Normal linear Markov model with applications to polygenic inheritance

Jesse Murray

University of Oxford, Oxford, UK,
jesse.murray@stats.ox.ac.uk

Abstract. A Markov model of polygenic inheritance is proposed. The initial condition is a population with a normal distribution of scores. Child scores relate to parent scores through a normal linear model, which is parametrized by a regression and residual coefficient. The normal population distributions for all generations and the normal conditional distributions for all descendants and ancestors are obtained. The conditional distributions exponentially converge to the population distributions and the two coefficients determine the rate constants such that a measure of intergenerational mobility is introduced as the ratio of the residual to the regression coefficient. The Markov model is reversible and stationary when the population variance is stable (constant) between generations. Percentile transition matrices enable visualizations of the model. From two large datasets on the heights of parents and their children, the proposed model is verified through statistical tests, and the regression and residual coefficients are estimated.

Keywords: Markov model, Normal, Gaussian, Polygenic

1 Introduction

Polygenic inheritance occurs when a trait is determined by many genes—in the range of hundreds or thousands. In general, the observed phenotypic scores of these traits are normally distributed, an effect of the central limit theorem [13, 6]. That is, many genes act largely independent of one another and combine through an additive sum to determine phenotypes, resulting in a normal distribution of the trait in the population [7].

One example of a polygenic trait is human height. At the time of writing, 697 variants have been identified at genome-wide significance, together explaining only one-fifth of the heritability of adult height [12, 15]. Indeed, height has an observed normal distribution in the population [8]. Furthermore, the normal linear model has been used to predict adult child height from parent height [8]. In this paper, the normal linear model is applied over multiple generations in the form of a Markov model to reveal insights into polygenic inheritance.

2 Model Formulation

In this paper, a novel Markov model is proposed that seeks to describe the inheritance of a univariate polygenic trait (such as human height) over multiple generations.

2.1 Markov Model of Inheritance

The Markov process exists in discrete time $i \in \mathbb{N}$, where i is the generation index (parent, child, grandchild, etc.); and continuous space $X_i \in \mathbb{R}$, where X_i is the phenotypic score of the trait. The terms *score* and *state* are used interchangeably. The proposed Markov model can be entirely derived from the initial condition and the one-step downward transition.

Conventions

- Let the tilde symbol, e.g. $\tilde{\mu}_{i+n} := \mathrm{E}(X_{i+n}|X_i)$, denote conditional parameters, such as conditional expectation and standard deviation (SD), that describe an upward or downward conditional distribution.
- Let the parameters μ_i and σ_i be the expectation and SD, respectively, of the marginal (population) distribution of state X_i.

1. Initial Condition. The initial score is drawn from a standard normal, denoted as

$$X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{1}$$

with $\mu_0 := 0$ and $\sigma_0^2 := 1$. This can be thought of as the marginal distribution of the initial generation, and a normally distributed univariate polygenic trait can be standardized to meet this condition. Due to the standardization of the initial generation, the scores of all generations are given in z-scores relative to the initial generation.

2. One-step Downward Transition. The conditional one-step downward transition, i.e., the score of a child, given the score of its parent, is

$$X_{i+1}|X_i \sim \mathcal{N}(\tilde{\mu}_{i+1}, \tilde{\sigma}_{i+1}^2)\ . \tag{2}$$

The one-step downward transition is parametrized by the regression and residual coefficients.

1. A child's score is proportional to the score of its parent by r—the *regression coefficient*:

$$\tilde{\mu}_{i+1} := rX_i\ . \tag{3}$$

We require

$$0 < r < 1 \tag{4}$$

for there to be regression towards the population mean.

2. There is a normally distributed residual about the mean that has an SD proportional to the marginal SD of the parent generation by s—the *residual coefficient*:

$$\tilde{\sigma}_{i+1} := s\,\sigma_i \ . \tag{5}$$

We require

$$s > 0 \tag{6}$$

for the residual SD to be positive.

The conditional one-step downward transition is in the form of a normal linear model, which is commonly used to model the relation between parent and adult child height [8]. This paper makes the novel generalization of the normal linear model into a Markov process, enabling the modeling of relations across multiple generations.

3 Properties of the Model

A variety of useful properties follow directly from the model formulation.

3.1 Marginal Population Distribution

The marginal distribution of the score of generation $i\,, \forall i \in \mathbb{N}$ is

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \tag{7}$$

with

$$\mu_i = 0 \tag{8}$$

and the marginal variance of generation $i + n\,, \forall n \in \mathbb{N}$ related to that of generation i is

$$\sigma_{i+n}^2 = (r^2 + s^2)^n\,\sigma_i^2 \ . \tag{9}$$

The marginal distribution can be obtained by considering the child score X_{i+1}, which can be written unconditionally in terms of the variance of the parent generation as

$$X_{i+1} = r\sigma_i Z_a + s\sigma_i Z_b \tag{10}$$

with Z_a and Z_b independent standard normals. Using the theorem of the sum of independent normal random variables (RVs), X_{i+1} is clearly described by Eqs. 7 - 9. Then, using induction, Eq. 9 can be easily shown for all $i, n \in \mathbb{N}$ (beginning with $n = 1$). Crucially, if the residuals were not normally distributed, the population would depart from the normal distribution in the next generation. Equation (9) can also be verified by the law of total variance (with Eqs. 13 and 14).

3.2 Covariance Between Ancestor and Descendant

The regression coefficient describes the covariance between the scores of an ancestor and its descendant, which is given by

$$\mathrm{Cov}(X_{i+n}, X_i) = r^n \sigma_i^2 \ . \tag{11}$$

This expression can be quickly verified from the conditional ancestor (section 3.7) and descendant (section 3.3) distributions, using basic properties of covariance.

3.3 Conditional Descendant Distribution

A general expression can be obtained for the conditional distribution of a descendant's score given its ancestor's score, given by

$$X_{i+n}|X_i \sim \mathcal{N}(\tilde{\mu}_{i+n}, \tilde{\sigma}_{i+n}^2) \ , \tag{12}$$

with

$$\tilde{\mu}_{i+n} = r^n X_i \tag{13}$$

and

$$\tilde{\sigma}_{i+n}^2 = [(r^2 + s^2)^n - r^{2n}]\sigma_i^2 \ . \tag{14}$$

This can be obtained by continually re-applying the one-step downward transition, resulting in

$$X_{i+n}|X_i = r^n X_i + s\sigma_i \sum_{j=1}^{n} r^{n-j}(r^2 + s^2)^{\frac{j-1}{2}} Z_j \tag{15}$$

where the Z_j are independent standard normals. Instead of making the unpleasant summation of the squared coefficients that multiply each Z_j, we use Eq. 9 and the variance of Eq. 15 to obtain $(r^2 + s^2)^n \sigma_i^2 = r^{2n}\sigma_i^2 + \tilde{\sigma}_{i+n}^2$. Rearranging, we obtain Eq. 14. This result is consistent with the following fact, which can be shown by induction to hold for all $a, b \in \mathbb{R}$:

$$(a+b)^n - a^n = b \sum_{j=1}^{n} a^{n-j}(a+b)^{j-1} \ . \tag{16}$$

Alternatively, the conditional descendant distribution (CDD) can be obtained by conditioning on the bivariate distribution, discussed in section 3.7.

3.4 Downward Convergence

The CDD converges to the marginal population distribution of X_{i+n} for large n. This can be shown by considering that for large n, we have $r^n \to 0$ by Eq. 4. Therefore, as $n \to \infty$:

$$\tilde{\mu}_{i+n} = r^n X_i \to 0 \tag{17}$$

and

$$\tilde{\sigma}^2_{i+n} = \sigma^2_{i+n} - (r^n \sigma_i)^2 \to \sigma^2_{i+n} \ , \tag{18}$$

thus

$$X_{i+n}|X_i \xrightarrow{d} X_{i+n} \ . \tag{19}$$

By the polygenic application, this result means that after many generations, a population member's descendants will have scores that become asymptotically indistinguishable from the marginal scores of the population.

3.5 Role of the Coefficients in Downward Convergence

The rate at which the CDD converges to the population distribution is determined by the regression and residual coefficients. From Eqs. 17 and 18, it is clear that the greater the regression towards the mean (*smaller* r), the faster the downward convergence to the marginal distribution, and vice versa. Likewise, Eq. 18 can be re-written as

$$\frac{\tilde{\sigma}^2_{i+n}}{\sigma^2_{i+n}} = 1 - (\frac{r^2}{r^2 + s^2})^n \ , \tag{20}$$

which makes it clear that the lesser the residual variance (*smaller* s), the slower the convergence to the marginal distribution, and vice versa.

3.6 Degenerate Lower Bound for the Residual Coefficient

We can show that $r \geq r^2 + s^2$ is degenerate, resulting in a lower bound on s:

$$s > \sqrt{r(1-r)} \ . \tag{21}$$

The possible values for r and s due to their initial requirements and this newly given lower bound on s are shown in Fig. 1. The degenerate lower bound is shown by breaking up $r \geq r^2 + s^2$ into two cases.

1. In the first case: $r = r^2 + s^2$, then there is no regression towards the mean relative to the population variances, i.e., a descendant has the same expected z-score as its ancestor:
$$\frac{\tilde{\mu}_{i+n}}{\sigma^2_{i+n}} = \frac{X_i}{\sigma^2_i} \ . \tag{22}$$
No regression towards the mean in the z-scores can be considered degenerate. Equations (22 and 23) are obtained by taking the ratio of Eqs. 13 and 9.
2. In the second case: $r > r^2 + s^2$, then we have $r^2 + s^2 < 1$ by Eq. 4 and $\sigma^2_{i+n} \to 0$ for large n, as is shown in section 3.11. Here, the marginal variance converges faster than the conditional expectation does, causing the expected z-score to diverge:
$$\frac{\tilde{\mu}_{i+n}}{\sigma^2_{i+n}} = (\frac{r}{r^2 + s^2})^n \frac{X_i}{\sigma^2_i} \to \infty \ . \tag{23}$$
As X_i was arbitrary, all descendants' expected z-scores would diverge regardless of their ancestors' scores, which can be considered degenerate.

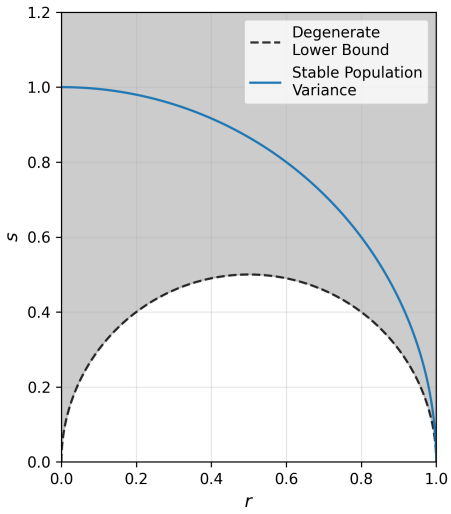

Fig. 1. Possible values for the regression and residual coefficients are given by the shaded area, which extends upwards for $s \to \infty$. The degenerate lower bound has $s > \sqrt{r(1-r)}$ and stable population variance has $s = \sqrt{1-r^2}$.

3.7 Conditional Ancestor Distribution

A general expression can be obtained for the conditional distribution of an ancestor's score given its descendant's score, given by

$$X_{i-n}|X_i \sim \mathcal{N}(\tilde{\mu}_{i-n}, \tilde{\sigma}^2_{i-n}) \ , \tag{24}$$

with

$$\tilde{\mu}_{i-n} = (\frac{r}{r^2+s^2})^n X_i \tag{25}$$

and

$$\tilde{\sigma}^2_{i-n} = [1 - (\frac{r^2}{r^2+s^2})^n]\sigma^2_{i-n} \ . \tag{26}$$

In order to remain consistent with our earlier notation, it will be easiest to think of the generation number of the ancestor as also $b = i - n$. That is, the ancestor existed at time b, which is n generations *before* the current time i. One way to obtain the distribution of $X_{i-n}|X_i$ is with Bayes' rule: $f(x_{i-n}|x_i) \propto f(x_{b+n}|x_b)f(x_b)$. This is similar, though not identical, to finding the posterior distribution for μ of a normal distribution when the prior distribution for μ is normal. Then, a normal prior times the normal likelihood gives a normal posterior. However, because $X_{b+n}|X_b \sim \mathcal{N}(r^n X_b, \tilde{\sigma}^2_{b+n})$, the r^n term obfuscates the conjugate prior.

An alternative approach is to recognize that X_{i-n} and X_i form a bivariate normal distribution. Then, conditioning on X_i, we can use the following standard

results for bivariate normals:

$$\mathrm{E}(X_{i-n}|X_i) = \mathrm{E}(X_{i-n}) + \frac{\mathrm{Cov}(X_{i-n}, X_i)}{\mathrm{Var}(X_i)}[X_i - \mathrm{E}(X_i)] \tag{27}$$

and

$$\mathrm{Var}(X_{i-n}|X_i) = \mathrm{Var}(X_{i-n}) - \frac{[\mathrm{Cov}(X_{i-n}, X_i)]^2}{\mathrm{Var}(X_i)} \ . \tag{28}$$

Plugging in Eqs. 7 - 9 and Eq. 11, we obtain the conditional ancestor distribution (CAD). This approach can also be used to confirm Eq. 12.

3.8 One-step Upward Transition

We can use Eqs. 24 - 26 to obtain the one-step upward transition, i.e., the conditional distribution of a parent score given the child score:

$$X_{i-1}|X_i \sim \mathcal{N}(\tilde{\mu}_{i-1}, \tilde{\sigma}^2_{i-1}) \tag{29}$$

with

$$\tilde{\mu}_{i-1} = \frac{r}{r^2 + s^2} X_i \tag{30}$$

and

$$\tilde{\sigma}^2_{i-1} = \frac{s^2}{r^2 + s^2} \sigma^2_i \ . \tag{31}$$

This has a very similar form to the one-step downward transition (Eq. 2), only the coefficients have changed.

3.9 Upward Convergence

The CAD converges to the marginal population distribution of X_{i-n} for large n. This can be shown by first considering that $r < r^2 + s^2$ by the lower bound (Eq. 21). Therefore, as $n \to \infty$,

$$(\frac{r}{r^2 + s^2})^n \to 0 \ , \tag{32}$$

and by Eq. 25,

$$\tilde{\mu}_{i-n} \to 0 \ . \tag{33}$$

Furthermore, $r^2 < r^2 + s^2$ as $r^2 < r$ by Eq. 4. Therefore, as $n \to \infty$,

$$(\frac{r^2}{r^2 + s^2})^n \to 0 \ , \tag{34}$$

and by Eq. 26,

$$\tilde{\sigma}^2_{i-n} \to \sigma^2_{i-n} \ . \tag{35}$$

In summary, as $n \to \infty$:

$$X_{i-n}|X_i \xrightarrow{d} X_{i-n} \ . \tag{36}$$

This is another way of confirming the lower bound on s (Eq. 21), as the upward conditional expectation would not converge under the degenerate condition.

3.10 Role of the Coefficients in Upward Convergence

As with downward convergence, the rate at which the CAD converges to the population distribution is determined by the regression and residual coefficients in the same general ways. From Eqs. 25 and 26, it is clear that smaller r means faster convergence, and vice versa; and that smaller s means slower convergence, and vice versa. The ratio given by Eq. 20 holds for the CAD (only with $i - n$ rather than $i + n$), which is clear from Eq. 26.

3.11 Stable and Unstable Population Variance

From Eq. 9, it is clear that if $r^2 + s^2 < 1$ then $\sigma^2_{i+n} \to 0$; whereas if $r^2 + s^2 > 1$, then $\sigma^2_{i+n} \to \infty$. These cases are called *unstable population variance*. Such degenerate limits may be prevented if there were some negative feedback between r or s and σ^2_i. However, the simplest resolution is that $r^2 + s^2 = 1$. Then, every generation has the same population variance, called *stable population variance* (SPV). Formally, $\sigma^2_i = 1\ ,\ \forall i \in \mathbb{N}$.

By the polygenic application, for the child generation to have the same population variance as the parent generation (see Eq. 2), indeed for all generations to have the same population variance, it must be the case that $r^2 + s^2 = 1$.

3.12 Stationary Distribution

The stationary distribution of the Markov model can be shown to have a couple of key properties.

1. *SPV is a necessary and sufficient condition for the stationary distribution.* It is clear from section 3.11 and Eq. 7 that only under SPV, is $X_i \sim X_{i+n} \sim \mathcal{N}(0,1)\ ,\ \forall\, i, n \in \mathbb{N}$.
2. *The stationary distribution is reversible.* It will be shown in section 3.7 that the model behaves identically going forwards and backward in time when there is SPV. It is straightforward to show that the conditional ancestor and descendant distributions (Eqs. 12 and 24) are identical under SPV:

$$\tilde{\mu}_{i-n} = \tilde{\mu}_{i+n} = r^n X_i \tag{37}$$

$$\tilde{\sigma}^2_{i-n} = \tilde{\sigma}^2_{i+n} = 1 - r^{2n} \tag{38}$$

3.13 Exponential Functions for Downward Convergence

The role of the coefficients can also be seen in the equations for the rate of change over generations in the parameters of the CDD.

The rate of change in the conditional expectation is negative and increasing with n (positive second partial derivative). That is, the conditional expectation asymptotically approaches zero by

$$\frac{\partial}{\partial n}[\tilde{\mu}_{i+n}] = -\log(\frac{1}{r})\,\tilde{\mu}_{i+n} \tag{39}$$

where 'log' is the natural logarithm. Likewise, the rate of change in the ratio of the conditional variance to the marginal variance is positive and decreasing with n (negative second partial derivative). Therefore, the ratio asymptotically approaches one by

$$\frac{\partial}{\partial n}[\frac{\tilde{\sigma}^2_{i+n}}{\sigma^2_{i+n}}] = \log(1 + \frac{s^2}{r^2})\,[1 - \frac{\tilde{\sigma}^2_{i+n}}{\sigma^2_{i+n}}]\ . \tag{40}$$

Equations (39 and 40) confirm the remarks on r and s made in section 3.5. They also make it clear that we can rewrite Eqs. 13 and 14 as the following exponential functions:

$$\tilde{\mu}_{i+n} = X_i\ \exp[-\log(\frac{1}{r})\,n] \tag{41}$$

and

$$\tilde{\sigma}^2_{i+n} = \sigma^2_{i+n}\,(1 - \exp[-\log(1 + \frac{s^2}{r^2})\,n])\ . \tag{42}$$

These show that the parameters of the CDD exponentially converge on the parameters of the population distributions as the generation gap n widens.

3.14 Exponential Functions for Upward Convergence

As with downward convergence, upward convergence has analogous equations, which similarly confirm the remarks on r and s made in section 3.10. These are given as:

$$\frac{\partial}{\partial n}[\tilde{\mu}_{i-n}] = -\log(\frac{r^2 + s^2}{r})\,\tilde{\mu}_{i-n} \tag{43}$$

$$\frac{\partial}{\partial n}[\frac{\tilde{\sigma}^2_{i-n}}{\sigma^2_{i-n}}] = \log(1 + \frac{s^2}{r^2})\,[1 - \frac{\tilde{\sigma}^2_{i-n}}{\sigma^2_{i-n}}] \tag{44}$$

$$\tilde{\mu}_{i-n} = X_i\ \exp[-\log(\frac{r^2 + s^2}{r})\,n] \tag{45}$$

$$\tilde{\sigma}^2_{i-n} = \sigma^2_{i-n}\,(1 - \exp[-\log(1 + \frac{s^2}{r^2})\,n])\ . \tag{46}$$

Therefore, there is also exponential convergence of the CAD as the generation gap n widens.

4 Mobility and Information-loss

We would like to find some way to quantify the degree of intergenerational *mobility* in the score of a descendant, given the score of its ancestor. This would also correspond to the *loss of information* that could be used to predict a descendant's score, given its ancestor's score, or vice versa.

4.1 Mobility as the Rate of Convergence to the Population Distribution

For the CDD, faster convergence means that the CDD quickly becomes indistinguishable from the population distribution of its generation. The information provided by an ancestor's score that might predict its descendant's score is quickly *lost* over the intervening generation(s). Faster convergence means greater mobility, in that within a small number of generations the CDD moves 'towards' the population distribution (both in expectation and variance). Likewise, for the CAD, faster convergence means that a descendant's score provides less information about its ancestor's score.

4.2 Proposed Mobility Measure

We have previously shown that convergence of the conditional distribution to the marginal distribution (both upward and downward) occurs faster with smaller r and larger s; and occurs slower for larger r and smaller s. Accordingly, a simple measure of mobility and information-loss would be proportional to s and inversely proportional to r.

Definition 1. *Let m be the measure of mobility, or equivalently of information-loss, for the Markov model with parameters r and s, given by*

$$m := \frac{s}{r} \ . \tag{47}$$

In general, a value of m is not unique to a given formulation of the Markov model, because infinitely many values for r and s can have the same ratio. Although it is not investigated here, different Markov models with the same m may turn out to have varied rates of convergence, upwards or downwards, or for expectation or variance. However, for SPV, *every possible Markov model has a unique m.*

Though not discussed here, it is straightforward to show that under the limiting cases for r and s, the desired effects in mobility and information-loss occur, consistent with the proposed measure of mobility. For example, as $r \to 0$, there is rapid convergence both upward and downward consistent with the proposed $m \to \infty$. This can be also shown for $s \to \infty$. Interestingly, in the case where $s \to 0$, then either $r \to 0$ or $r \to 1$ by Eq. 21. In the former, r shrinks faster than s does, causing $m \to \infty$, and the CDD and CAD rapidly converge to the marginals. In the latter combination of r and s, mobility is minimized to the limit of zero, and the behavior of the CDD and CAD are consistent with that.

5 Application for Modeling a Polygenic Trait

The Markov model proposed in this paper has features that should be noted for its application to model a polygenic trait, such as height.

5.1 Assumptions and Limitations

Assumptions of the model:

1. Normal distribution of the polygenic trait in the population, verified for human height [8].
2. Normal linear relationship between parent and child scores, verified for human height [8].
3. The probability of successful reproduction for all members of the population is the same for all scores, i.e., no natural selection takes place.
4. The SD of the conditional child distribution is proportional to the marginal SD of the parent generation (Eq. 5), though this assumption vanishes under SPV, i.e. Eq. 5 holds for all $i \in \mathbb{N}$.

Limitations of the model:

1. Marginal populations exist in discrete time and are non-overlapping.
2. Changes in population size are not described.
3. Only the relationship between a single ancestor-descendant pair is modeled, such as mother and daughter or grandfather and granddaughter. Mating partners and other ancestors are excluded.

5.2 Adjustment for Uniform Environmental Effects

The model described thus far does not describe the average movement of the population mean between generations as a result of external effects that would uniformly shift the conditional mean of each child. This has occurred in human height, which increased between generations in developed countries during the 19th and 20th centuries, an observation that has been attributed to population-wide improvements in the environment for growth, relating to nutrition and health [2, 11].

Such effects can easily be included in the model. Denoting the uniform increase in height between generations by c, we have

$$X_{i+1}|X_i \sim \mathcal{N}(r\,X_i + c,\, s^2\,\sigma_i^2)\ , \tag{48}$$

where we assume μ_i is zero due to standardization of the initial distribution. Then, the only change to the model is in the population mean, which is given by $\mu_{i\pm n} = \mu_i \pm n\,c$, if c is constant between generations. Alternatively c could vary between generations, which is also easy to parametrize.

5.3 Estimation of Parameters

The regression and residual parameters can be estimated with a normal linear model from one-step transition data, which consists of the paired heights of parents and children. When handling this data, the parent and child scores are standardized to the sample mean and SD of the measured parent scores, to

match the form of the model. Then, the vector of coefficients $\beta = (c, r)$ is given by Eq. 48 and is estimated through ordinary least squares. That is, in a plot of parent versus child scores, the estimate for c is the intercept and the estimate for r is the slope of the regression line. Additionally, the unbiased estimate of s is obtained from $\sqrt{RSS/(n-p)}$, where $p = 2$ and RSS is the residual sum of squares.

The regression and residual coefficients provide the SD ratio of the child-generation population to the parent-generation population through the following relation: SD ratio $= \sqrt{r^2 + s^2}$ (see Eq. 9). Therefore, having estimates of any two of the three terms immediately provides an estimate of the third. Furthermore, if SPV can be assumed, then estimating either r or s immediately yields an estimate of the other. It is worth noting that under SPV, r is also the correlation coefficient between parent and child scores (see Eq. 11). Estimates of r and s are important because they parametrize the entire model.

6 Model Verification for Human Height

From two large datasets on the heights of parents and their children, the proposed model is verified through statistical tests for the one-step transition, and the regression and residual coefficients are estimated.

6.1 Requirements for Model Verification

All of the properties of the Markov model follow from the initial condition and the one-step downward transition, confirming their accuracy confirms the model properties. Thus, model verification consists of two parts:

1. The parental heights are modeled by the initial condition.
2. The conditional distribution of the adult child heights is modeled by the one-step downward transition.

6.2 Swedish Dataset

The study *Target Height as Predicted by Parental Heights in a Population-Based Study*, published in *Nature—Pediatric Research* [8], used a normal linear model to predict the heights of adult children from the heights of their parents from a large sample ($n = 2402$) of normal Swedish children born in the 1970s.

The authors did not report on the normality of the parental heights in the dataset but cited a paper that claims height is normally distributed in a population [12]. A uniform increase in height between generations was observed between the parent and child generations, amounting to 0.7 cm for male and 1.0 cm for female subjects.

The authors confirmed the validity of a normal linear model for the one-step transition of parent to child heights through the following tests:

1. The residuals were normally distributed (tested by skewness and kurtosis).

2. The mean residual values were fairly constant over the range of midparental heights, fluctuating around zero, and were only statistically significantly above zero ($p < 0.05$) for very short midparental heights (below -2 SDs).
3. The mean residual values were constant over the range of difference in parental heights.
4. The residual SDs were constant over the range of midparental heights.
5. The residual SDs were constant over the range of difference in parental heights.

Varied values of r and s were estimated by least squares for using mothers' heights to predict sons' heights, fathers' heights to predict sons' heights, mothers' heights to predict daughters' heights, etc. These four possibilities averaged $r = 0.49$ and $s = 0.88$, which essentially produces SPV, producing by Eq. 9 an SD ratio of 1.01. The observed SD ratio was 0.978 for fathers and sons, and 1.16 for mothers and daughters, somewhat consistent with SPV.

6.3 English Dataset

Karl Pearson organized the collection of data on over 1,100 families in England in the 1890s. The dataset he collected contains the heights of mothers, fathers and their adult children with no more than two adult children per family [10].

The model was verified (by section 6.1) for the daughter-mother data ($n = 1375$) with statistical tests at the 5% level. The data was standardized as described in section 5.3. The mother's scores were normally distributed, passing statistical tests of skewness and kurtosis, as well as D'Agostino and Pearson's test, and normality was also clear from a Q-Q plot. The normality of the residuals and of the daughter generation was confirmed by the same methods. As in the Swedish data, an increase in height between generations was observed, amounting to 3.3 cm, therefore the adjustment of section 5.2 was made. The residuals were uncorrelated with the parent scores, as shown in Fig. 2. Therefore, the adjustment constant was valid as the 3.3 cm shift was uniform for all children.

Though not shown here, the same statistical tests were applied to the son-father data ($n = 1078$). All tests passed except the kurtosis and D'Agostino and Person's test for normality of the residuals. The failures occurred because of deviations from normality in the edges (more than two standard deviations from the mean) as apparent from the Q-Q plots.

The Pearson female data had the estimates $r = 0.54$ and $s = 0.96$ (by section 5.3), which nearly produce SPV (SD ratio of 1.10). The female data failed the F-test for equality of variances between the mother and daughter generation ($p < 0.001$). The Pearson male data had the estimates $r = 0.51$ and $s = 0.89$, which nearly produce SPV (SD ratio of 1.03). The male data passed the F-test for equality of variances between the father and son generation ($p = 0.20$).

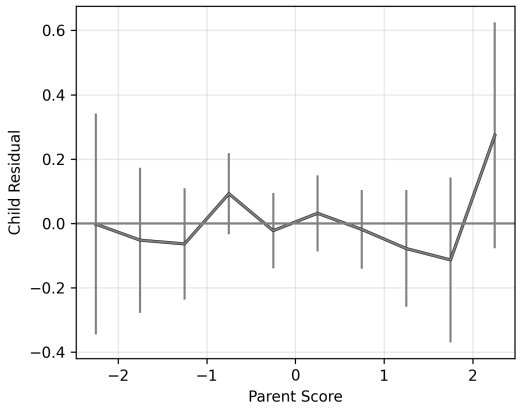

Fig. 2. Mean values of the residuals by parent score, with normal 95% confidence intervals for the Pearson mother-daughter data [10].

6.4 Stable Population Variance for Height

We have confirmed from the Pearson female data both the initial condition and the one-step downward transition, and thus the entire model—as long as there is SPV. This is because, under SPV, the assumption of proportionality between residual and population variance of the previous generation vanishes, as described in assumption 4 of section 5.1.

Overall, SPV seems *a priori* to be a reasonable condition for human height (e.g. continually increasing variance over generations would eventually lead to absurdly large differences in height between the tallest and shortest members of the population). One possible explanation for why the SD ratios were sometimes larger than one is that the adult children tended to be younger than their parents. Therefore, the adult child heights may have continued to regress towards the mean after the study. In the Swedish data, all the adult children had their heights measured at around age eighteen [4]. Similarly, in the Pearson data, all of the adult children were at least age 18 years old, and all of the parents were at most 65 years old [10].

7 Probability Kernels

Probability kernels consist of state-to-set and set-to-set kernels, which can be given conditionally as attributable and destined kernels that form the building blocks of percentile transition matrices.

7.1 State and Set Kernels

Definition 2. *Let the probability kernel $P_n(D, x_i)$ be the n-step probability of reaching the set $D \subseteq \mathbb{R}$ in the descendant population from the state x_i in the*

ancestor population, given by

$$P_n(D, x_i) = \int_{x_{i+n} \in D} f_d(x_{i+n}|x_i) f_m(x_i)\, dx_{i+n} \tag{49}$$

where $f_d(x_{i+n}|x_i)$ is the conditional probability density function (pdf) of the CDD (Eq. 12) and $f_m(x_i)$ is the pdf of the marginal distribution (Eq. 7).

If D is an uninterrupted set (D_{min}, D_{max}), Def. 2 can be simplified to

$$P_n(D, x_i) = f_m(x_i)[\,\Phi(\frac{D_{max} - \tilde{\mu}_{i+n}}{\tilde{\sigma}_{i+n}}) - \Phi(\frac{D_{min} - \tilde{\mu}_{i+n}}{\tilde{\sigma}_{i+n}})\,] \tag{50}$$

where Φ is the standard normal cumulative density function (cdf).

Definition 3. *Let the probability kernel $P_n(D, A)$ be the n-step probability of reaching the set $D \subseteq \mathbb{R}$ in the descendant population from the set $A \subseteq \mathbb{R}$ in the ancestor population, given by*

$$P_n(D, A) = \int_{x_i \in A} P_n(D, x_i)\, dx_i \ . \tag{51}$$

7.2 Kernel Reversibility

Under reversibility, we have

$$f_d(x_{i+n}|x_i) f_m(x_i) = f_d(x_i|x_{i+n}) f_m(x_{i+n}) \tag{52}$$

$\forall i, n \in \mathbb{N}$, and $\forall x_i, x_{i+n} \in \mathbb{R}$, which means by Eqs. 49 and 51 we have

$$P_n(D, A) = P_n(A, D) \tag{53}$$

$\forall A, D \subseteq \mathbb{R}$.

It should also be noted that when A and D are percentile-sets, Eq. 53 holds even under unstable population variance because the values are normalized to the differing variances of the two generations. This results in the persymmetry property of percentile transition matrices.

7.3 Attributable and Destined Kernels

Definition 4. *Let $P_{\alpha,n}(A|D)$ be the conditional probability that a state $X_{i+n} \in D \subseteq \mathbb{R}$ is descendant from (attributable to) a state $X_i \in A \subseteq \mathbb{R}$, given by*

$$P_{\alpha,n}(A|D) = \frac{P_n(D, A)}{P_n(D)} \ , \tag{54}$$

where $P_n(D)$ is the marginal probability of $X_{i+n} \in D$ (see Eq. 7).

Definition 5. *Let $P_{\delta,n}(D|A)$ be the conditional probability that a state $X_i \in A \subseteq \mathbb{R}$ is ascendant from a state $X_{i+n} \in D \subseteq \mathbb{R}$ (i.e. X_{i+n} is destined for D), given by*

$$P_{\delta,n}(D|A) = \frac{P_n(D, A)}{P_i(A)} \ , \tag{55}$$

where $P_i(A)$ is the marginal probability of $X_i \in A$ (see Eq. 7).

8 Percentile Transition Matrices

Percentile transition matrices make interpretable predictions over generations, enabling visualizations of multi-generational regression towards the mean and the effects of the mobility measure.

8.1 Formulation

The calculations of attributable and destined probabilities can be defined by percentile sets. For example, in $P_{\delta,n}(Q_j|Q_k)$, the percentile set Q_k would be converted to a real number set through the cdf $F_i^{-1}(Q_k)$, using the marginal variance σ_i^2. It should be noted that when attributable and destined probabilities are calculated this way, they are not affected by unstable population variance because using percentiles is a way of standardizing to different variances.

Percentile transition matrices result from equal-size, continuous percentile sets Q_1, Q_2, ..., Q_m. For example, for $m = 5$, we have: $Q_1 = (0, 0.2]$, $Q_2 = (0.2, 0.4]$, ..., $Q_5 = (0.8, 1]$. As all percentile sets have the same size, e.g., $P_\cdot(Q_j) = P_\cdot(Q_k) = 0.2$, the calculations can be equivalently interpreted as attributable and destined probabilities. That is, we have

$$P_{\alpha,n}(Q_j|Q_k) = P_{\delta,n}(Q_j|Q_k) = P_{\cdot,n}(Q_j|Q_k) \tag{56}$$

$\forall j, k \in \{1, 2, \ldots, m\}$.

Definition 6. *Let a percentile transition matrix be given by*

$$\mathcal{P}(i, n, m) := \begin{pmatrix} P_{\cdot,n}(Q_m|Q_1) & P_{\cdot,n}(Q_m|Q_2) & \ldots & P_{\cdot,n}(Q_m|Q_m) \\ P_{\cdot,n}(Q_{m-1}|Q_1) & P_{\cdot,n}(Q_{m-1}|Q_2) & \ldots & P_{\cdot,n}(Q_{m-1}|Q_m) \\ \vdots & \vdots & \ddots & \vdots \\ P_{\cdot,n}(Q_1|Q_1) & P_{\cdot,n}(Q_1|Q_2) & \ldots & P_{\cdot,n}(Q_1|Q_m) \end{pmatrix} \tag{57}$$

where $Q_1, Q_2, \ldots, Q_m$ are continuous, equal-size percentile sets of size $1/m$, such that

$$P_u(Q_j) = \frac{1}{m} \tag{58}$$

$\forall j \in \{1, 2, \ldots, m\}$ and $\forall u \in \{i, n\}$, where $P_u(Q_j)$ is the marginal probability of $X_u \in F_u^{-1}(Q_j)$.

8.2 Persymmetry

Percentile transition matrices are persymmetric, i.e., symmetric along the north-east to southwest diagonal. This results from Eq. 53, where if Q_j and Q_k have the same size, we get the persymmetry property

$$P_{\cdot,n}(Q_j|Q_k) = P_{\cdot,n}(Q_k|Q_j) \ . \tag{59}$$

Persymmetry says, for example, that the probability relating a parent in the bottom quintile to a child in the second quintile is the same as that relating a parent in the second quintile to a child in the bottom quintile.

8.3 Symmetry

Percentile transition matrices are symmetric, i.e., symmetric along the northwest to southeast diagonal. This results from the fact that normal distributions are symmetric, which means that

$$f_d(-x_{i+n}|-x_i)f_m(-x_i) = f_d(x_{i+n}|x_i)f_m(x_i) \ . \tag{60}$$

It is also trivially true that

$$f_a(x_i|x_{i+n})f_m(x_{i+n}) = f_d(x_{i+n}|x_i)f_m(x_i) = f(x_{i+n}, x_i) \tag{61}$$

where $f_a(x_{i+n}|x_i)$ is the conditional pdf of the CAD (Eq. 24). Combining Eqs. 60 and 61, we get

$$f_d(x_{i+n}|x_i)f_m(x_i) = f_a(-x_i|-x_{i+n})f_m(-x_{i+n}) \ , \tag{62}$$

which leads immediately to the symmetry property

$$P_{\cdot,n}(Q_j, Q_k) = P_{\cdot,n}(Q_{m-k+1}, Q_{m-j+1}) \tag{63}$$

by Eqs. 49 and 51. Symmetry says, for example, that the probability relating parents in the bottom quintile to children in the fourth quintile is the same as that relating parents in the second quintile to children in the top quintile.

8.4 Bisymmetry and Number of Unique Entries

Percentile transition matrices are bisymmetric, i.e, symmetric along both diagonals, which we just showed. A result of bisymmetry is that the number of unique entries in a percentile transition matrix is $m + (m-2) + ... + 4 + 2$ for even m, and $m + (m-2) + ... + 3 + 1$ for odd m. This is clear by the following argument: filling in the entries by row, the number of non-unique entires in a row is twice the row index (starting at zero). Then, the first row has m unique entries, the second row has $m-2$ unique entries, etc. By the arithmetic series, the number of unique entries is $m(m+2)/4$ for even m and for $(m+1)^2/4$ for odd m.

9 Visualizations of Percentile Transition Matrices

Percentile transition matrices are calculated from the model as well as estimated (through proportions) from observed data, enabling comparisons to the model. The matrices are visualized with the rows indicated by color, and the sizes proportional to the probabilities in Figs. 3, 4, 5, and 6. The integral in Eq. 51 cannot be solved analytically, and instead is calculated numerically to very high fidelity (see section 11.2).

The calculated *quintile*-transition matrices ($m = 5$) have 9 unique entries (section 8.4). One way to understand these figures is to consider an entry as a proportion of its column sum (probability destined) or row sum (probability attributable). Because of the symmetry property, the calculated matrices can equivalently be re-written with the x-axis as *Descendant's Quintile*, and the y-axis as *Cumulative Probability of Ancestor's Quintile*.

9.1 Comparison to Pearson Data

The observed percentile transition matrices of the Pearson mother-daughter and father-son data are shown along with a simulation from the estimated parameters in Fig. 3. A reasonable similarity is observed between the calculated and observed matrices. It should be noted that for the Pearson data, there was a relatively small number of observations for each entry of the quintile-quintile matrices (minimum of 10, average of 49, and maximum of 149). This may explain some of the deviations between the male and female matrices, and potentially, the deviations from the calculated matrix.

The rightmost column in Fig. 3 shows that for height, a parent in the top quintile has about a 43% chance of having a child also in (destined for) the top quintile, a 25% chance of having a child in the fourth quintile, and so on. Equivalently, by Eq. 56, these numbers can be interpreted as follows: about 43% of the children in the top quintile are attributable to a parent also in the top quintile, about 25% of the children in the fourth quintile are attributable to a parent in the top quintile, and so on.

9.2 Multigenerational Regression Towards the Mean

Between the Swedish and English datasets, the estimates for r were about 0.5, with approximate SPV. Using $r = 0.5$ and SPV ($s \approx 0.866$), quintile transition matrices are calculated, showing in Fig. 4 the regression towards the mean that takes place over multiple generations. The exponential downward convergence can be seen in the asymptotic approach of the probabilities towards uniform 20%, with slowing rates of convergence as the generation-gap widens.

9.3 Visualization of the Mobility Measure

The effect of the mobility measure (Def. 1) is visualized in quintile transition matrices over varying m in Fig. 5 under SPV, in which m uniquely determines r and s.

9.4 Comparison to US Family Income Mobility

A quintile transition matrix relating the family incomes of 9,867,736 US children born between 1980-82 and their parents was calculated from income tax data by Chetty et al. and is reproduced in Fig. 6 [3]. This transition matrix is crucially different than the other ones shown in this paper because it relates measures of the *families* of children and their parents, rather than relating measures of an individual child to one of its parents (of the same sex). Nonetheless, reasons for similarity can be explained by the fact that the distribution of income can be approximated by a log-normal distribution [1, 9]. Then, if the relation between the (log) income of parents and their children can be roughly approximated by a normal linear model, the Markov model described in this paper could reasonably be applied.

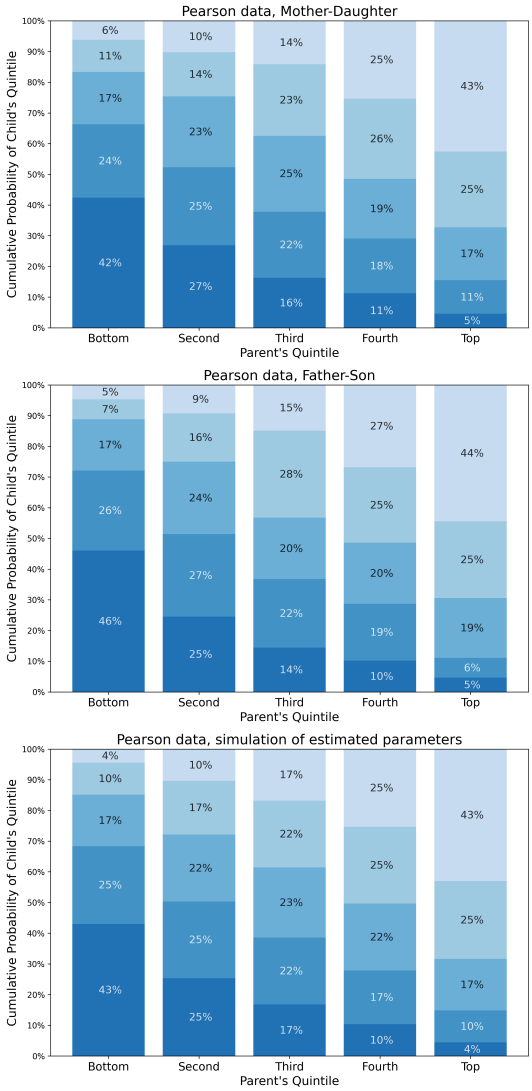

Fig. 3. Observed and calculated quintile transition matrices from the Pearson data, with estimated parameters $r = 0.54$, $s = 0.96$ for female and $r = 0.51$, $s = 0.89$ for male [10]. These parameters result in identical calculated matrices (to the first decimal place), therefore only one is shown.

Using the parent's family income to predict the child's income might be expected to increase the correspondence (and reduce mobility). On the other hand, predicting the child's family income might be expected to reduce the correspondence (and increase mobility), due to regression towards the mean. These effects

are generally unknown, thus the caveat must be made that the transition matrix in Fig. 6 cannot be equivalently compared to the other transition matrices shown in this paper. Nonetheless, it can be generally remarked that there is a greater degree of observed mobility in family income (Fig. 6) than in height (Fig. 3). The mobility measure for income appears to be best approximated by $m \approx 3$ (see Fig. 5). This can be compared with the mobility in height of about $m \approx$ 1.7-1.8, obtained from the estimates of the regression and residual coefficients.

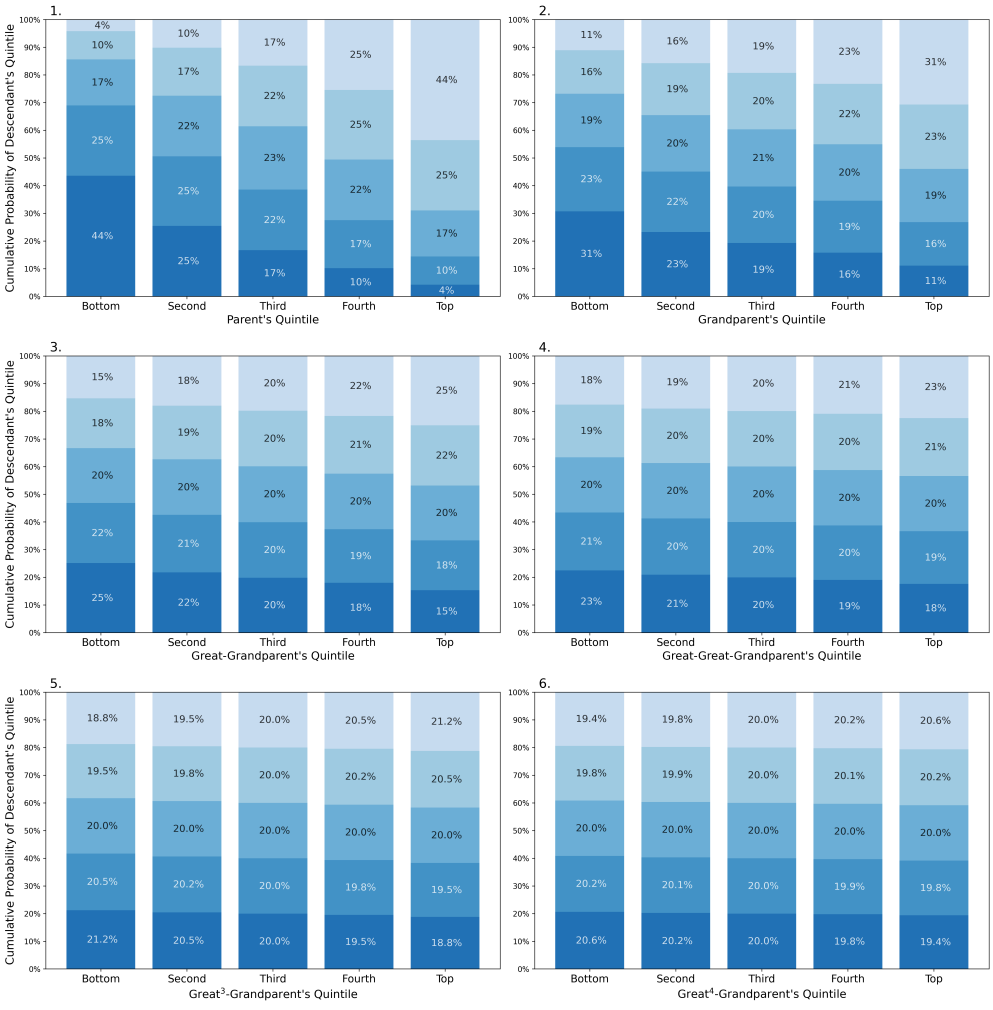

Fig. 4. Quintile transition matrices over generation-gaps n from 1 through 6, with SPV at $r = 0.5$ ($s \approx 0.866$).

The heritability of height has been well established [8, 12, 15], though much of the correspondence between the incomes of parents and their children may be due to environmental factors. Nonetheless, studies of genome-wide SNPs have identified polygenic scores that account for a small amount ($\sim 2.5\%$) of the variance in family socioeconomic status (SES), including income [14, 5]. A normal linear model (and the corresponding Markov model introduced here) may be useful to describe both the genetic *and* environmental factors that relate the incomes of parents and their children.

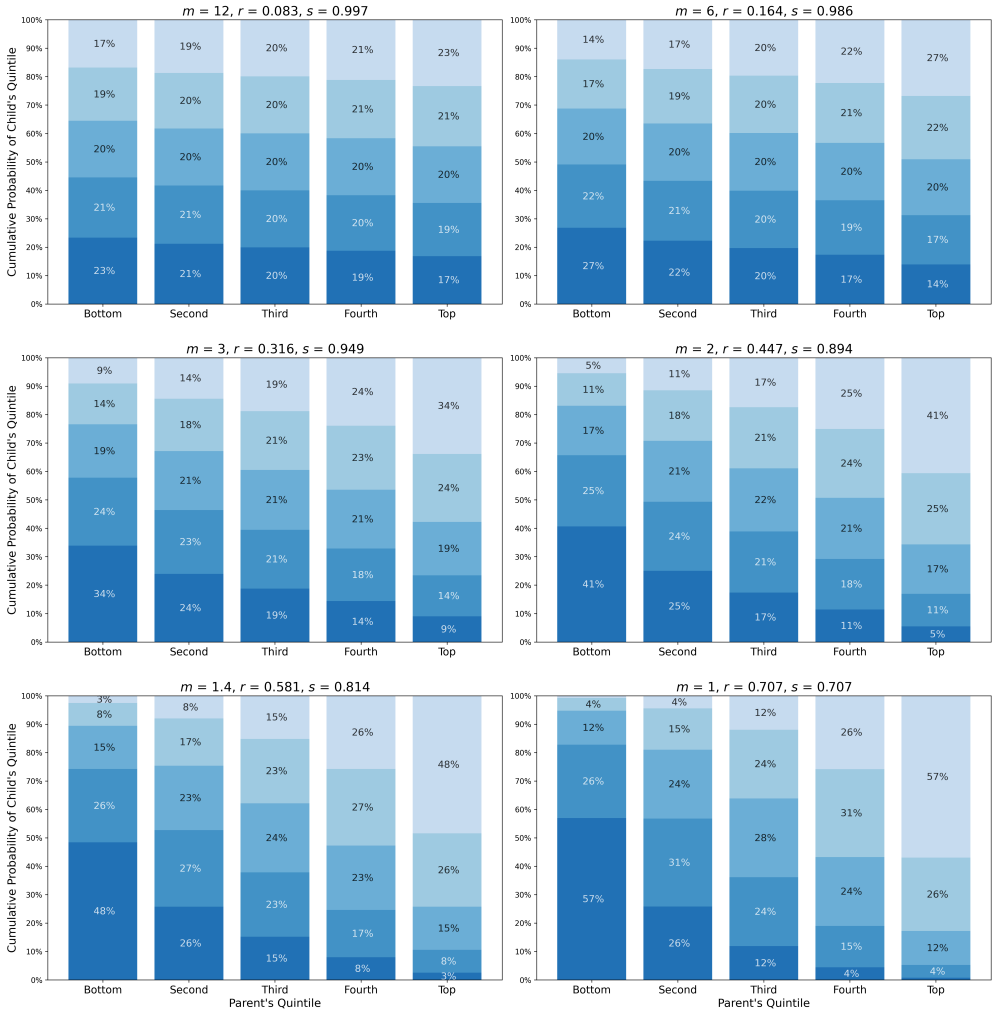

Fig. 5. Parent-to-child quintile transition matrices, over decreasing levels of mobility, with SPV.

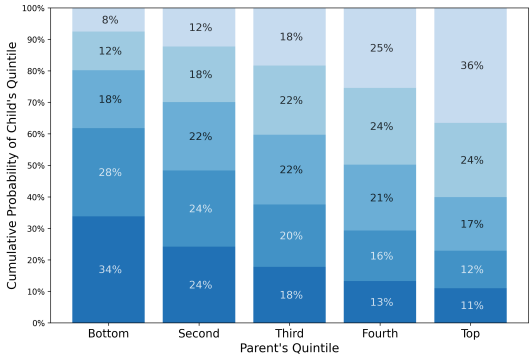

Fig. 6. Parent-to-child quintile transition matrix for US family income data for the 1980-82 birth cohort [3].

10 Discussion

The model proposed in this paper extends the normal linear model (linear regression with normally distributed residuals) into a Markov model. The initial state is a normal distribution, and all states after that are also normally distributed. It has immediate application for a polygenic trait such as height, which is normally distributed, and for which normal linear models have are an established method of predicting adult child height from parent height [8].

10.1 Model Properties

Under a linear model, the conditional expectation for a child's score is the weighted average of the parent's score and the mean population score, where the weight of the parent's score is given by the regression coefficient r. Around this expectation, or prediction, there is random normal variation with an SD that is assumed in the model to be proportional to the marginal SD of the parent-generation population, where the degree of proportionality is given by the residual coefficient s.

This simple formulation relating parents and their children is re-applied through induction as generations reproduce to produce successive generations. The result is a Markov model in discrete time (generations) and continuous space (phenotypic score). From this model, conditional normal distributions are derived for the scores of descendants and ancestors of any generation-gap n apart. Furthermore, the conditional distribution of an ancestor or descendant's score follows an exponential function of convergence to population distributions as n increases, with rate constants determined by r and s. That is, as the distance to a descendant or ancestor increases, prediction-power (information) is lost as an exponential function of the generation-gap.

The simple formulation is also consistent with the overall normal distribution of the polygenic trait in the population. Assuming the initial generation is normally distributed, all future and past generations are also normally distributed. To obtain the marginal distribution of the child's generation, the RV representing a parent—scaled by the regression coefficient, is combined additively with the standard normally distributed RV representing the variation about the prediction—scaled by the residual coefficient. The sum of independent normally distributed RVs is a normal RV with expectation equal to the sum of the expectations and variance equal to the sum of the variances. An immediate result is therefore that the population variance of the child-generation is proportional to that of the parent-generation by r^2+s^2. This means that for there to be constant population variance between generations (SPV), $r^2 + s^2$ must equal one.

SPV results in the stationary distribution. That is, the conditional descendant and ancestor distributions are identical for the same generation-gap n. Equivalently, the system behaves the same with time going forwards or backward. When scores are indexed by percentile (relative to the population of their generation) the Markov chain is stationary and reversible in all cases (for both stable and unstable population variance). This is because indexing by percentile standardizes the differing population variances of the generations. Reversibility means that predicting a child's or grandchild's score has the same form as predicting a parent's or grandparent's score (the CDD and CAD are the same).

10.2 Intergenerational Edge Persistence

Consider those with scores on the edge of the population distribution, where the edge boundary is defined by some threshold. What proportion of these members (e.g. the tallest members of the population) are the children of parents also above the threshold (e.g. the tallest members of the last generation)? There are two counteracting effects: probability and number. Parents above the threshold have a higher probability of having children also above the threshold. However, there are many parents from the rest of the distribution whose children can end up above the threshold. The amount of edge persistence thus results from a competition of sorts between the greater probability for the tall parents and the greater number of shorter parents. Intuitively, when mobility (m) is greater, there is less edge persistence and vice versa. These questions can be answered for quintiles from the visualizations in section 9, and the software is available to examine other thresholds (see section 11.2).

10.3 Potential to Model Genetic and Environmental Factors

The observed approximation of the log-normal distribution to income (Figs. 5 and 6) suggests an application of the model to this domain. That is, a person's income could result from the multiplicative combination of many factors that correlate between parents and children. Then, a person's log-income would be normally distributed due to the central limit theorem, as with height, and his or her adult child's log-income could be predicted with a normal linear model.

The example of US family income introduced in the section on visualizations of percentile transition matrices suggests that a normal linear model may be useful to describe the relation between parents and their children for a trait with genetic or environmental determinants, or a combination thereof. This is especially reasonable to expect if the effects of the environment are also numerous and, on average, small, as are the effects of genes for a polygenic trait. Then, the trait should be normally distributed in the population and might fall under the paradigm of the linear model, in which conditional expectation for a child's score is a weighted average (by r) of the parent's score and the mean population score.

10.4 Future Work

In future work, the Markov model described in this paper could model traits that are approximately normally distributed and for which child scores can be predicted from parent scores with a normal linear model. These traits could have some arbitrary combination of genetic or environmental determinants, as long as they meet those basic two requirements. For these traits, this paper introduces a vocabulary of sorts: the regression coefficient r and the residual coefficient s. Estimates of the mobility measure could enable a standard comparison between the mobility of these traits, such as between height and income as discussed here.

Additional future work could apply the multi-generational extension of the normal linear model shown in this paper to other generalized linear models. This extension was justified for describing a population that reproduces between generations, in which the forward one-step transition between generations is well-modeled by a normal linear relationship. However, other general linear models, such as the Poisson or the Gamma, might also be extended to a Markov model, with potential applications as well. Lastly, this paper described a univariate polygenic trait. It may be interesting to explore a multivariate normal linear Markov model, which could model a set of related polygenic traits with non-zero covariances.

11 Appendix

11.1 Abbreviations

The abbreviations used are: SPV, stable population variance; CDD, conditional descendant distribution; CAD, conditional ancestor distribution; RV, random variable; pdf, probability density function; cdf, cumulative density function.

11.2 Software and Statistical Tests

The software for calculating and estimating the probability transition matrices (section 9) and the full analysis of the English data (section 6.3) are both available with extensive documentation at https://github.com/jessebmurray/polygenic. The results from these sections can be fully reproduced using this repository.

References

1. 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.
2. Bogin, B. and Rios, L. (2003). Rapid morphological change in living humans: implications for modern human origins. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, 136(1):71—84.
3. 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.
4. Karlberg, J. and Albertsson-Wikland, K. (1995). Growth in full- term small-for-gestational-age infants: From birth to final height. *Pediatric Research*, 38:733–739.
5. 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.
6. Lange, K. (1997a). An approximate model of polygenic inheritance. *Genetics*, 147(3):1423–1430.
7. Lange, K. (1997b). *The Polygenic Model. In: Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health.* Springer, New York, NY.
8. 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.
9. Neal, D. and Rosen, S. (1998). Theories of the distribution of labor earnings. Working Paper 6378, National Bureau of Economic Research.
10. Pearson, K. and Lee, A. (1903). On the laws of inheritance in man: I. inheritance of physical characters. *Biometrika*, 2(4):357–462.
11. Perkins, J. M., Subramanian, S. V., Davey Smith, G., and Özaltin, E. (2016). Adult height, nutrition, and population health. *Nutrition reviews*, 74(3):149–165.
12. Preece, M. A. (1996). The genetic contribution to height. *Hormone Research in Pediatrics*, 45(Suppl. 2):56–58.
13. Rieger, R., Michaelis, A., and Green, M. (1968). *A Glossary of Genetics and Cytogenetics.* Springer, New York, NY.
14. 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.
15. 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.