Connor Brown

Professor Adam Kapelner, Ph.D.

Math 390.4

4/29/2019

The Human Body is Not Understood as Height Cannot be Predicted

The medical establishment does not understand the human body. Contrary to popular belief, physicians do not hold every answer about the state of your health, how lifestyle choices affect health, how genes affect health, risk levels of developing certain diseases, or stages of life when you're most likely to develop certain diseases and conditions. In fact, physicians cannot even provide an accurate answer to one of the most common questions asked by children… "how tall am I going to be?"

Some traits, like sickle-cell anemia, are determined entirely by a single gene, and are called, "monogenic" (Driss, 2009). Other traits, like eye color, are determined by interactions between multiple genes, and are called, "polygenic" (Pospiech, 2011). Height is a quantitative trait that is believed to be determined by multiple genes as well as interactions with the environment and is called "multifactorial".

Human height has long been a fascination of scientists. Although the heredity of human height has long been acknowledged, only recently have scientists been able to identify individual genes that influence adult height, or specific mutations that result in abnormal stature. It is believed approximately 80% of height is determined by genetics. The remaining 20% is

determined by environmental factors, which may include nutrition, disease, exercise, and pollution (Visscher, 2006).

"Models", in science, are representations that are approximations to reality, absolute truth, systems, or phenomena. A mathematical model aims to describe the dynamics of these phenomena utilizing mathematics, more specifically, the relationship between a set of variables, or features, and a specified output(s), by an equation or set of equations (Quarteroni, 2009).

"Essentially, all models are wrong, but some are useful" (Box, 1987), is a famous quote by British statistician George Box. Box's quote can be interpreted as meaning every model is wrong because it is a simplification of reality that ignores some feature of reality. Models are limited, as it is almost impossible for a single model to perfectly describe a real-world phenomenon. The prediction of monogenic traits is an exception to this rule, as only one feature (gene variant) is necessary for perfect predictive accuracy. The reality of height, and other multifactorial traits, is infinitely complex, and a model for prediction of these traits is just an approximation of their reality. Nonetheless, a model does not need to be correct to be useful. If the approximations of a model are similar to reality, the model can be used to make predictions, or better understand how the universe works.

One of the biggest assumptions in developing mathematical models is stationarity, which implies that the relationship between phenomenon and causal inputs does not change, i.e. the relationship is constant (ex: gravity does not change over time). It also implies that the mechanism used to measure phenomenon and features does not change. If data is used to model a relationship, and that relationship changes, the model is useless because it models a relationship that no longer exists.

Models can be developed using an approach called, “learning from data”, which is an empirical (observation-based) approach. The specific type of learning from data we employ here is called, “supervised learning”, which infers a function that maps an input to an output based on multiple observations, or input-output pairs. Supervised learning requires “training data”, denoted D, which contains these input-output pairs, a set of candidate functions that can fit the data, denoted *H*, and an algorithm, *A*, that takes D and *H* and provides the best *approximate* of the function for the data.

Because height is determined by multiple gene variants, it is difficult to accurately predict how tall a child will be. The inheritance of these variants from one’s parents helps explain why children usually grow to be close in height to their parents. However, because of homologous recombination and independent assortment in meiosis, each gamete (sperm, egg) contains a different set of DNA (Riley, 1965). This produces a unique combination of genes in the resulting zygote, which explains why siblings can be of different heights, even if they are the same sex.

There exists a couple of common, simple models currently used by physicians to predict adult height. One is used when the child is an infant: double a boy's height at age 2 or a girl's height at age 18 months; this model is believed to be less accurate (MayoClinic.org). The most commonly used model is called, “mid-parental height”, and is calculated by adding the mother's height and the father's height in either inches or centimeters, adding 5 inches (13 centimeters) for boys or subtracting 5 inches (13 centimeters) for girls, and dividing by two (MayoClinic.org). Studies have shown that 90% of children’s heights will be within 1.5 SD’s (standard deviations) of their mid-parental height (Wright, 1999).

The purpose of this paper is to formulate a more complex mathematical model to predict adult height.

In this model, the phenomenon measured, human final adult height (height), is a quantitative, continuous, trait, that is expressed in units of centimeters (cm), and is denoted as y. Height is measured in physicians' offices at age 25 years old, as growth is assumed complete for most humans. This is one potential flaw in the measurement of the phenomenon, as we are assuming final adult height has been reached by this age.

Reality can be written as $y = t(z_1, \ldots, z_t)$, where t represents the function that mother nature uses to create the phenomenon, and $z_1, \ldots, z_t$ represents all of the causal inputs that contribute to the phenomenom. As reality is infinitely complex, we are ignorant of all of the features that contribute to height (z's), as well as the relationship between them (t).

The next best thing to do is obtain features that approximate the information in the z's (features believed to contribute to height), and combine them together to model y; these features are denoted as $x_1, \ldots, x_p$. Their raw dimensionality, or, the number of features used that are believed to contribute to height, is denoted p.

To build this predictive height model, data must be collected from observations (people) that contain information for $x_1, \ldots, x_p$. This data is collected immediately following height measurement in physicians' offices. The information for the features for each person is contained in a vector $\mathbf{x}_i := [x_{i1}, \ldots, x_{ip}]$. The collection of all of the people's feature information is $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and are contained in a matrix, X. The heights of all the people are denoted $y_1, \ldots, y_n$ and are contained in a single-column matrix, Y. The dataset, D, consists of X and Y together. The number of people included in D is denoted n.

Feature selection is a critical component of developing mathematical models. Oftentimes, strong domain knowledge is necessary to have an understanding of which features are worth pursuing measurement. After a set of features have been selected utilizing domain knowledge, forward stepwise regression can be used to trim the list of features down to prevent overfitting of the model. An overfit model is an overly complex model that has very low in-sample RMSE (IS-RMSE), fits the data too closely, increases estimation error, and results in future predictions suffering as a result, characterized by high out-of-sample RMSE (OOS-RMSE). Overfitting can be caused by an overly complicated model, like a model with high-degree polynomials, too little data (small n), or too many features (large p).

Although the heredity of height has been long-acknowledged, extensive research has explored the contribution of environmental factors.

A 2016 publication estimated mean height for people born between 1896-1996 in 200 different countries. The researchers found that height increase in some countries far outpaced height increase in others, which supports the notion that height contains strong environmental influence, as genes would not be expected to result in significant changes over such a short period of time (NCD Risk Factor Collaboration, 2016).

It is believed that an increase in protein quality and consumption is a strong contributor to increases in height in many of these countries. The consumption of rice is highest in Asia and is accompanied by low total protein and energy intake, and also has one of the shortest statures in the world (~162–168 cm). The highest animal protein consumption is in Northern and Central Europe, with the tallest men in the world in the Netherlands at 184 cm (Grasgruber, 2016).

Within the U.S., differences in average height are present between gender and races. Black and White females stand comparably, at 64.2 and 64.1 inches, respectively, while Asian

and Hispanic females stand similarly, at 61.8 and 61.9 inches, respectively. The propensity for Blacks and Whites, as well as Asians and Hispanics, to share similar average heights, is also seen in men. Black and White men stand 69.3 and 69.6 inches, respectively, while Asian and Hispanic men stand 67.0 and 67.3 inches (CDC, 2017). Some studies suggest the actual heredity of height can differ depending on ethnicity.

A study on mono- and di-zygotic twins suggested heritability ranges from 73 to 86% in American/European Caucasian female adolescents, and 68 to 86% in East Asian female adolescents. Similar results were obtained from male data: 75 to 81% in Caucasians, 68 to 79% in East Asians (Hur, 2008). Another study suggested that height heritability is as low as 65% for West African populations (Roberts, 1978).

Infection during childhood is also believed to affect adult height. Monozygotic twin studies suggest that increased exposure to infection throughout childhood can result in shorter adult height. In one study, the twin with the higher frequency of childhood infection was twice as likely to be the shorter twin, irrespective of gender (Hwang, 2013).

Specific genes have been identified as being associated with height. In 2017, GIANT (International Genetic Investigation of Anthropometric Traits Consortium) identified close to 700 gene variants believed to affect height, most of which have a small effect, typically a millimeter or less. They reported on 83 height-associated coding variants, some with effects of up to 2cm per allele, including genes *IHH*, *STC2*, *AR* and *CRISPLD2*. Unfortunately, these genes were identified through genome-wide *association* studies, so how they act biochemically is unknown (Marouli, 2017).

From this research, race, sex, protein intake, childhood infection, and gene variants are identified as features of interest, in addition to, of course, maternal and paternal height. Immediately, it is clear the majority of features are categorical (race, sex, and gene variants), while maternal and paternal height are quantitative. The modeler must decide how to measure protein intake and childhood infection.

In the research describing protein intake and height increases in different countries, it is inferred that increased animal protein intake results in increased height. How this feature is measured must be specified. Reliable data on a person's lifetime animal protein intake is practically infeasible, unless they keep a log of everything they eat from birth to adulthood. Instead, self-reported data must be used, which is notoriously less reliable (Smith, 2018). Patients are asked to complete a survey where they estimate how often (and in what quantities) they ate common animal-protein-containing foods like chicken, fish, red meat, eggs, milk, and cheese during distinct age brackets (0-3, 4-7, 8-11, 12-15, 16-19 years old). A mean monthly animal protein intake in units of grams (g) over the course of their lifetime is calculated to use as a feature in the model.

Childhood infection must also be quantified. Patients are asked to review medical records from their pediatrician, and recall sicknesses during childhood, to self-report the total number of days they were too ill to attend school/extracurricular activities. It is obvious that this method of measurement is susceptible to inaccuracies. In addition, this method may not even capture the true contribution of illness to height, as the *severity* of the illness possibly holds the true affect. A severe illness like leukemia or cancer may have a significantly greater effect on height than a common cold or the flu.

The measurement of animal protein intake and childhood infection is an example of infidelity in the measurement process. Because this data is both self-reported and estimated, there is a high probability of inaccuracy in the measurements. The measurement of height is also susceptible to infidelity. It is possible the height scale used in some physicians' offices are old and worn down and could result in inaccurate measurements. Human error is also possible, as some nurses or physicians may not correct a patients' posture when standing, or they may simply not place the scale at a 90 degree angle, or mis-read the height. The time of day height is measured can also affect the measurement, as humans are known to be taller in the morning and get shorter as the day progresses; the difference between morning height and night height can be up to 2.7 cm (Vuvor, 2017). The measurement of parental height is collected from parents medical records at age 25 years old, and potential infidelity in these measurements is equivalent to those described above.

Race and sex are both self-reported. For the purposes of this study, only biological sex (not identity) will be utilized. Some people are mixed-race, so they may report multiple races, which is addressed when dummifying variables.

Gene variant data must be collected by having the patients' genome sequenced by a whole genome sequencer, followed by filtering out of the gene variants of interest; if the gene variants of interest are found, they are mapped to a 1 in the dataframe. If the gene variants are not found, the feature is mapped to a 0. The error rate of whole genome sequencing is approximately 0.0001% (Schmutz, 2004). In a genome of 3 billion base pairs, this equates to approximately 300,000 base pair errors in any single sequencing. A single base pair error can potentially alter the measurement of the gene variant, although this is really case-dependent. This risk of measurement error is mitigated by sequencing each person's genome 10 times. If base

pair error is treated as a binomial random variable, the probability of there being more than one base pair error for any given base pair after 10 sequencings can be represented as $\sum_{k=1}^{10} \binom{10}{k}(0.000001)^k(1 - 0.000001)^{10-k}$, which is computed in RStudio as pbinom(1, size = 10, prob = 10E-6, lower.tail = FALSE), and is equal to 4.49976e-09 (near-zero probability).

Because the dependent variable in this model, height, is continuous, and the independent variables are both continuous and categorical, an Ordinary Least Squares (OLS) Algorithm called, Analysis of Coviance (ANCOVA) can be utilized. The initial data frame is visualized in Figure 1, below, where p = 706 (700 of the 'height-associated' gene variants identified in the GIANT study, plus 6 other features). Each feature can be represented by a vector, $\mathbf{x}_{.i} := x_{1i}, \ldots, x_{ni}$, and the collection of all feature vectors is denoted $\mathbf{x}_{.1}, \ldots, \mathbf{x}_{.p}$

Figure 1: The initial design matrix, X, containing p_{raw} = 706 features, alongside the **y** vector, containing the heights for all observations.

y

Height
y_1
. . .
y_n

X

Mom Height	Dad Height	Protein Intake (g/month)	Childhood Infection (total #)	Race	Sex	Gene Variant 1	…	Gene Variant 700
x_{11}	.	.	.	.	.	.	.	x_{1p}
. . .								. . .
x_{n1}	.	.	.	.	.	.	.	x_{np}

Regression on categorical variables that contain text cannot be run, so the categorical features, race and sex, are dummified. There are only 2 levels for sex (male, female), so female is mapped to 1, male to 0, and p is still equal to 706. Race is limited to 6 levels: White, Black,

Latino, Asian, American Indian, and Pacific Islander. After dummification, there are 5 race columns, and White used in the reference level with female. All categorical variables have been quantified so $p = 710$.

Now, $p = 710$ with the data frame. A high number of features will result in overfitting of the model without a significantly multiple number of observations. As mentioned in the GIANT study on gene variants that influence height, most of the 700 variants reported influence height by less than a millimeter, while a few influenced height by up to 2 cm. Feature selection can be trimmed down by using a machine learning method called forward stepwise regression. If interested in interactions between features, a linear model can be built that includes multiple interactions, which will significantly increase the number of features in the matrix, and should be considered if run-time is important to the modeler. Forward stepwise regression starts by building all possible single-variable models by regressing y on every x individually, and selecting the feature that produced a model with the lowest out-of-sample standard error (OOS-SE). Then, another model is built utilizing the previous feature chosen with every second feature. Then, the second feature that produced a model with the lowest OOS-SE is chosen. This iterates repeatedly until OOS-SE bottoms out and starts to increase, and the final p_{raw} features for the model can be selected by identifying the OOS-SE minima. Selecting the models with the lowest OOS-SE (within stepwise regression) involves a process called validation.

Validation is the process of checking model performance on data. To validate a model, the data, D, must be split in two subsets, D_{train} and D_{test}, where D_{test} commonly contains 10% of the total data. D_{train} is used to build the model, and D_{test} is used to test the model. Validation can be separated into two categories, in-sample and out-of-sample validation (IS-Validation, OOS-Validation). IS-Validation uses the model to predict y-values for observations present in the

training set. IS-Validation provides metrics for model evaluation called in-sample R^2 and RMSE (IS-R^2, IS-RMSE). Using a model to predict for observations that were used to create the model does not provide an honest metric of the model's ability to predict for future observations. Therefore, OOS-Validation is conducted, using the model to predict on observations *not* in the training set. The OOS-R^2 and OOS-RMSE obtained are the honest metrics of model performance and are therefore used in model selection.

Now that final raw features have been selected for the model, different model complexities can be explored in a process called model selection. Phenomenon can be written as

$$y = g(\mathbf{x}) + (h^*(\mathbf{x})\text{-}g(\mathbf{x})) + (f(\mathbf{x})\text{-}h^*(\mathbf{x})) + (t(\mathbf{z})\text{-}f(\mathbf{x}))$$

where g is the model produced via the learning process, h^* is the best function within the model candidate set, f is the best function for the data (regardless of the candidate set), and t is reality. $(h^*(\mathbf{x})\text{-}g(\mathbf{x}))$ is called, "estimation error", and is minimized with increased data (higher n), $(t(\mathbf{z})\text{-}f(\mathbf{x}))$ is called, "error due to ignorance", is denoted as δ, and can be reduced with the measurement of more relevant variables. $(f(\mathbf{x})\text{-}h^*(\mathbf{x}))$ is called, "mis-specification error", and can be reduced by utilizing a more flexible model candidate set, which can include polynomial terms, interactions, variable transformations, etc. The sum of mis-specification error and error due to ignorance is denoted ε, which are the differences between the observed and predicted output values from h^*. The sum of all three errors is denoted e, and is known as the "residuals", which are the differences between the observed and predicted output values from g.

"Model Selection" is a procedure for picking a model to use out of a set of models. Machine Learning in R (MLR) can explore different learning algorithms. "Model Averaging" is a reasonable strategy for selection. In this procedure, D is split into three subsets, D_{train}, D_{select},

and D_{test}. D_{select} is just testing data for D_{train}. First, the design matrices for the different models must be made. Then, the matrices are fit to the training data to get models, predicted on the selection set, and the model with the lowest OOS-SE, the "best" model, is selected. You can fit many, many models but should be careful not to optimize to the selection set. If the selection set has a lot of data, this should not be an issue. The "best model" is then predicted on the D_{test} to get OOS-SE. If the SE on the test set is similar to the selection estimate, overfitting to the selection set did not occur. Finally, build the model on *all* of the data and ship it.

MLR can be used to test multiple different models on the training set. The inner loop will cross-validate every single model on the training and selection set. Cross-validation is done to reduce variation in the SE. The outer loop re-samples the tuning-wrapper to prepare for the next iteration of the inner loop. The model chosen can potentially be different after every iteration of the outer loop. It then aggregates all of the OOS predicted *y*-values together, and computes the final SE, which provides the best guess of how the model *selection algorithm* will perform, not the model itself. Then, the final model is built on the entire dataset, but SE for the final model is expected to be similar to the SE from the selection algorithm.

Now that the final model has been made, it can be used for real-world prediction. It is important to understand that extrapolation when using predictive models is very dangerous and can result in extremely poor predictions, especially with highly complex models where the 'weights' of each feature are not understood. Extrapolation is when new input data for prediction is outside the range of the input data used to create the model. Interpolation is much more reliable, and ideally, the values for all of the *x*'s will be within the ranges of the corresponding data used to make the model.

Height is an extremely complex trait, which is influenced by hundreds of different genes, in addition to a number of environmental factors. The genetics of human height are just starting to be greater understood. However, understanding of the complexity of human height will likely take many more years, as we have almost no understanding at all of how height-associated genes interact with one another biochemically to produce the affect they have. A useful mathematical model for the prediction of height cannot be made with our current understanding of the causal inputs. A model with $R^2 > 0.98$ would constitute "understanding" of human height. The R^2 for a model with this measured data would likely be greater than 0.7 but less than 0.9, as the heredity of height is believed to account for ~80% of variation, and the 700 gene variants identified in the GIANT study are believed to be the primary contributors to final adult height. Nonetheless, because height cannot be predicted with low error, the medical establishment does not understand the human body.

Bibliography

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Chan, Y., Salem, R. M., Hsu, Y. H. H., McMahon, G., Pers, T. H., Vedantam, S., ... & Smith, G. D. (2015). Genome-wide analysis of body proportion classifies height-associated variants by mechanism of action and implicates genes important for skeletal development. *The American Journal of Human Genetics*, *96*(5), 695-708.

Driss, A., Asare, K. O., Hibbert, J. M., Gee, B. E., Adamkiewicz, T. V., & Stiles, J. K. (2009). Sickle cell disease in the post genomic era: a monogenic disease with a polygenic phenotype. *Genomics insights*, *2*, GEI-S2626.

Grasgruber, P., Sebera, M., Hrazdíra, E., Cacek, J., & Kalina, T. (2016). Major correlates of male height: A study of 105 countries. *Economics & Human Biology*, *21*, 172-195.

Jelenkovic, A., Hur, Y. M., Sund, R., Yokoyama, Y., Siribaddana, S. H., Hotopf, M., ... & Pang, Z. (2016). Genetic and environmental influences on adult human height across birth cohorts from 1886 to 1994. *Elife*, *5*, e20320.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., ... & Rüeger, S. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, *542*(7640), 186.

NCD Risk Factor Collaboration. (2016). A century of trends in adult human height. *Elife*, *5*, e13410.

Pośpiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A., & Branicki, W. (2011). Gene–gene interactions contribute to eye colour variation in humans. *Journal of human genetics*, *56*(6), 447.

Quarteroni, A. (2009). Mathematical models in science and engineering. *Notices of the AMS*, *56*(1), 10-19.

Riley, R., & Law, C. N. (1965). Genetic variation in chromosome pairing. In *Advances in genetics* (Vol. 13, pp. 57-114). Academic Press.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., ... & Escobar, J. (2004). Quality assessment of the human genome sequence. *Nature*, *429*(6990), 365.

Smith, C., Edwards, P., & Free, C. (2018). Assessing the validity and reliability of self-report data on contraception use in the MObile Technology for Improved Family Planning (MOTIF) randomised controlled trial. *Reproductive health*, *15*(1), 50.

Visscher, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., ... & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, *2*(3), e41.

Vuvor, F., & Harrison, O. (2017). A study of the diurnal height changes among sample of adults aged 30 years and above in Ghana. *Biomedical and Biotechnology Research Journal (BBRJ)*, *1*(2), 113.

Wright, C. M., & Cheetham, T. D. (1999). The strengths and limitations of parental heights as a predictor of attained height. *Archives of disease in childhood*, *81*(3), 257-260.