A Population Model of Polygenic Inheritance

Jesse Murray

March 2020

Abstract

A population model of polygenic inheritance is derived from the linear regression and normality of polygenic traits as first discovered by Francis Galton in the late 19th century. A simulation of the model is run with the measured parameter values r and r_s in the data from Galton's famous study on the heights of adult children and their parents. The simulation is shown to successfully construct an offspring generation that highly replicates the parent generation - indicating a stable population distribution between generations. A quintile transition matrix of intergenerational mobility is created, obtaining an R^2 of 0.81 ($p < 0.001$) with Galton's data and an R^2 of 0.96 ($p < 0.001$) with US family income data. Calculations of intergenerational upward persistence are made, which highly correspond to measurements in Galton's data. An equilibrium point is obtained at the percentile score of 71.7%, which compares well with the measured equilibrium percentile score of 72.2% in Galton's data. (The equilibrium indicates that those above the 72nd percentile are equally likely to be the children of parents above the 72nd percentile and of parents below the 72nd percentile.) Furthermore, the paper demonstrates the roles r and r_s have in determining population variance stability between generations.

I received helpful comments from my research advisor Dr. Minjoon Kouh and fellow student Aidan Carter. I also benefited from a presentation of this research to the Drew University Math Club.

1 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 [Schacherer, 2016]. However, important phenotypic traits are often determined by many genes - in some cases hundreds or thousands. These traits are termed polygenic traits. In general, the phenotypic population distribution for polygenic traits follows 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]. The approximation of normality 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].

Human stature is known to be a non-sex-linked, nondominant, polygenic trait [Preece, 1996]. 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 average of the individual gene effects - the optimal phenotype - regardless of the distribution of the individual gene effects - uniform, normal, etc. The inheritance of height can 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 the work presented here does not discuss such cases.

As the phenotypes of a population follow a normal distribution, their frequencies can be modeled by the equation for a normal distribution (1). In this equation, the parameter μ is the mean of the distribution (and also its median and mode); σ 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 μ_{pd} and σ_{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 898 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 α and β 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 and s_x, s_y denote the standard deviations of the x and y data, respectively.

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

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

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 father's and adjusted mother'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 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 μ_{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. It is worth reiterating that this correlation will be the same between the post-matched parent generation and the post-matched offspring generation, assuming the offspring generation exhibits similar amounts of regression in

mother-father pairings. Therefore, when comparing the parent generation to the offspring generation, it is equivalent to compare the post-matched generations with one another, and the pre-matched generations with one another. The conversion from pre-matched to post-matched consists of a linear transformation of reducing the variance by the same proportion, about one-half. Then, $\bar{x}_o$ can be considered pre-matched or post-matched, as long as x_p is considered the same. While equation 6 gives the mean of the distribution of the post-matched phenotypic values of the offspring of parents at x_p, it fails to describe the shape of the distribution. An analysis of the Galton data reveals the distribution to be closely approximated by a normal distribution about the expected offspring value $\bar{x}_o$, displayed in Figure 1. The offspring height in the Galton data is predicted with equation 4, in which the correlation between mid-parental and offspring height is found to be 0.51 ($p < 0.01$) [Galton, 2017]. Furthermore, when combining the residuals of the male and female adult children using the Wright and Cheetham normalization method, the standard deviation of the residuals is found to be 2.15 in, which compares to the parent generation standard deviation of 2.39 in. The ratio of the standard deviation of the offspring distribution residuals to the standard deviation of the parent generation distribution is then 0.90. This quantity will from now on be indicated by r_s.

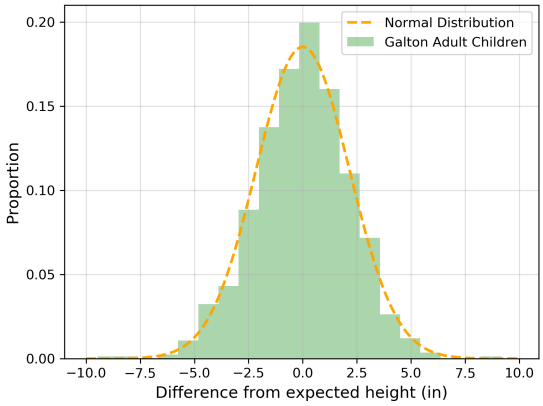


Figure 1: Combined residuals for all adult children in the Galton dataset. Male and female heights normalized with the Wright and Cheetham method [Galton, 2017, Wright and Cheetham, 1999].

A linear regression equation of the form of equation 6 dates back at least to Francis Galton. From the data he collected on height, Galton concluded that a person's characteristics are positively correlated with those of his parents. However, personal characteristics also "regress to mediocrity" so that, on average, the personal characteristics of a child are less extreme (i.e., closer to the mean) than those of his parents [Galton, 1877]. Galton's model can be represented as equation 7 in which x denotes an adult's personal characteristic such as height, and ϵ represents determinants of the adult's personal characteristic that is uncorrelated with the parent's personal characteristic.

$$x_{t+1} = \alpha + \beta x_t + \epsilon_{t+1} \ , \ \beta \in (0, 1) \tag{7}$$

Galton suggested that the appropriate model of inheritance has $\beta \in (0, 1)$ and estimated β for height to be about 2/3 when x_t was measured as an average of maternal and paternal characteristics and 1/3 when x_t was measured as a single parent's characteristic. Galton even suggested that the same 2/3 dictates the inheritance of any personal characteristic [Galton, 1889]. Indeed, Mulligan writes that Galton's model gives similar predictions to other models of intergenerational income mobility. Mulligan concludes that "...the challenge facing economists is to produce a model of intergenerational mobility with predictions that are (a) distinct from Galton's, and (b) true." [Mulligan, 1999]. Any population model of polygenic inheritance should thus be assed by how well it predicts the mobility of socioeconomic characteristics in addition to polygenic ones.

In the population model proposed in this paper, the individual offspring distributions from each x_p are distributed normally about the expected offspring value $\bar{x}_o$ with a standard deviation $r_s \, \sigma_{pd}$. These

distributions sum to form the distribution of the offspring generation. By integrating the contributions from sections of the parent distribution to sections of the total offspring distribution, important predictions can be made about the intergenerational movement of traits for populations that reproduce with regression to the mean. The model can be assessed by how closely its constructed offspring generation aligns with the measured offspring generation from the Galton data, assuming the measured parameter values for r and r_s in the Galton data. Furthermore, the model can be assessed by how well it predicts the intergenerational mobility of characteristics both for height and for income. Finally, the model can provide a theoretical answer to the original question that motivated the creation of the model: Consider the tallest members of a population; are most of them the children of the last generation's tallest members or shorter/average members? Although the tall members of the last generation have a higher probability of having tall children, there are many more average/short members than tall members. More specifically, at what SDS are those above the SDS equally the children of parents above the SDS and parents below the SDS? The population model can be validated by how well its answer to this question corresponds to the measured value in the Galton data.

2 Derivation of the Model

2.1 Individual Offspring Distribution

In the proposed model, the post-matched phenotypic values for the offspring of parents at x_p is normally distributed about $\bar{x}_o$, from equation 6, with a standard deviation $r_s\,\sigma_{pd}$. The frequencies are scaled by the frequency of the matched parent phenotypic value $f(x_p)$, from equation 2, to reflect the frequency of the mid-parental phenotype x_p. The distribution is then given by equation 8.

$$g(x) = f(x_p)\frac{1}{r_s\sigma_{pd}\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\bar{x}_o}{r_s\,\sigma_{pd}})^2} \tag{8}$$

If r_s was 1, then the variance of the offspring from parents at x_p would be equal to the variance of the parent generation population. Instead, the measured $r_s = 0.90$ from the Galton data is used. Furthermore, if r was 1, then the center of the individual offspring distribution would be at the phenotype of the parents x_p, this would indicate complete inheritance and no regression toward the mean. On the other hand, if r were to be 0, then then parents at x_p would have offspring centered at the population mean, which would indicate no effect of inheritance and complete regression toward the mean. Instead, r is set to be 0.50, which is similar to the measured value of 0.51 from the Galton data. (The value 0.50 is used instead of 0.51 to run time, described in section 3.) An example of an individual offspring distribution is displayed in Figure 2. It is important to note that $g(x)$ does not indicate the absolute frequency but rather the relative frequency of the individual offspring. When the individual offspring distributions are combined to form the combined offspring generation distribution, they are scaled by a multiplicative factor m in equation 12, such that the size of offspring generation is equal to the size of the parent generation, for a stable population across generations.

The phenotypic value x_p corresponds to the z-score z_p - relative to the parent generation population. A complete description of the individual offspring distribution can then be made with equations 9, 10 and the following statement: The distribution of the offspring of parents at x_p is a normal distribution with frequencies proportional to $f(x_p)$.

$$z_o = r\,z_p \tag{9}$$

$$\sigma_o = r_s\,\sigma_{pd} \tag{10}$$

The statement and two equations provide an alternative - perhaps simpler - way of describing the individual offspring distribution than equation 8 and elucidates the role of r and r_s.

2.2 Offspring Generation 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

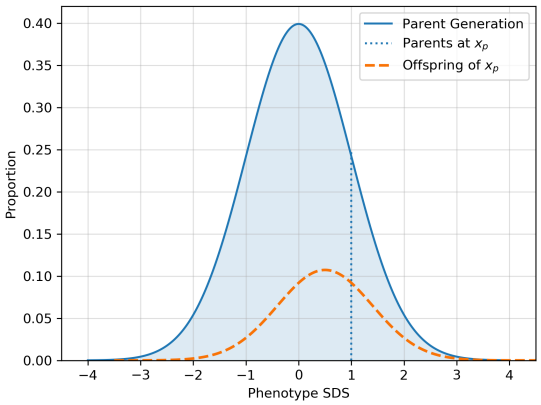


Figure 2: The individual offspring distribution of parents with the mid-parental SDS of 1. The parameter values $r = 0.50$ and $r_s = 0.90$ used in the simulation are drawn from the measured Galton data parameters [Galton, 2017].

individual 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 the probability density function in equation 11.

$$G(x) = m \int_{-\infty}^{\infty} g(x)\, dx_p \tag{11}$$

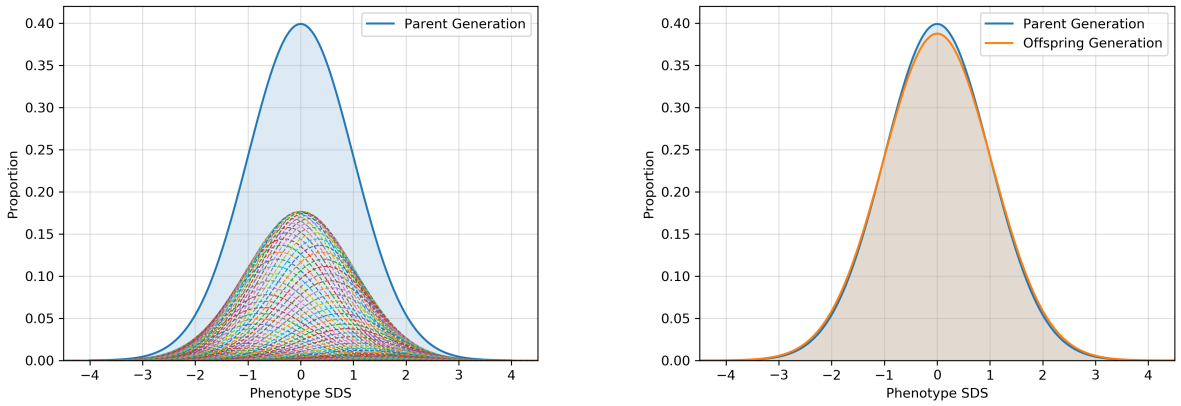
$$m = \frac{\int_{-\infty}^{\infty} f(x_p)\, dx_p}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)\, dx_p\, dx} \tag{12}$$

The frequency of each phenotype x in the offspring generation population is obtained by summing the frequency at x from the individual offspring distribution $g(x)$ of each x_p in the parent generation, see equation 8. In order for the size of the offspring generation to be equal to the size of the parent generation, each frequency in the offspring generation distribution is scaled by the multiplicative factor m in equation 12.

In Figure 3a, 100 example individual offspring distributions are shown that combine through equation 11 to form the offspring generation shown in Figure 3b. (Figure 3b is actually simulated from 1000 individual offspring distributions.) The parameters of $r = 0.50$ and $r_s = 0.90$ from the Galton data produce an offspring generation that has a standard deviation only 2.99% greater than that of the parent generation. The offspring generation in the Galton data had a similar result, with the standard deviation of the normalized adult children 4.98% greater than that of the parents with the measured values $r = 0.511$ and $r_s = 0.902$. As described later the standard deviation of the offspring generation distribution varies proportionally with r and r_s. Thus, it is not surprising that the Galton data, with greater values for r and r_s, had a larger ratio of the offspring generation to the parent generation. It is worth noting that the total offspring generation $G(x)$ as shown in Figure 3b follows a normal distribution, which may not have been an obvious result from the form of equations 11 and 8. The close approximation of the offspring generation distribution to the parent generation distribution lends credence to the model as the offspring generation should be normally distributed, and with the parameters of a stable population, have a mean and standard deviation equal to those of the parent generation distribution.

2.3 Intergenerational Movement of Two Forms

With the model created thus far, it is now possible to make calculations about intergenerational mobility. There appear to be two useful forms for describing intergenerational mobility. The first form is to give the proportion of an arbitrary section of the offspring generation distribution that is the offspring of another



(a) Individual offspring distributions.

(b) Combined offspring distribution.

Figure 3: The individual offspring distributions from 100 mid-parental values x_p in the parent generation (3a). The offspring generation made up of 1000 combined individual offspring distributions through equation 11 (3b). The parameter values $r = 0.50$ and $r_s = 0.90$ used in the simulation are drawn from the measured Galton data parameters [Galton, 2017].

arbitrary section of the parent generation distribution. For example, one could provide what proportion of the offspring generation population with SDSs of between 1 and 1.5 are the offspring of members of the parent generation population with SDSs of between -0.5 and 0. The motivating question was of this type, as it asked what proportion of a top section of the offspring generation distribution was from the same top section of the parent generation distribution. This form is referred to as the proportion attributable. The second form is to give the proportion of the offspring of parents in an arbitrary section of the parent generation distribution that end up in another arbitrary section of the offspring generation distribution. For example, one could provide, for parents with SDSs of between -2 and -1, the proportion of their offspring that have SDSs of between 1 and 2. This form is referred to as the proportion destined. In answering these questions, it is helpful to define the Φ term in equation 13.

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

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. This is because the area under a specific section of a population distribution bounded by phenotypic values represents the size of the population with those phenotypic values.

Proportion Attributable

The proportion attributable is the proportion of an arbitrary section of the offspring generation that is the offspring of or 'attributable to' an arbitrary section of the parent generation. The proportion is equivalent to the probability that a given member of the arbitrary section of the offspring generation is the offspring of a member of the arbitrary section of the parent generation. The proportion attributable is given by equation 14.

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

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

Proportion Destined

The proportion destined is the proportion of the offspring of parents in an arbitrary section of the parent generation that end up in or are 'destined to' the arbitrary section of the offspring generation. The proportion is equivalent to the probability that a given offspring of a parent in the arbitrary section of the parent generation will be a member of the arbitrary section of the offspring generation. The proportion destined is given by equation 15.

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

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.

3 Simulation of the Model

The equations in the model described in previous sections do not have closed form solutions and are be simulated with code in Python. The results presented in this paper are then approximations as the simulations are limited by computational speed. The simulation of the model went as follows: A two-dimensional array was created of n values for x ranging across an SDS range of -4 to $+4$ and n values for y calculated from the x values by the equation of a normal distribution (2). The parameter n was set to be 1000 for simulations in this paper under the trade-off between run time and accuracy. This initial distribution represented the parent generation. Then, similarly constructed arrays for the n individual offspring distributions (one from each x, y point in the parent generation distribution) were generated by equation 8. Thus, the individual offspring distributions consisted of n^2 (1,000,000) x, y pairs in memory.

It is important to note that choices of r and r_s often result in the x values of the individual offspring distributions failing to correspond with the x values of the parent distribution. For example, if the x value from the parent distribution, x_p, was 1.0 and $r = 0.51$, then the individual offspring distribution would be centered about 0.51. However, 0.51 is not one of the x values in the parent distribution, which ranges from -4 to $+4$ in increments of 0.008 (when $n = 1000$). This mis-alignment can be further exacerbated by the chosen value for r_s.

To allow for the summation of the individual offspring distributions to form the offspring generation distribution under conditions of mis-alignment between individual offspring x values and parent generation x values, the disparate individual offspring x values (and their corresponding y pairs) were binned together with the nearest x value from the parent distribution. They are then said to be normalized to the parent distribution increment, which in the case of the simulations in this paper is 0.008. Then, the offspring generation distribution is multiplied by the multiplicative value m given by equation 12, such that the area under the offspring generation distribution is equal to the area under the parent generation distribution.

The values for proportion attributable and proportion destined are both calculated in the same initially: by finding the Φ term in the numerator of equations 14 and 15, which is the same. This is accomplished by calculating the offspring distributions for only those parent x values between k_1 and k_2 and only calculating the offspring x values between k_3 and k_4, the rest are set to zero. Then, the area under those individual offspring distributions is calculated to obtain the Φ term. A similar procedure is executed to obtain the Φ terms in the denominators of equations 14 and 15. The standard deviation of the offspring generation distribution is not assumed to be equal to that of the parent generation distribution. Thus, if the k values are given as SDSs or percentiles instead of simple values, the standard deviation of the offspring generation distribution must first be calculated to make the conversions to simple values.

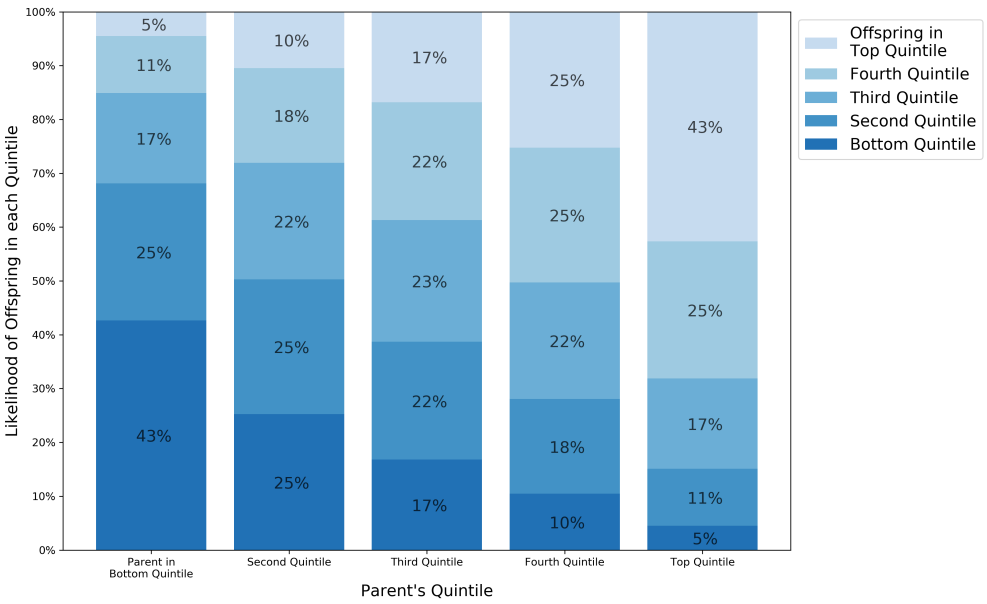
In making the simulation, consideration was taken of the effects on computational complexity that result from values of r and r_s that do not correspond with the parent generation increment, which is the quotient of the SDS range divided by the number of parent generation x values, n. The parameter values for the simulation were then chosen to be $r = 0.50$ and $r_s = 0.90$, in close alignment with the measured values from the Galton data of $r = 0.511$ and $r_s = 0.902$.

The simulations run to generate the standard deviation errors in section 6 used an n and SDS range of -3 to $+3$ to minimize run time.

4 Intergenerational Mobility Between Quintiles

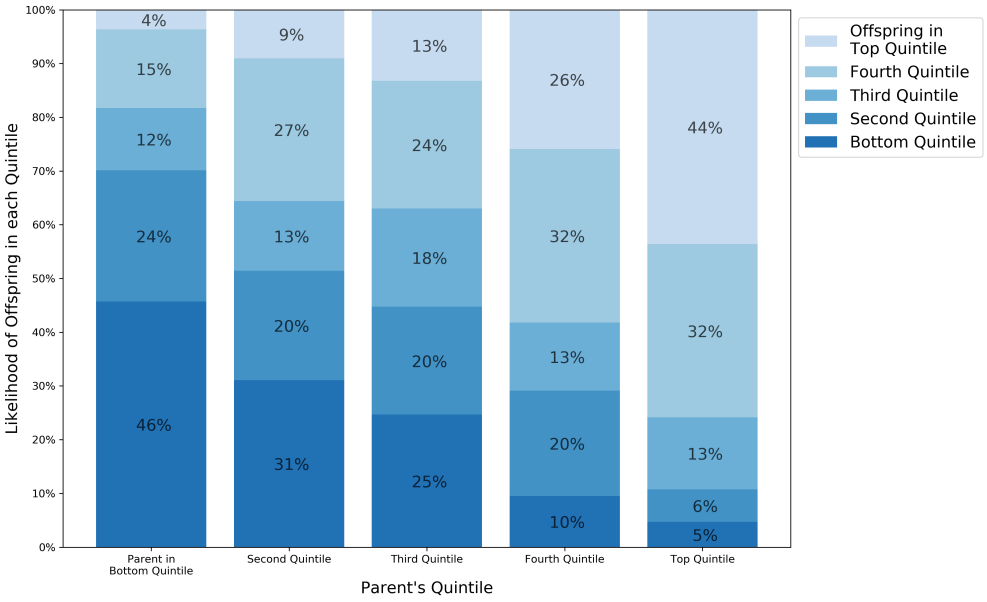
The model can now be used to make predictions about intergenerational mobility. A widely-used metric of intergenerational mobility is a quintile transition matrix [Chetty et al., 2014, Isaacs, 2008]. One advantage of this approach is that it clearly indicates how likely it is that someone born to parents in a given quintile will end up in another given quintile as adults. To obtain this quintile transition matrix, P_d was calculated for each quintile of the parent generation and total offspring generation and the results are plotted in Figure 4a. As a side note, simple arithemetic can show that the proportion destined and proportion attributable are equivalent in a percentile transition matrix. Thus, the transition matrix in Figure 4a also gives the proportion of offspring in each quintile of the offspring generation from parents in each quintile of the parent generation.

Some striking results are obtained, for example: Offspring of parents in the bottom quintile or top quintile are about 9 times more likely to end up in the same quintile as their parents than move to the opposite extreme quintile. On the other hand, offspring of parents in the middle quintile have a roughly uniform chance of ending up in any of the five quintiles. The transition matrix can be compared with that of the Galton height data shown in Figure4b. The correspondence between the two is high $R^2 = 0.81$ ($p < 0.00001$). Additionally, the transition matrix can be compared with that of US family income data from federal income tax returns and W2s as collected by Chetty et al. [Chetty et al., 2014]. This data is desired because it consists of the family income of adult children and their parents rather the income of individual adult children or individual parents. Unlike the Galton dataset of 898 children, the income dataset is large, consisting of 9,867,736 children born between 1980 and 1982. Its transition matrix is shown in Figure 4c. The correspondence between its transition matrix and that of the proposed model is very high $R^2 = 0.96$ ($p < 0.00001$). Additionally, the simulated transition matrix can be compared with that of US family income data from the smaller Panel Study of Income Dynamics (PSID) dataset of roughly 10,000 families [Isaacs, 2008]. The correspondence with this transition matrix is also high $R^2 = 0.91$ ($p < 0.00001$).

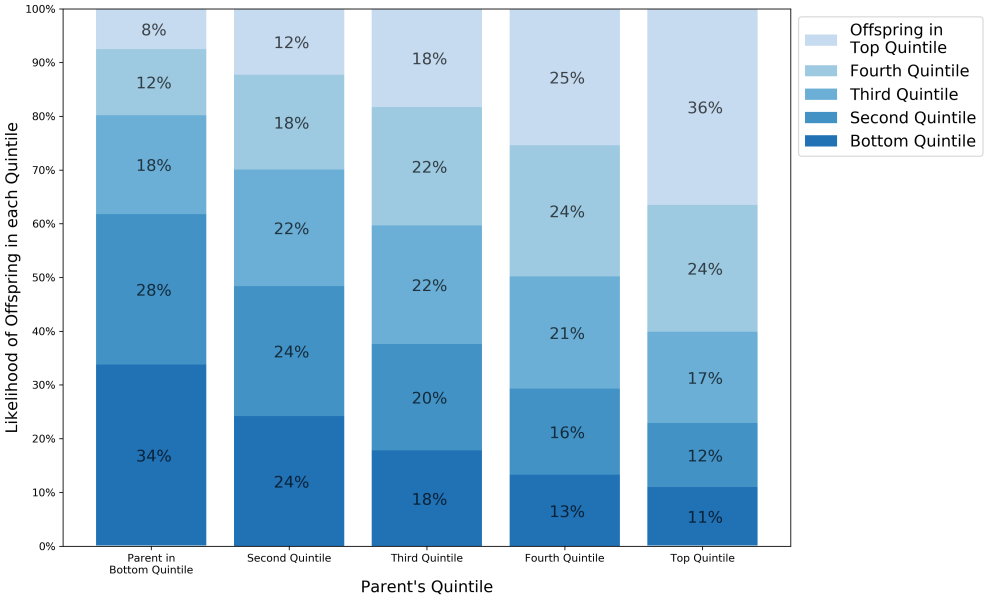


(a) Proposed Model.

Figure 4: Quintile transition matrices of intergenerational mobility. The transition matrix calculated from a simulation of the proposed model use parameter values $r = 0.50$ and $r_s = 0.90$ (4a). A transition matrix is measured from the Galton height data (4b) [Galton, 2017]. A transition matrix is provided by Chetty et al. for US family income data for the 1980-82 birth cohort (4c) [Chetty et al., 2014].



(b) Galton Height Data.



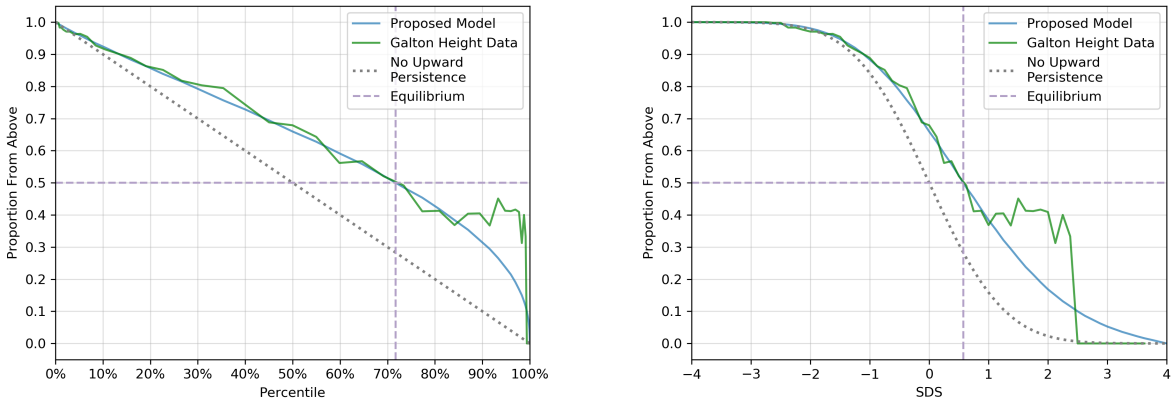
(c) US Family Income Data.

Figure 4: Quintile transition matrices of intergenerational mobility. The transition matrix calculated from a simulation of the proposed model use parameter values $r = 0.50$ and $r_s = 0.90$ (4a). A transition matrix is measured from the Galton height data (4b) [Galton, 2017]. A transition matrix is provided by Chetty et al. for US family income data for the 1980-82 birth cohort (4c) [Chetty et al., 2014].

5 Intergenerational Upward Persistence

Finally, a simulation of the model can provide an answer to the question that first motivated the creation of the model. Of those members at the top of a population distribution, for example, the tallest members of the population, to what extent are they the children of parents who were also in the same top section of the distribution? And to what extent are they the children of parents who were in lower sections of the population distribution? This question ultimately gets at the upward persistence of traits between generations. To be precise, intergenerational upward persistence measures the proportion of members of the offspring generation above a given SDS or percentile that are the offspring of members of the parent generation also above the same SDS or percentile.

Two effects appear to have counteracting effects on the persistence of traits between generations: likelihood and number. On the one hand, members of the top section of the parent generation distribution have a higher likelihood of having offspring that end up in the same top section. On the other hand, there are far fewer of those members of the top section of the parent generation distribution than there are of members of lower sections of the parent generation distribution. The intergenerational upward persistence is then a competition between higher likelihood for the top section of the parent generation and greater number for the lower sections of the parent generation. Of specific interest is the point (SDS or percentile) of equilibrium, that is, the point at which the two competing effects balance one another. Members of the offspring generation above this point are equally likely to be the offspring of parents above this point and below this point.



(a) Percentile Scores

(b) SDSs

Figure 5: Intergenerational upward persistence is given for each SDS in 5b and each percentile in 5a from a simulation of the proposed model and measurements from the Galton data. Proportions are calculated or measured in SDS increments of 0.125. The parameter values $r = 0.50$ and $r_s = 0.90$ used in the simulation are drawn from the measured Galton data parameters [Galton, 2017].

Calculations and measurements of intergenerational upward persistence are given from a simulation of the proposed model and the Galton height data, respectively, in Figure 5. A striking correspondence is observed between the two sources of data. It should be noted that in the Galton height data, there are only 51 adult children with an SDS equal to or greater than 1.5. Beyond this point, the quality of the data likely begins to suffer as a result of the small number of data points. This may partly explain the deviation between the Galton height data and the proposed model for SDSs or percentile scores above this point. Some observations can be made. The proposed model finds that about 30% of those above the 90th percentile are from parents with a mid-parental value that was also above the 90th percentile. In addition, about 67% (two-thirds) of those above the 50th percentile are from parents with a mid-parental value that was also above the 50th percentile. Finally, the proposed model finds the equilibrium point to be an SDS of 0.575, which corresponds to a percentile score of 71.7%. That is, those above the 72nd percentile are equally likely to be from parents below the 72nd percentile and above the 72nd percentile. This result corresponds

highly with the measured equilibrium point in the Galton data of an SDS of 0.590, which corresponds to a percentile score of 72.2%.

A depiction of the counteracting effects of likelihood and number can be seen clearly in Figure 6. The offspring of parents above the equilibrium SDS is depicted alongside the offspring of parents below the equilibrium SDS. Both offspring populations boast an equal number of offspring with SDSs above the equilibrium SDS, i.e., the area to the right of the equilibrium line is equal under both offspring distributions. However, this equality is reached differently for both offspring populations. The offspring distribution of parents above the equilibrium SDS is shifted to the right - indicating a greater likelihood of having a phenotype above the equilibrium SDS. Meanwhile, the offspring distribution of parents below the equilibrium SDS is much larger - indicating a greater number of total offspring, some of which could have a phenotype above the equilibrium SDS.

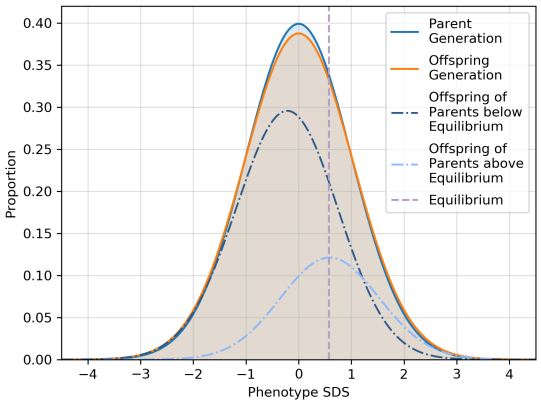
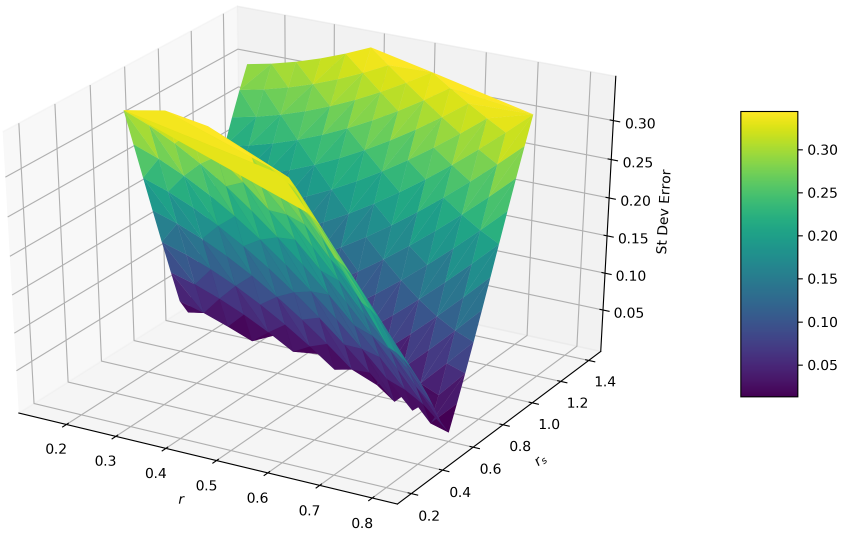


Figure 6: The offspring of parents above the equilibrium SDS and the offspring of parents below the equilibrium SDS. The parameter values $r = 0.50$ and $r_s = 0.90$ used in the simulation are drawn from the measured Galton data parameters [Galton, 2017].

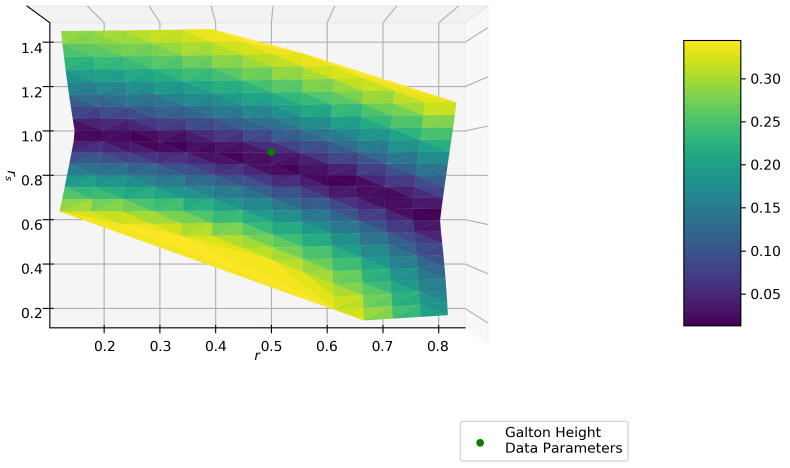
6 Parameters of a Stable Population

The main simulations presented in this study have used the parameters $r = 0.50$ and $r_s = 0.90$ drawn from the measured ones in the Galton height data, as discussed in section 2.2. Under both the model and Galton's height data, these parameter values generate an offspring generation distribution with only a slightly greater standard deviation than that of the parent generation distribution. That is, these parameters generate a stable population variance between generations. An unstable population variance between generations would be characterized by a gradual spreading out or condensing in of the population distribution about the optimal phenotype.

It is natural to ask then, what combination of r and r_s produce a stable population variance between generations? As a reminder, r and r_s are the parameters that describe the normal distribution of the individual offspring distribution from parents at a given phenotypic value. Their mathematical definitions are given by equations 10 and 9. Qualitatively, r is the regression coefficient between the parental and offspring phenotypes, r_s is the ratio of the standard deviation of the individual offspring distribution residuals to that of the standard deviation of the parent generation distribution. Figure 7 gives the error of the offspring generation standard deviation with that of the parent generation for varying r and r_s values in increments of 0.05. The left and right sides of the valley in Figure 7a indicate r and r_s values when the offspring generation standard deviation was less than and greater than the parent generation standard deviation, respectively. The combination of r and r_s that generate a stable population variance are shown by the dark purple line in Figure 7b, along with the r and r_s combination measured in the Galton height data.



(a) Side view.



(b) Top view.

Figure 7: The error of the offspring generation standard deviation with the parent generation standard deviation for varying r and r_s parameters.

The combination of r and r_s values provided by Figure 7 could be used to predict either the r value, r_s value, or variance stability of a population given two out of three of these pieces of information. For example, one could imagine a scenario in which the r and r_s values were measured from a small number of parents and their offspring. The model proposed in this study could then provide how the populations' variance would change between generations.

7 Multigenerational Mobility

Instead of simply obtaining the amount of intergenerational mobility between the parent and offspring generation, it is possible to obtain the mobility across multiple generations. That is, given an initial mid-parental (family) phenotype in some section of the population distribution, say above the 90th percentile, what is the likelihood that the family's descendant generations are found in any given section of the population distribution, say in each decile. (An in depth explanation of the math here is coming soon.)

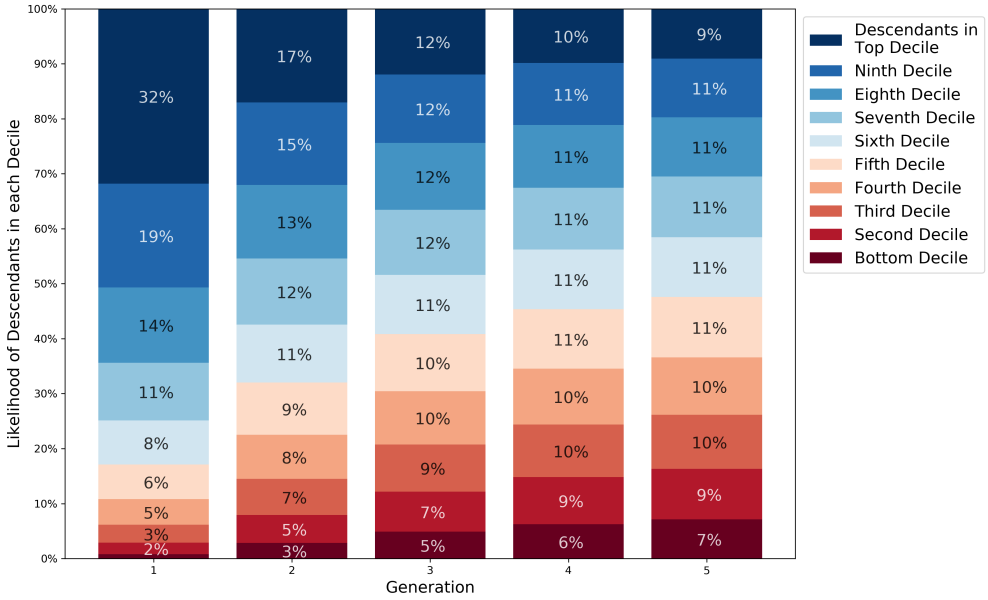


Figure 8: Multigenerational mobility. The multi-generational descendant matrix is calculated from a simulation of the proposed model use parameter values $r = 0.50$ and $r_s = 0.90$ [Galton, 2017].

8 Discussion

This study presents a population model based on the normality of individual offspring distributions - as demonstrated from the Galton height data - about an expected phenotype given by the linear regression model of polygenic inheritance [Galton, 2017, Galton, 1877, Luo et al., 1998, Wright and Cheetham, 1999]. The individual distributions are shown to combine to form the offspring generation. A simulation run on the measured parameters r and r_s from the Galton height data successfully produces an offspring generation that highly replicates the parent generation - indicating a stable population distribution across generations. The study shows how families regress to the mean across generations, but that the effects of an extreme family starting point can be detected at least three generations after. The study also demonstrates the roles r and r_s have in determining the stability in population variance between generations.

The model makes predictions of the intergenerational mobility of traits through the construction of a quintile transition matrix, which measures the likelihood that someone born to parents in a given quintile ends up in another given quintile as an adult; and a plot of intergenerational upward persistence, which measures the likelihood that someone in a top section of the population distribution was born to parents

also in that top section. The model's predictions are shown to highly correspond with measurements from the Galton height data - some of the inconsistencies are likely due to the limited size of the Galton height data (197 families and 898 children).

Most strikingly, the predictions of intergenerational mobility through the quintile transition matrix are shown to correspond very highly with quintile transition matrices of intergenerational income mobility. This result may be surprising because unlike height, income is not a polygenic trait. However, previous studies of genome-wide SNPs have in fact shown a significant genetic influence on family socioeconomic status (SES), including income [Trzaskowski et al., 2014, Krapohl and Plomin, 2016]. Krapohl and Plomin write that a high degree of heritability in SES is usually interpreted to mean a low degree of social mobility. However, removing environmental sources of variation will not remove genetically driven resemblance between parents and offspring. To the contrary, as environmental differences diminish, individual differences that remain will to a larger proportion be due to genetic differences, and the heritability would increase. As a result, Krapohl and Plomin argue that the heritability of SES could be seen as a positively correlating index of social mobility [Trzaskowski et al., 2014, Krapohl and Plomin, 2016].

In this study, the intergenerational mobility transition matrix produced by the proposed polygenic model was shown to agree well with the observed transition matrix of income mobility. If the proposed model's transition matrix over-predicted the amount of mobility between quintiles, then that would indicate low social mobility in the US. Instead, the opposite is observed. The transition matrix created by the proposed model run on Galton's parameters slightly under-predicts the amount of mobility between quintiles, especially for bottom and top quintile families. Therefore, this study finds there to be slightly more mobility in family income in the United States than would be predicted by the mobility of a polygenic trait using parameters for the inheritance of height.

The high correspondence between the model's transition matrix and that of US family income is also surprising given the non-normal distribution of income. Instead, income is generally considered to be best approximated a log-normal distribution [Battistin et al., 2007, Neal and Rosen, 1998]. The high correspondence may have resulted from the parent values in the income distribution producing sub-offspring distributions that regressed towards the mean, even though the overall shape of the income distribution differed from normal.

References

[Battistin et al., 2007] Battistin, E., Blundell, R., and Lewbel, A. (2007). Why is consumption more log normal than income? gibrat's law revisited. *Battistin, E. and Blundell, R. and Lewbel, A. (2007) Why is consumption more log normal than income? Gibrat's Law revisited. Working paper. IFS Working Papers (W08/07). Institute for Fiscal Studies, London, UK.*, 117.

[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.

[Chetty et al., 2014] Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, 129(4):1553–1623.

[Galton, 1877] Galton, F. (1877). Typical laws of heredity. *Proceedings of the Royal Institution*, 339(8):282–301.

[Galton, 1889] Galton, F. (1889). *Natural inheritance*. Macmillan, London.

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

[Isaacs, 2008] Isaacs, J. (2008). Economic mobility of families across generations. The Brookings Institution.

[Krapohl and Plomin, 2016] Krapohl, E. and Plomin, R. (2016). Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide snps. *Molecular Psychiatry*, 21(3):437–443.

[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.

[Neal and Rosen, 1998] Neal, D. and Rosen, S. (1998). Theories of the distribution of labor earnings. Working Paper 6378, National Bureau of Economic Research.

[Preece, 1996] Preece, M. A. (1996). The genetic contribution to stature. *Hormone Research in Pediatrics*, 45(Suppl. 2):56–58.

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

[Schacherer, 2016] Schacherer, J. (2016). Beyond the simplicity of mendelian inheritance. *Comptes Rendus Biologies*, 339(7):284 – 288.

[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.

[Trzaskowski et al., 2014] Trzaskowski, M., Harlaar, N., Arden, R., Krapohl, E., Rimfeld, K., McMillan, A., Dale, P. S., and Plomin, R. (2014). Genetic influence on family socioeconomic status and children's intelligence. *Intelligence*, 42(100):83 – 88.

[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.