Abstract:

Biologists have long wondered what degree a parent’s features can influence their child’s characteristics. This curiosity sparked those such as Gregor Mendel, who studied the genetics of a pea to see how characteristics are passed down from generation to generation. These observations, as well as those presented by Charles Darwin, allow the chance for environmentally pressured evolution of species, has made become the basis of modern biology. In this study, researchers hope to study the ability for modern statistical techniques to predict a child’s height based off of their parents. In an effort to determine which parent has the most effect on predicting the child’s height, various transformations of the data describe different features, such as which parent is taller, or the difference in height of the parent’s. With these extra variables in hand, researchers have the ability to reduce the size of the model by hand or using algorithms designed to remove the influences of the researcher from the process. Both of these methods are employed to determine the extent of each parent’s influence on their child. Some of these models are heavily mathematically rigorous, and others utilize the multiprocessing features of our modern CPU’s to create a large number of models, and then pick the best of all of the possible combinations. This allows the researcher to be removed from the process of variable selection in attempts to reduce bias. Each of these different models is tested, then presented to the reader.

Introduction:

In efforts to describe this data, much credit is due to the providers of the software which allowed this research to take place, SAS was used heavily to study the statistical aspects of the models, including the P-Values, residuals, and made fitting a model very straightforward. Python, in conjunction with Pandas and Numpy, helped to clean data and transform it to attain different features to used in the models. Using different matrix decomposition algorithms was made very easy by SciKit Learn, as well as fitting Linear Models. Using Pipelines in this way allows for quick access to a wide variety of models very quickly. When going through the references and code, readers should note that each technique is very similar to the next, and the only difference is calling a different object to do a different job. It is in this same fashion this software can be used to utilize more complex tools, as is the heart of object oriented programming.

Analysis:

The first model to be examined is the full model. This a standard linear regression model applied to all of the supplied dependent variables. These include mother and fathers height and race, child’s nutrition and gender, which parent is taller, and the difference in heights between the parents. Researchers find an R Square value of .61. From here the researchers investigate whether or not the model breaks the assumptions necessary for an unbiased linear model. It is discovered observation #222 has a very large Cooke's Distance, and must be removed from the data set. All further analysis is performed on the data which does not include this point. This immediately results in a .03 increase in both our R-Sqquare value and the Adjusted R-Square value in all models which were studied.

Next researchers attempt to decipher whether or no the model is colinear. Colinearity is when we have a singular matrix, which implies that one of our dependent variables is linearly dependent on another. This is a common concept in linear algebra, that describes a n-vectors, where at least 2 vectors can be described as a linear combination of the remaining vectors. To study this effect, a metric is utilized called the Variance Inflation Factor (hereby referred to as VIF), which is taken for each dependent variable being studied. It is quickly uncovered that father's height, father's race, mother's race and distance between the height of the parents all share a VIF of nearly 10, which is imples these factors are linearly dependent on one another. We remove the factor with the highest P-Value, mother's race. We observe the newly reduced model and find that we still have a linear dependency between distance, father's height and mother's height, though it is not nearly as strong as before. We remove distance as it has the highest P-Value, and observe that each variable is now less than 2 in regards to it's VIF.

Feature selection was also done with a variety of non-empircal algorithms including PCA, backwards selection, stepwise selection factor analysis, and random forest classification t provided by SciKitLearn and SAS. It should be noted that lower R Square values are seen across the board, ranging from roughly 0.5-0.55 for models reduces using SciKitLearn.. The explained variance ratio is studied when decomposition techniques are utilized, to show which variable have a higher likelihood of being reduced. The larger the ratio, the higher chance they have of being used in the reduced model. To remove bias from these selection techniques, we use a confidence level of 0.1. For the PCA techniques, this eliminates every variable but father’s height and father’s race. We see Factorial Analysis has an R Square value of -0.087, so we ignore this model, though it is included in the source code. This concludes the use of decomposition in efforts to reduce the dimension of the data set.

Random Forest algorithms are different as the benefit form the use of graphs, a very heavily studied branch of Mathematics, heavily utilized by Computer Science, Finance, and Physics. In this computational approach, many different children of a decision are birthed with mutations, and these mutations are studied to see which performs the best, then other branches are made off of the better nodes in the tree. This is a Random Tree approach, but we are using a random forest approach, and just like a forest, this algorithm uses many trees to find the best fit. In this particular instance, the classifier chooses the mother and father’s height as well as the father’s race. In the polynomial version, it uses one feature, The father’s height times the fathers race, which is mapped to an integer value, though this model has an R Square value of 0.03 and should not be considered. The Random Forest classifier for the full linear model produces a R Square value of 0.51 when only using the first three features, father's height, father's race, and mother's height.

SAS is used to study stepwise selection, R-Squared Selection, and CP. The first three actually produce exactly the same results, reducing the model to mother and father's heights, child's gender, and which parent is taller. They achieve the same R-Square Value of 0.641, when the full model achieves an R-Square value of 0.649. The R-Square selection model and CP model choose the same variables, which include all features except for father's race. As a result the achieve the same R-Square, at 0.644. This unfortunately does not reduce our model enough for effective results in efficiency or accuracy.

Conclusion:

Many of the models used to explore the data set result in very similar accuracies. This is great for the purpose of the researchers as it implies that even when reducing the model to use less dependent variables, which gives faster results, we maintain a very similar accuracy as before. This has shown that mother and father's height are both very important for predicting their child's future height, as the always have comparatively low P-Values, even in the full model we find that both scores are below 0.025, mother's height going as low as .008. Child's gender is also used in each model , as well as which parent is taller. That being said, which parent is taller often has a high P-Value in the reduced models, and if researchers were to reduce the model even further, this feature would always be the first one to go. It seems studying the outliers of the model provides the biggest increase in accuracy, though this could be described by over fitting the model to the supplied data set. A way to avoid this problem would be to either collect more data, find more factors to study, or to use k-fold cross validation.

References:

<https://scikit-learn.org/stable/modules/classes.html>

Data:

https://github.com/g8r-b8/regressionLab