

Polygenic Markov Chain

Jesse Murray

October 11, 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 generally normally distributed [Lange, 1997a, Lange, 1997b]. Their normal distribution is theorized to result from the central limit theorem - as the phenotype of a polygenic trait (referred to as the score) 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 $\{0, 1, 2, \dots\}$, representing familial generations (parent, offspring, etc.); and continuous space \mathbb{R} , representing a continuous one-dimensional polygenic score.

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

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

1. The initial score is drawn from a normal distribution: $X_0 \sim \mathcal{N}(0, \sigma_0^2)$.

$$X_0 = \sigma_0 Z$$

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: $X_{n+1}|X_n \sim \mathcal{N}(\tilde{\mu}_{n+1}, \tilde{\sigma}_{n+1}^2)$.

$$X_{n+1}|X_n = \tilde{\mu}_{n+1} + \tilde{\sigma}_n Z$$

- (a) An offspring's score is expected to be similar to that of its parent, although scaled by r - the *expectation regression coefficient*. Typically, $0 < r < 1$, such that there is - on average - regression towards the population mean.

$$\mathbb{E}(X_{n+1}|X_n) = \tilde{\mu}_{n+1} = r X_n$$

- (b) An offspring's score has a standard deviation that is similar to the standard deviation of the parent generation's population, although scaled by r_s - the *standard deviation scaling coefficient*.

$$\sqrt{\text{Var}(X_{n+1}|X_n)} = \tilde{\sigma}_{n+1} = r_s \sigma_n$$

2.3 Further one-step properties

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

$$\text{Cov}(X_{n+1}, X_n) = r\sigma_n^2$$

$$\text{Corr}(X_{n+1}, X_n) = r \frac{\sigma_n}{\sigma_{n+1}}$$

2.4 Arbitrary step-length properties

$$E(X_{n+j}|X_n) = \tilde{\mu}_{n+j} = r^j X_n$$

$$\sqrt{\text{Var}(X_{n+j}|X_n)} = \tilde{\sigma}_{n+j} = \sigma_n \sqrt{(r^2 + r_s^2)^j - r^{2j}}$$

For a stable population variance, the expression under the square root asymptotically approaches 1.

2.5 The marginal random state

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

$$X_n = \sigma_n Z$$

Then, for Z_a, Z_b iid standard normal rvs:

$$X_{n+1} = r\sigma_n Z_a + r_s\sigma_n Z_b$$

2.6 Marginal variance

It follows immediately that σ_{n+1}^2 has the following one-step relationship:

$$\sigma_{n+1}^2 = (r^2 + r_s^2)\sigma_n^2$$

By induction, σ_n^2 can also be stated in terms of the initial variance:

$$\sigma_n^2 = (r^2 + r_s^2)^n \sigma_0^2$$

3 Stable population variance

Let a stable population variance be defined as follows:

$$\sigma_{n+1}^2 = \sigma_n^2$$

The following are important properties that occur under stable population variance:

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

$$\text{Corr}(X_{n+1}, X_n) = r$$

For an arbitrary state $i \geq 0$:

$$\sigma_i^2 = \sigma_0^2$$

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 $P(A, x_n)$ gives the one-step probability of reaching the set A from the state x_n .

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

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

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 .

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

$$P(A, B) = \int_{x_n \in B} P(A, x_n) dx_n$$

4.3 Probability attributable

Define the probability that a current state X_{n+1} in set A resulted from or is 'attributable' to a previous state or parent score X_n 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_{n+1} is in the space A , which is given by $P(A)$:

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

Therefore:

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

4.4 Probability destined

Define the probability that a previous state X_n in set B will result in or is 'destined' for a current state or offspring score X_{n+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_n is in the space B , which is given by $P(B)$:

$$P(B) = \int_{x_n \in B} f(x_n) dx_n$$

Therefore:

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

5 Linear regression

We have that:

$$E(X_{n+1}|X_n) = rX_n, \quad X_{n+1} = rX_n + \epsilon$$

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

$$E(Y|X) = bX, \quad Y = bX + \epsilon, \quad E(\epsilon) = 0$$

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.

$$X_n = \frac{1}{r}X_{n+1} + \frac{r_s}{r}\sigma_n Z$$

$$X_n|X_{n+1} \sim \mathcal{N}\left(\frac{1}{r}X_{n+1}, \frac{r_s^2}{r^2}\sigma_n^2\right)$$

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

$$\text{Var}(X_n|X_{n+1}) = \frac{r_s^2}{r^2(r^2 + r_s^2)}\sigma_{n+1}^2$$

7 Eve's law

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

$$\text{Var}(X_{n+1}) = \text{E}(\text{Var}(X_n|X_{n+1})) + \text{Var}(\text{E}(X_{n+1}|X_n))$$

We have that:

$$X_{n+1} = rX_n + \tilde{\sigma}_{n+1}Z$$

$$\tilde{\sigma}_{n+1} = r_s\sigma_n$$

Therefore, each term in Eve's law is:

$$\text{E}(\text{Var}(X_{n+1}|X_n)) = \text{E}(r_s^2\sigma_n^2) = r_s^2\sigma_n^2$$

$$\text{Var}(\text{E}(X_{n+1}|X_n)) = \text{Var}(rX_n) = r^2\sigma_n^2$$

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

$$\sigma_{n+1}^2 = (r^2 + r_s^2)\sigma_n^2$$

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:

$$\frac{\sigma_{n+1}^2}{\sigma_n^2} = r^2 + r_s^2$$

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: $\tilde{\sigma}_{n+1} = r_s \sigma_n$.

Alternatively, after having only measured the $\frac{\sigma_{n+1}^2}{\sigma_n^2}$, 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). *TEX: 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.