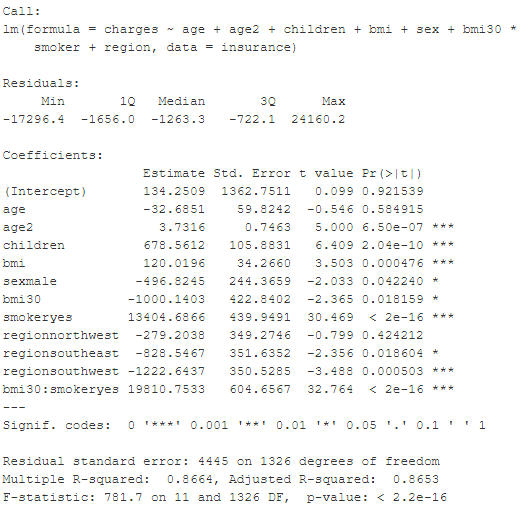
Linear model:



Linear model’s performance:



Improved (with nonlinearity and interaction terms) performance:



Interpretation of the Results

Note 1: Even just the (multi)linear model is fit with good R-squared (about 0.75).

Note 2: Since age is a factor that would naturally give increase in rate of charges with constant increase of age, it makes sense to add at least one more polynomial term of increasing order (as in Taylor expansion) to improve performance.

Note 3: Since we would expect smoking and obesity to amplify one another with respect to insurance charges, we add an interaction term between the two to improve performance.

Note 4: The increase in R-squared from about 0.75 to about 0.865 demonstrates that the common sense guided addition of terms was not misguided.

Note 5: We can keep adding terms, but we do not want to overfit. So, to do so, we should look for good reason to as we did above.