Polygenic Markov Chain

Jesse Murray

October 13, 2020

1 Introduction

A simple Markov chain model is proposed that seeks to describe the movement of polygenic scores between familial generations. Some potentially useful properties are shown to result from this model. The model is compared with the data on the heights of 898 adult children and their parents, as collected by Francis Galton.

2 Model

2.1 Conceptual basis for polygenic analogy

Polygenic traits are determined by many genes (hundreds or thousands), and their observed phenotypic scores are generally normally distributed [Lange, 1997a, Lange, 1997b]. This is thought to result from the central limit theorem – as the phenotype of a polygenic trait results from the additive sum of many gene-effects, which are akin to independent random draws from an arbitrary distribution [Rieger et al., 1968].

2.2 Markov chain

The proposed Markov chain exists in discrete time $i \in \{0, 1, 2, ...\}$, where i is the familial generation (parent, offspring, etc.); and continuous space $X_i \in \mathbb{R}$, where X_i is the phenotypic score of the analogized one-dimensional polygenic trait. The terms *score* and *state* are used interchangeably.

Convention Let Z indicate a standard normal $Z \sim \mathcal{N}(0, 1)$. All Z are taken to be independent unless specified otherwise.

A complete description of the chain can be given in two parts: (1.) the initial condition and (2.) the one-step transition.

1. The initial score is drawn from a standard normal distribution: $X_0 \sim \mathcal{N}(0, 1)$. A normally distributed one-dimensional polygenic trait can be standardized to meet this condition.

$$X_0 = Z$$

2. The conditional one-step transition, i.e., the score of an offspring, given the score of its parent, is also drawn from a normal distribution: $X_{i+1}|X_i \sim \mathcal{N}(\tilde{\mu}_{i+1}, \tilde{\sigma}^2_{i+1})$.

$$X_{i+1}|X_i = \tilde{\mu}_{i+1} + \tilde{\sigma}_{i+1} Z$$

   (a) An offspring's score is expected to be similar to that of its parent, scaled by r – the *expectation regression coefficient*. There is no strict requirement on r, other than that it is a real number. However, in order for there to be – on average – regression towards the population mean, $0 < r < 1$.

$$\mathrm{E}(X_{i+1}|X_i) = \tilde{\mu}_{i+1} = rX_i$$

   (b) An offspring's score has a standard deviation that is similar to the marginal standard deviation of the parent generation, scaled by r_s – the *standard deviation scaling coefficient*. For standard deviation to be positive and degenerate normal distributions avoided, we have $r_s > 0$.

$$\sqrt{\mathrm{Var}(X_{i+1}|X_i)} = \tilde{\sigma}_{i+1} = r_s \sigma_i$$

Combining these two parts we have the initial condition:

$$X_0 = Z$$

And the one-step transition:

$$X_{i+1}|X_i = rX_i + r_s\sigma_i Z$$

These two equations give a complete description of the Markov chain and all of its properties can be derived from them.

2.3 Marginal random state

It can be shown by induction and the theorem of the sum of independent normal random variables (rvs) that $X_n \sim \mathcal{N}(0, \sigma_n^2)$:

$$X_n = \sigma_n Z$$

Convention Let σ_i^2 – the *population variance* – refer to the marginal variance of X_i, i.e., not given any previous score X_j with $j < i$. The population variance is shown to have a one-step relationship with the population variance of the previous generation.

Then, for Z_α, Z_β iid standard normals, the marginal random state X_{i+1} can be written in terms of the population variance of the previous state.

$$X_{i+1} = r\sigma_i Z_\alpha + r_s\sigma_i Z_\beta$$

2.3.1 Population variance

From the equation for the marginal random state X_{i+1}, it can be shown that σ_{i+1}^2 has the following one-step relationship with the previous population variance:

$$\sigma_{i+1}^2 = (r^2 + r_s^2)\sigma_i^2$$

By induction, the population variance can be computed from a previous population variance for an arbitrary step-length:

$$\sigma_{i+n}^2 = (r^2 + r_s^2)^n\sigma_i^2$$

Setting $i = 0$:

$$\sigma_n^2 = (r^2 + r_s^2)^n$$

2.4 Conditional transition of an arbitrary step-length

The conditional distribution of X_{i+n} given X_i can be shown to be normally distributed: $X_{i+n}|X_i \sim \mathcal{N}(\tilde{\mu}_{i+n}, \tilde{\sigma}_{i+n})$.

With the following conditional expectation:

$$\mathrm{E}(X_{i+n}|X_i) = \tilde{\mu}_{i+n} = r^n X_i$$

And the following conditional standard deviation :

$$\sqrt{\mathrm{Var}(X_{i+n}|X_i)} = \tilde{\sigma}_{i+n} = \sigma_i\sqrt{(r^2 + r_s^2)^n - r^{2n}}$$

The equation for conditional standard deviation results from consecutively re-applying the one-step transition, which leads to the following equation for $X_{i+n}|X_i$:

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

It might appear that to obtain $\tilde{\sigma}_{i+n}^2$, an unpleasant summation needs to made of the squared coefficients that multiply each Z_j. However, the earlier equation for population variance can be deployed, in which $r^n X_i$ is treated as a random variable instead of a constant. Therefore, the population variance is equal to the sum of marginal variance of $r^n X_i$ and $\tilde{\sigma}_{i+n}^2$.

$$\sigma_{i+n}^2 = (r^2 + r_s^2)^n \sigma_i^2 = r^{2n} \sigma_i^2 + \tilde{\sigma}_{i+n}$$

After rearranging and taking the square root, we obtain the equation for $\tilde{\sigma}_{i+n}$.

This is – coincidentally – a proof of the following fact, which can be shown by induction to hold any real numbers a, b:

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

2.5 One-step dependency

The expectation regression coefficient describes the covariance and correlation between parent and offspring scores.

$$\mathrm{Cov}(X_{i+1}, X_i) = r\sigma_i^2$$

$$\mathrm{Corr}(X_{i+1}, X_i) = r \frac{\sigma_i}{\sigma_{i+1}}$$

3 Population variance and the stationary distribution

3.1 Unstable population variance

We have shown that:

$$\sigma_{i+n}^2 = (r^2 + r_s^2)^n \sigma_i^2$$

Taking the limit $n \to \infty$:

- For $r^2 + r_s^2 < 1$ we have $\sigma_{i+n}^2 \to 0$.
- For $r^2 + r_s^2 > 1$, we have $\sigma_{i+n}^2 \to \infty$.

These degenerate limits could be prevented if r or r_s were inversely proportional to some power of the population variance σ_i^2. However, this negative feedback is not explored here. Instead, we discuss the simplest resolution to population variance instability, which is that $r^2 + r_s^2 = 1$. Then, every generation has the same population variance.

3.2 Stable population variance

We are particularly interested in the case where the population variance remains constant between generations. This would model a population in which the phenotypes of the polygenic trait do not become increasingly or decreasingly spread out between successive generations.

$$\sigma_{i+1}^2 = \sigma_i^2$$

This occurs if and only if:

$$r^2 + r_s^2 = 1$$

By induction, the population variance of any arbitrary generation i is equal to the initial variance.

$$\sigma_i^2 = 1$$

3.2.1 Stable population variance implies stationary distribution

It can be shown that the chain is at a stationary distribution when there is stable population variance:

$$X_i \sim \mathcal{N}(0, 1)$$

$$X_{i+1} = rZ_\alpha + r_s Z_\beta$$

$$X_{i+1} \sim \mathcal{N}(0, r^2 + r_s^2)$$

$$X_{i+1} \sim \mathcal{N}(0, 1)$$

3.2.2 Convergence of the conditional transition to the stationary distribution

The n-step conditional transition $X_{i+n}|X_i \sim \mathcal{N}(\tilde{\mu}_{i+n}, \tilde{\sigma}_{i+n})$ converges to the marginal population distribution for large n.

From stable population variance, we have that $r^2 + r_s^2 = 1$; and by definition, we have that $r_s > 0$. Together, these imply $|r| < 1$. Then by the ratio test, we have the following convergence:

$$\lim_{n\to\infty} r^n \to 0$$

As a result, the expectation and variance of the conditional normal distribution converge to the expectation and variance of the stationary population.

$$\lim_{n\to\infty} \mathrm{E}(X_{i+n}|X_i) = r^n X_i \to 0$$

$$\lim_{n\to\infty} \mathrm{Var}(X_{i+n}|X_i) = (r^2 + r_s^2)^n - r^{2n} \to (r^2 + r_s^2)^n = 1$$

Then as $n \to \infty$:

$$X_{i+n}|X_i \to \mathcal{N}(0, 1)$$

By the polygenic analogy, this result says that after many generations, a population member's descendants will have scores that become asymptotically indistinguishable from the scores of the population at large. Regardless of how average or extreme the ancestor's score was, it eventually becomes irrelevant.

3.2.3 Stationary distribution use case

Under stable population variance, observing or measuring one of r or r_s immediately gives away the other. As an example, r could be obtained by the correlation coefficient between parent and offspring.

$$\mathrm{Corr}(X_{i+1}, X_i) = r$$

Alternatively, r could be obtained by ordinary least squares, as described later.

When there is stable population variance, knowledge of either r or r_s provides sufficient information to model the population with the proposed Markov chain model.

4 Transition kernel

Let A be a subset of the state space:

$$A \subseteq \mathbb{R}$$

4.1 State to set transition

Then the transition kernel $P(A, x_i)$ gives the one-step probability of reaching the set A from the state x_i.

$$P(A, x_i) = \int_{x_{i+1} \in A} f(x_{i+1}|x_i) f(x_i)\, dx_{i+1}$$

Where $f(x_{i+1}|x_i)$ is the conditional probability density function (pdf) of $X_{i+1}|X_i \sim \mathcal{N}(\tilde{\mu}_{i+1}, \tilde{\sigma}^2_{i+1})$, and $f(x_i)$ is the pdf of $X_i \sim \mathcal{N}(0, \sigma_i^2)$.

4.2 Set to set transition

A similar transition kernel $P(A, B)$ can be used to obtain the one-step probability of reaching the set A from the set B.

$$B \subseteq \mathbb{R}$$

$$P(A, B) = \int_{x_i \in B} P(A, x_i)\, dx_i$$

4.3 Probability attributable

Define the probability that a current state X_{i+1} in set A resulted from or is 'attributable' to a previous state or parent score X_i in set B.

$$P_\alpha(A, B) = \frac{P(A, B)}{P(A, \mathbb{R})}$$

By the law of total probability, $P(A, \mathbb{R})$ is the marginal probability that the state X_{i+1} is in the space A, which is given by $P(A)$:

$$P(A) = \int_{x_{i+1} \in A} f(x_{i+1})\, dx_{i+1}$$

Therefore:

$$P_\alpha(A, B) = \frac{P(A, B)}{P(A)}$$

4.4 Probability destined

Define the probability that a previous state X_i in set B will result in or is 'destined' for a current state or offspring score X_{i+1} in set A.

$$P_\delta(A, B) = \frac{P(A, B)}{P(\mathbb{R}, B)}$$

Because the integral over the entire support of a pdf equals 1, $P(\mathbb{R}, B)$ is the marginal probability that the state X_i is in the space B, which is given by $P(B)$:

$$P(B) = \int_{x_{i+1} \in B} f(x_i)\, dx_i$$

Therefore:

$$P_\delta(A, B) = \frac{P(A, B)}{P(B)}$$

5 Linear regression of the one-step transition

We have that:

$$\mathrm{E}(X_{i+1}|X_i) = rX_i, \qquad X_{i+1} = rX_i + \epsilon$$

This has the same form as the linear regression model where b can be estimated by minimising the sum of the squared errors.

$$\mathrm{E}(Y|X) = bX, \qquad Y = bX + \epsilon, \qquad \mathrm{E}(\epsilon) = 0$$

This means that r can be estimated from existing one-step transition data through the least-squares method.

6 Reverse one-step transition

The parent's phenotypic score is also a rv that is dependent on the offspring's phenotypic score.

$$X_i = \frac{1}{r}X_{i+1} + \frac{r_s}{r}\sigma_i Z$$

$$X_i|X_{i+1} \sim \mathcal{N}(\frac{1}{r}X_{i+1}, \frac{r_s^2}{r^2}\sigma_i^2)$$

Where the variance can be rewritten in terms of the nth generation's variance:

$$\mathrm{Var}(X_i|X_{i+1}) = \frac{r_s^2}{r^2(r^2+r_s^2)}\sigma_{i+1}^2$$

7 Eve's law

It is possible to compute $\mathrm{Var(X_{i+1})}$ through Eve's law:

$$\mathrm{Var(X_{i+1})} = \mathrm{E}(\mathrm{Var}(X_i|X_{i+1})) + \mathrm{Var}(\mathrm{E}(X_{i+1}|X_i))$$

We have that:

$$X_{i+1} = rX_i + \tilde{\sigma}_{i+1}Z$$

$$\tilde{\sigma}_{i+1} = r_s\sigma_i$$

Therefore, each term in Eve's law is:

$$\mathrm{E}(\mathrm{Var}(X_{i+1}|X_i)) = \mathrm{E}(r_s^2\sigma_i^2) = r_s^2\sigma_i^2$$

$$\mathrm{Var}(\mathrm{E}(X_{i+1}|X_i)) = \mathrm{Var}(rX_i) = r^2\sigma_i^2$$

Combining these terms, we confirm the variance of X_{i+1}:

$$\sigma_{i+1}^2 = (r^2 + r_s^2)\sigma_i^2$$

8 Use cases

With the two parameters r and r_s, we can perfectly describe the variance of the offspring's generation relative to that of the parent's generation through the following relation:

$$\frac{\sigma_{i+1}^2}{\sigma_i^2} = r^2 + r_s^2$$

Simply knowing or having measured two of the three values: the ratio parent generation and offspring generation variance, the expectation regression coefficient (r), the standard deviation scaling coefficient (r_s); it is possible to obtain the third.

This can be useful, for example, to obtain the standard deviation of the offspring's polygenic score, after having measured the parent's polygenic score: $\tilde{\sigma}_{i+1} = r_s \sigma_i$.

Alternatively, after having only measured the $\frac{\sigma^2_{i+1}}{\sigma^2_i}$, which is 1 when there is stable population variance, and r, which can be obtained through ordinary least squares with the parent and offspring data or by the correlation coefficient between parent and offspring when there is stable population variance; the model can be constructed and applied.

Appendix

```
Put your R code here.
```

References

[Lamport, 1994] Lamport, L. (1994). *LaTeX: A Document Preparation System*. Addison Wesley, Reading, Massachusetts, 2nd edition.

[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.

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