# A statistical model of the intergenerational movement of traits for populations reproducing under polygenic inheritance

Jesse Murray, Minjoon Kouh

March 2020

## Abstract

This work describes and presents the results from a mathematical model based on the linear regression equation of polygenic inheritance. When applied to intergenerational movement between quintiles, the model obtained an $R^2$ of 0.92 and 0.93 with the Brookings Institution measures of intergenerational education and income mobility, respectively. The model better predicted measures of education and income mobility than those measures predicted one another: $R^2=$ 0.84. One original question motivated the creation of the model: consider the tallest one fifth of trees in a forest. Under polygenic inheritance, are a majority of them the offspring of the previous generation's tallest one fifth of trees or are a majority of them the offspring of the previous generation's shorter four fifths of trees? While tall trees are more likely to have tall offspring, there are far more average/short trees than tall trees. It is not immediately clear whether or at what point the effect of a higher probability of tall offspring outweighs the effect of a far greater number of offspring. A simulation of the model showed that a minority (43%) of trees above the 80th percentile are the offspring of the previous generation's tallest one fifth. The 72nd percentile is the equilibrium point at which the proportion is 50%. That is, of the trees above the 72nd percentile, half are the offspring of parents also above the 72nd percentile and half are the offspring of parents below the 72nd percentile.

# Introduction

In biology, a phenotypic trait is a measurable trait that results from the expression of genes. As an example, the phenotype of hair color is the observed color while the genotype is the underlying genes that determine the color. The phenotypic traits Mendel studied in pea plants were unique in that they were determined single genes. However, it is often the case that important phenotypic traits are determined by many genes - in some cases hundreds or thousands. These traits are termed polygenic traits.

In general, the population distribution for the phenotype of a polygenic trait falls into a normal distribution. This phenomenon has been observed by plotting the frequency of phenotypes for a polygenic trait and finding a close approximation to a normal distribution. As described by Lange in his work on polygenic inheritance models, as the number of genes influencing a trait increases, the phenotypes in a population tend towards normality [Lange, 1997a, Lange, 1997b]. This phenomenon is thought to occur because of the many possible allelic combinations among individual genes. In this additive genetic model, genes code for alleles with either positive or negative effects on a measurement of the trait [Rieger et al., 1968].

One example of a polygenic trait is height: there are roughly 700 genes known to influence human height, each of which has a very small positive or negative effect on the measured trait [Wood and Esko, 2014]. The resultant population distribution of height is then Gaussian. Note that the Gaussian shape of the distribution is not contingent on how the individual gene effects are distributed. This is because the central limit theorem implies a normal population distribution about the optimal phenotype, the average of the individual gene effects, regardless of their shape - uniform, normal, etc. Polygenic inheritance of height can loosely be compared to flipping 700 coins and recording the number of heads minus the number of tails. (In this case, the individual gene effects fall into a shifted Bernoulli distribution.) If one were to run this experiment many times - once for each member in the population - the distribution of experimental results would fall into a normal distribution. That is, the experiments would most frequently result in a balanced number of heads versus tails and occasionally result in a largely imbalanced number of heads versus tails. It is worth noting that in the case of height, the phenotype is univariate, meaning that it is measured by one value. However, traits are sometimes multivariate, and though the work presented here does not discuss such cases, future work likely will.

As the phenotypes of a population fall under a normal distribution, their frequencies can be modeled by the equation for a normal distribution (1). In this equation, the parameter $\mu$ is the mean of the distribution (and also its median and mode); $\sigma$ is its standard deviation.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{1}$$

When the population being described is the parent generation, the distribution is made up of all the parent phenotypic values $x_p$ and their corresponding frequencies $f(x_p)$ are given by equation 2. The parameters $\mu_{pd}$ and $\sigma_{pd}$ are the mean and standard deviation of the parent generation population.

$$f(x_p) = \frac{1}{\sigma_{pd}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_p-\mu_{pd}}{\sigma_{pd}})^2} \tag{2}$$

It is important to be precise about what is meant by the parent generation. When discussing intergenerational mobility and inheritance, it is common to separate the parent generation from the offspring/child generation for individual families [Becker et al., 2018, Mulligan, 1999]. For inheritance within a family, the phenotype of a single parent (the father or the mother) is distinguished from average phenotype of the parents, i.e., the mid-parental phenotype [Luo et al., 1998]. The calculation of mid-parental height has been a standard procedure for assessing the heights of individual children since it was first described by Tanner [Tanner et al., 1970]. The paradigm of considering the parent generation apart from the offspring generation can be extended to a population as a whole. Then, the distribution of the single parent phenotypes is distinguished from the distribution of the mid-parental phenotypes. In the latter case, the parents in the population are already matched to one another and their phenotypic values are averaged. The mid-parental phenotype distribution has a smaller variance than the pre-matched parent generation due to the regression that occurs in parent to parent matching. Parameter values on mid-parental phenotypes can be estimated from the data on the heights of adult children and their parents for 197 families collected by Francis Galton in 1885. In this dataset, the standard deviation scores (SDSs) of fathers and mothers are not found to correlate (r = 0.074, p = 0.304) and the distribution of mid-parental phenotypes is found to have roughly one half (0.54) the variance of the distribution of individual parent phenotypes [Galton, 2017].

To the extent that polygenic traits are heritable, there exists a correlation between the phenotypes of parents and offspring. Previous studies have sought to measure the correlation between mid-parental and child SDSs for height. A pediatric growth study of English children found the correlation to be 0.47 (p $< 0.01$) [Wright and Cheetham, 1999]. An additional pediatric growth study of Swedish children found the correlation to be 0.59 ($p < 0.01$) [Luo et al., 1998]. To the extent that parent and offspring generation populations have the same variance, the correlation coefficient can be used to estimate the expected offspring/child height from the mid-parental height. This results from linear regression, in which there is a straight regression line (3) that provides the 'best' fit for related data points.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{3}$$

In the case of polygenic inheritance, $x$ refers to a mid-parental phenotypic value and $\hat{y}$ refers to the expected phenotypic value of the offspring of the parents (which from now on will be indicated by $\bar{x}_o$). The parameters $\alpha$ and $\beta$ are found by minimizing the sum of squared residuals between the mid-parental and offspring phenotypes. It can be shown that $\hat{y}$ is given by equation 4, in which $r$ is the correlation coefficient given by equation 5.

$$\frac{\hat{y}-\bar{y}}{s_y} = r\,\frac{x-\bar{x}}{s_x} \tag{4}$$

$$r = \frac{Cov[x,y]}{Var[x]} \tag{5}$$

In polygenic inheritance, the $x$ data are the mid-parental heights of the parent generation population. Mid-parental height can be calculated by taking the arithmetic mean of the mother's and father's height [Tanner et al., 1970]. Alternatively, male and female heights can first be normed by adding to all female heights the positive difference between the average male and average female height, then the arithmetic mean of the normalized mother's and father's height can be taken [Wright and Cheetham, 1999]. This method used by Wright and Cheetham to normalize male and female heights can be applied separately to the parents and adult children in the Galton dataset to compare the mean and standard deviation of the parents and adult children. Then, the mean and standard deviation of the normalized adult children is found to be 0.00583% less than and 4.98% greater than those of the normalized parents, respectively. That is, the distribution of the parent generation appears to be more or less identical to the offspring generation, implying a stable population between generations. Nevertheless, the distribution of mid-parental heights will necessarily have a smaller standard deviation than the distribution of offspring heights because the mid-parental heights regress to the mean in parent to parent matching, as discussed earlier. However, assuming the offspring generation population exhibits similar levels of regression in offspring to offspring matching as the parent generation does, the mean and standard deviation of the post-matched offspring generation population and the post-matched parent generation population are equal. Furthermore, because offspring to offspring matching involves a simple linear transformation, the well measured correlation between post-matched (mid-parental) heights and offspring heights is the same between the post-matched parent heights and the post-matched offspring heights.

With this understanding, the expected post-matched phenotypic value for the offspring of the parents at the post-matched phenotypic value $x_p$ is given by equation 6 and is similar to previous linear regression equations [Luo et al., 1998, Wright and Cheetham, 1999].

$$\bar{x}_o = \mu_{pd} + r(x_p - \mu_{pd}) \tag{6}$$

The parameter $\mu_{pd}$ is the mean phenotype of the parent population and the parameter $r$ is the well-measured correlation coefficient between the mid-parental (post-matched) and child/offspring phenotypes. While equation 6 gives the mean post-matched phenotypic value of the offspring of parents at $x_p$, it fails to describe its general distribution. In this work, it is suggested that the offspring of members of the parent population with phenotypic value $x_p$ are normally distributed with a mean at $\bar{x}_o$. The offspring distributions from each $x_p$ in the parent distribution sum to form the total offspring distribution. By keeping track of the contribution of sections of the parent distribution to sections of the total offspring distribution, it is possible to make meaningful statements about the intergenerational movement of traits for reproducing populations in nature and society.

## One Offspring Distribution

This work proposes that the frequency of the phenotypic values for the offspring of parents at $x_p$ is normally distributed about $\bar{x}_o$. The distribution of the phenotypic values of the offspring of parents at $x_p$ is then given by the following equation:

$$g(x) = f(x_p)\frac{1}{r_s\sigma_{pd}\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\bar{x}_o}{r_s\,\sigma_{pd}})^2} \tag{7}$$

The offspring distribution is a centered at $\bar{x}_o$. Its standard deviation is the parent generation population standard deviation $\sigma_{pd}$ scaled by $r_s$ and each of its values are scaled by the frequency of the parent phenotypic value f($x_p$).

If $r_s = 1$, then the variance of the offspring from parents at $x_p$ is equal to the variance of the entire parent generation population. While there are not yet literature measurements of $r_s$, it would seem to be more likely that the variance is less than, and almost certainly not greater than that of the entire parent population. In that case, $r_s$ is more likely less than 1 as opposed to equal to or greater than 1. In a more complicated scenario not considered here, $r_s$ varies with $x_p$.

Note that the phenotypic value $x_p$ corresponds to the z-score $z_p$ - relative to the parent generation population. A complete description of the one offspring distribution can be made with the following statement and two equations:

The distribution of the offspring of parents at $x_p$ is a normal distribution centered at z-score $z_o$ (relative to the parent generation population), with standard deviation $\sigma_o$, and proportional to the value at f($x_p$).

$$z_o = r\,z_p \tag{8}$$

$$\sigma_o = r_s\,\sigma_{pd} \tag{9}$$

The statement and two equations do not supply any additional information about the one offspring distribution. Instead, they provide an alternative way of describing the one offspring distribution that more clearly indicates the role of $r$ and $r_s$.

## Total Offspring Distribution

While $g(x)$ describes the distribution of offspring from only one $x_p$, a function is needed to describe the distribution of the entire offspring-generation population. This distribution is made up of the combined one-offspring-distributions from each $x_p$ in the parent generation population. The frequencies of the phenotypes of the offspring-generation population can then be described by the following probability density function.

$$G(x) = \int_{-\infty}^{\infty} g(x)\, dx_p \tag{10}$$

The frequency of each phenotypic value $x$ in the offspring-generation population is obtained by summing the frequency at $x$ for each one-offspring-distribution $g(x)$.

It is important to remark that this distribution $G(x)$ appears by all measures to be a normal distribution. This lends credence to the model as the offspring-generation population should indeed be normally distributed, and in most cases have a mean and standard deviation equal to those of the parent generation distribution. The mean of the total offspring distribution is always equal the mean of the (total) parent distribution. On the other hand, the standard deviation of the total offspring distribution varies proportionally with both $r$ and $r_s$.

## Answering the Motivating Question

At this point, it would seem to be possible to answer the motivating question: Are a majority of the tallest one fifth of trees in a forest the offspring of the previous generation's tallest one fifth? It is important to recognize that the area under a specific section of a population distribution bounded by phenotypic values represents the size of the population with those phenotypic values. In the case of the tallest one fifth of trees in a forest, the section is bound by $k_2$ and $\infty$, where $k_2$ represents the phenotypic value (height) at the 80th percentile of the population distribution. For a given phenotypic value $x_p$ in the parent generation population, it is necessary to find the size of its offspring population that is located in the top quintile. This is achieved by integrating $x_p$'s one offspring distribution from $k_2$ to $\infty$:

$$f(x_p) \frac{1}{\sigma_o \sqrt{2\pi}} \int_{k_2}^{\infty} e^{-\frac{1}{2}(\frac{x-\bar{x}_o}{\sigma_o})^2} dx \tag{11}$$

The integral provides the amount of offspring with a phenotypic value above $k_2$ from parents with the phenotypic value $x_p$ .

To find what proportion of the offspring in the top fifth of the offspring-generation population are from parents in the top fifth of the parent generation population, it is necessary to divide the amount of top fifth offspring from only those $x_p$ in the top fifth of the parent population by the amount of top fifth offspring from all $x_p$ in the parent population. This fraction gives the proportion of top fifth offspring from top fifth parents, the answer to the motivating question. The $x_p$ in the top fifth of the parent distribution are bounded by $k_1$ and $\infty$, where $k_1$ represents the height at the 80th percentile of the parent distribution. The following expression gives the amount of top fifth offspring from the top fifth parents.

$$\int_{k_1}^{\infty} f(x_p) \frac{1}{\sigma_o\sqrt{2\pi}} \int_{k_2}^{\infty} e^{-\frac{1}{2}(\frac{x-\bar{x}_o}{\sigma_o})^2} dx\, dx_p \tag{12}$$

This expression is then divided by the amount of top fifth offspring from all parents, which is a similar expression. The only difference is that the outer integral ranges over all members of the parent distribution (-$\infty$ to +$\infty$). The inner integral can be simplified with the cumulative distribution function.

# Intergenerational Movement and Two Types of Questions

The calculations involved in answering the motivating question can be generalized to answer two types of questions.

The first type of question is to ask what proportion of an arbitrary section of the total offspring distribution is from another arbitrary section of the parent distribution. For example, one could ask what proportion of the offspring-generation population with z-scores of between 1 and 1.5 are the offspring of members of the parent generation population with z-scores of between -0.5 and 0. The motivating question was of this type, as it asked what proportion of a top section of the total offspring distribution was from the same top section of the parent distribution.

The second type of question is to ask what proportion of the offspring of parents in an arbitrary section of the parent distribution end up in another arbitrary section of the total offspring distribution. For example, one could ask what proportion of the offspring from parents with z-scores of between -2 and -1, have z-scores of between 1 and 2.

In answering these questions, it is helpful to define a $\Phi$ term as follows.

$$\Phi(k_1, k_2, k_3, k_4) \equiv \int_{k_1}^{k_2} f(x_p) \frac{1}{\sigma_o\sqrt{2\pi}} \int_{k_3}^{k_4} e^{-\frac{1}{2}(\frac{x-\bar{x}_o}{\sigma_o})^2} dx\, dx_p \tag{13}$$

This term gives the size of the population with phenotypic values between $k_3$ and $k_4$ that are the offspring of members of the parent generation with phenotypic values between $k_1$ and $k_2$. In other words, it provides the amount of a specific section of the offspring-generation population from a specific section of the parent generation population.

### Proportion Attributable

To answer the first type of question, it is necessary to find the ratio of the $\Phi$ term for the specific section of the parent and offspring-generation population divided by the $\Phi$ term for the specific section of the offspring-generation population, but the entire parent generation population. This gives the proportion of the arbitrary section of the total offspring distribution that is the offspring of or 'attributable to' the arbitrary section of the parent distribution. The proportion is equivalent to the probability that a given member of the arbitrary section of the total offspring distribution is the offspring of a member of the arbitrary section of the parent distribution. The proportion attributable is given by the following equation.

$$P_a(k_1, k_2, k_3, k_4) = \frac{\Phi(k_1, k_2, k_3, k_4)}{\Phi(-\infty, \infty, k_3, k_4)} \tag{14}$$

The parameters $k_3$ and $k_4$ give the bounds of the arbitrary section of the total offspring distribution and the parameters $k_1$ and $k_2$ give the bounds of the arbitrary section of the parent distribution.

### Proportion Destined

To answer the second type of question, it is necessary to find the ratio of the $\Phi$ term for the specific section of the parent and offspring-generation population divided by the $\Phi$ term for the specific section of the parent generation population, but the entire offspring-generation population. This gives the proportion of the offspring from the arbitrary section of the parent distribution that end up in or are 'destined to' the arbitrary section of the total offspring distribution. The proportion is equivalent to the probability that a given offspring of a parent in the arbitrary section of the parent distribution is a member of the arbitrary section of the total offspring distribution. The proportion destined is given by the following equation.

$$P_d(k_1, k_2, k_3, k_4) = \frac{\Phi(k_1, k_2, k_3, k_4)}{\Phi(k_1, k_2, -\infty, \infty)} \tag{15}$$

The parameters $k_3$ and $k_4$ give the bounds of the arbitrary section of the total offspring distribution and the parameters $k_1$ and $k_2$ give the bounds of the arbitrary section of the parent distribution.

## Comparison with Galton data

Data on the heights of adult children and their parents for 197 families was collected by Francis Galton in 1885 and is provided by the Harvard Dataverse [Galton, 2017].

Previous work by Mulligan (1997) compared the linear regression approach that Galton suggested to predict education the Panel Study of Income Dynamics (PSID) data [Mulligan, 1999].

# Discussion

While the equations in the model do not have closed form solutions, they can be simulated with code. As a result, the answers to the questions presented here are approximations as the simulations are limited by computational speed.

To obtain values for intergenerational movement between quintiles, $P_d$ was obtained for each quintile of the parent and total offspring distributions. The $P_d$'s were then compared to the measured values for education and income mobility provided by the Brookings Institution. If income and education are normally distributed in the population with regression towards the mean between parent and offspring, then a high correlation between the values provided by this model and those provided by the Brookings Institution might indicate that the equations presented here provide a good model of reproducing normal population distributions with regression towards the mean.

# References

[Becker et al., 2018] Becker, G. S., Kominers, S. D., Murphy, K. M., and Spenkuch, J. L. (2018). A theory of intergenerational mobility. *Journal of Political Economy*, 126(S1):S7–S25.

[Galton, 2017] Galton, F. (2017). Galton height data. Harvard Dataverse.

[Lange, 1997a] Lange, K. (1997a). An approximate model of polygenic inheritance. *Genetics*, 147(3):1423–1430.

[Lange, 1997b] Lange, K. (1997b). *The Polygenic Model. In: Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health.* Springer, New York, NY.

[Luo et al., 1998] Luo, Z. C., Albertsson-Wikland, K., and Karlberg, J. (1998). Target height as predicted by parental heights in a population-based study. *Pediatric Research*, 44(4):563–571.

[Mulligan, 1999] Mulligan, C. B. (1999). Galton versus the human capital approach to inheritance. *Journal of Political Economy*, 107(S6):S184–S224.

[Rieger et al., 1968] Rieger, R., Michaelis, A., and Green, M. (1968). *A Glossary of Genetics and Cytogenetics*. Springer, New York, NY.

[Tanner et al., 1970] Tanner, J. M., Goldstein, H., and Whitehouse, R. H. (1970). Standards for children's height at ages 2-9 years allowing for height of parents. *Archives of Disease in Childhood*, 45(244):755–762.

[Wood and Esko, 2014] Wood, A. R. and Esko, T., e. a. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186.

[Wright and Cheetham, 1999] Wright, C. M. and 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.

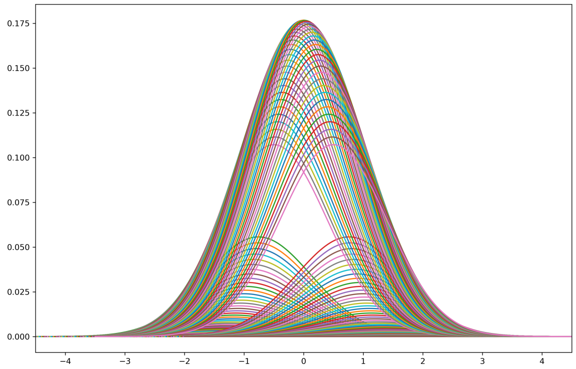

Figure 1: Example of a cool plot.