Appendix 2

Path, Causality, and Network Analysis

As humans, we cannot avoid thinking in terms of causality. (Asendorpf 2012.)

Version 22 August 2022

Wright (1921a) developed the method of **path analysis** as a means of interpreting the correlation between two variables in terms of hypothetical paths of causation between them. Initially, he was interested in the relative importance of general and specific growth factors for the variation of bones sizes in small mammals (Wright 1918), but he quickly realized the broad utility of his new technique and later applied it to many problems in genetics, agricultural economics, physiology, and ecology. More recently it has been extended and applied widely in systems biology and genomics. Provided the underlying assumptions are kept in mind, path analysis provides an extremely powerful, and conceptually simple, tool. Many of the fundamental principles of quantitative genetics can be derived by its use. It can also be a powerful approach for analyzing multiple levels of molecular data (such as joint data on QTL genotypes, mRNA and protein levels, and trait phenotypes).

The purpose of path analysis is the quantification of the relative contributions of causal sources of variance and covariance ***once a certain network of interrelated variables has been accepted*** (or assumed). While path analysis is not a formal technique proving the actual sources of causality (which can only come from careful experimentation), it does allows one to examine the level of support for different causality structures. In response to periodic abuses and criticism of the technique, Wright (1932, 1934e, 1968, 1983, 1984) repeatedly emphasized these points. Exceptionally lucid accounts of the theory and applications of basic path analysis are given by Li (1975) and Pedhazur (1982). Path analysis has grown from its simple roots and now forms the foundation for the rich fields of **causal analysis** (Spirtes et al. 2000; Pearl 2009; Pearl et al. 2016), **structural equation modeling** (**SEM**) (Bollen et al. 1989) and **graphical models** (Rosa et al. 2011, 2016; Sinoquet and Mourad 2014; Rohrer 2018), the later is also widely used in Bayesian modeling (Wu et al. 2010). Chapter 30 in WL examines applications of path analysis in determining the nature of selection on traits in the wild. A nice overview of many of these issues is given by Shipley (2016).

UNIVARIATE ANALYSIS

Through a graphical display, path analysis can greatly facilitate the analysis of a complex problem. Consider, for example, a system of four measurable variables, one (y) dependent and three $(z_1,\ z_2,$ and $z_3)$ potential explanatory values. Such a system can be displayed in the form of a **path diagram** (Figure A2.1). Single-headed arrows denotes a *direct path* from an explanatory variable to y, implying a cause-and-effect relationship. The connections between the explanatory variables are represented by double-headed arrows. It is assumed that y is a linear function of the z_i, although nonlinear relationships can be incorporated by considering z_i^2. Finally, unless y is known to be completely determined by the observed explanatory variables, an arrow is also drawn from the independent residual term, e.

Figure A2.1 displays only one of several possible path diagrams for our four-variable system. A great deal of information is contained in this diagram. For example, it can easily be seen that z_1 potentially influences y in three ways: directly by the path $z_1 \rightarrow y$ and indirectly by the paths $z_1 \leftrightarrow z_2 \rightarrow y$ and $z_1 \leftrightarrow z_3 \rightarrow y$. Similarly, z_2 influences y through paths $z_2 \rightarrow y$, $z_2 \leftrightarrow z_1 \rightarrow y$, and $z_2 \leftrightarrow z_3 \rightarrow y$, and z_3 influences y through $z_3 \rightarrow y$, $z_3 \leftrightarrow z_2 \rightarrow y$, and $z_3 \leftrightarrow z_1 \rightarrow y$. The independent residual term operates only through path $e \rightarrow y$.

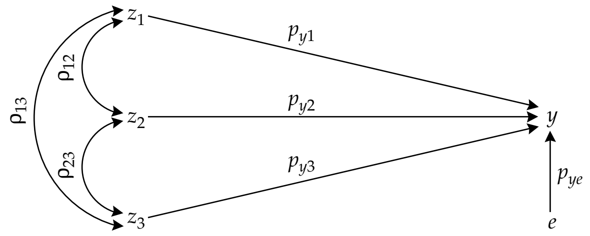


Figure A2.1 Path diagram for the variable y in terms of three explanatory variables (z_1, z_2, and z_3) and a residual error (e). p denotes a path coefficient and ρ a correlation (here between the z_i, standardized to have unit variance).

The labels on the double-headed arrows are the simple correlation coefficients (ρ) between the two denoted variables, while the quantities along single-headed arrows are **path coefficients** (p). The path coefficients for this diagram are standardized partial regression coefficients. If, prior to a multiple regression, each of the variables (y, z_1, z_2, and z_3) are standardized by subtracting the mean and dividing by the standard deviation so that the transformed variables all have zero means and unit variances, the subsequent partial regression coefficients equal the path coefficients.

To see this equivalence, we start with the general linear model

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + e \tag{A2.1a}$$

where $\beta_1, \beta_2, \cdots, \beta_n$ are partial regression coefficients. Subtracting $\overline{y}$ from the left and its equivalent $(\alpha + \beta_1 \overline{z}_1 + \beta_2 \overline{z}_2 + \cdots + \beta_n \overline{z}_n + \overline{e})$ from the right yields

$$y - \overline{y} = \beta_1(z_1 - \overline{z}_1) + \beta_2(z_2 - \overline{z}_2) + \cdots + \beta_n(z_n - \overline{z}_n) + (e - \overline{e}) \tag{A2.1b}$$

Squaring this expression and taking expectations, we obtain a general expression for the variance of y,

$$\sigma^2(y) = \sum_{i=1}^{n} \beta_i^2 \sigma^2(z_i) + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} \beta_i \beta_j \sigma(z_i, z_j) + \sigma_e^2 \tag{A2.2}$$

The residual variable is uncorrelated with the remaining variables under a least-squares analysis (Chapter 3), so covariance terms involving e do not appear in Equation A2.2. Dividing all terms in Equation A2.2 by $\sigma^2(y)$, recalling that $\sigma(z_i, z_j) = \rho_{ij}\sigma(z_i)\sigma(z_j)$, where ρ_{ij} is the correlation between variables i and j, and defining

$$p_{yi} = \beta_i \left[\frac{\sigma(z_i)}{\sigma(y)} \right] \tag{A2.3a}$$

$$p_{ye} = \frac{\sigma(e)}{\sigma(y)} \tag{A2.3b}$$

we obtain one of the fundamental equations of path analysis,

$$1 = \sum_{i=1}^{n} p_{yi}^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} p_{yi} \rho_{ij} p_{yj} + p_{ye}^2 \tag{A2.4}$$

This expression, known as the **equation of complete determination,** is a simple extension of the multiple regression equation. The p_{yi} are called **path coefficients**, and from Equation A2.3a can be seen to be standardized partial regression coefficients. Thus, the path coefficients are directly obtainable by multiplying partial regression coefficients by ratios of observed standard deviations (Equation A2.3a). Path coefficient p_{yi} may be interpreted as the expected change in y in standard deviations caused by a change in z_i in standard deviations when all other background variables are held constant.

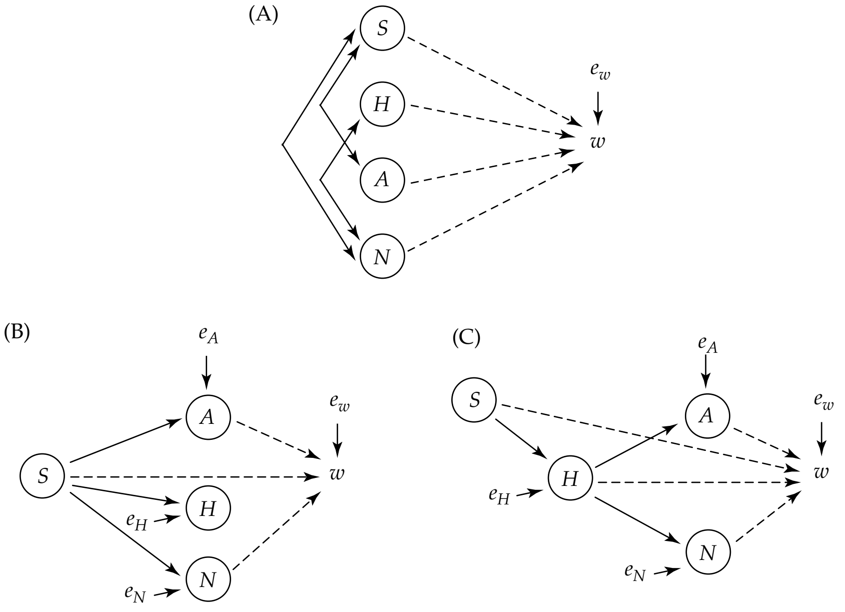
Equation A2.4 greatly expands the utility of multiple regression by explicitly partitioning the variance of y into proportional contributions from all of the direct and indirect paths of influence. The contribution from each path is a simple product of the correlation coefficients (for each double-headed arrow) and path coefficients (for each single-headed arrow) along a loop between variables. Thus, the contributions of the direct paths from z_1, z_2, z_3 and e to y are p_{y1}^2, p_{y2}^2, p_{y3}^2, and p_{ye}^2, respectively. The contributions from the indirect paths $y \leftarrow z_1 \leftrightarrow z_2 \rightarrow y$, $y \leftarrow z_2 \leftrightarrow z_3 \rightarrow y$, and $y \leftarrow z_1 \leftrightarrow z_3 \rightarrow y$ are, respectively, $2p_{y1}\rho_{12}p_{y2}$, $2p_{y2}\rho_{23}p_{y3}$, and $2p_{y1}\rho_{13}p_{y3}$. Each of the indirect paths is counted twice because they influence y in both directions (double-headed arrows). For example, the contribution of the path $y \leftarrow z_1 \leftrightarrow z_2 \rightarrow y$ to the variance of y is the same as path $y \leftarrow z_2 \leftrightarrow z_1 \rightarrow y$.

It is important to note that in computing the joint influence of two explanatory variables on y, only the direct correlation between the two variables is considered. Hence, $y \leftarrow z_1 \leftrightarrow z_2 \leftrightarrow z_3 \rightarrow y$ is not a contributing path in Figure A2.1. The entire correlation between z_1 and z_3 is contained in ρ_{13}. Thus, the general rules of path analysis are that: (i) ***there is only one two-headed arrow in any path***, and that (ii) the ***arrows change direction only once in a path***. Unlike correlation coefficients, path coefficients need not have absolute values less than one. Indeed, the total net contribution over a path can be negative (a negative correlation). The only constraint on Equation A2.4 is that the total contributions sum to one.

Regressions versus Path Analysis

The key distinction between a regression and a path analysis is that regression attempts to ***statistically account*** for the covariance between explanatory variables, while a path analysis further attempts to account for the ***processes generating correlations among the explanatory variables***. Regressions makes no assumption about the causality structure generating triat correlations, while Crespi (1990) noted that "path-analytic reasoning assumes that characters are correlated as a result of biological causes that should be used as information rather than adjusted away." A key limitation (and also strength) of a path analysis is that different assumed path diagrams for a set of variables (**causality structures**) can result in rather different biological interpretations (WL Chapter 30).

Example A2.1 As an example of different path diagrams for the same set of variables, consider the following example from Chapter 30 of WL.



The explanatory variables here are four plant traits: seed weight (S; the weight of the seed that gave rise to the focal individual), plant height (H), leaf area (A), and number of leaves (N). We seek to determine how these influence a measure of fitness, w, namely total seeds from a plant. For emphasis purposes, direct connections to fitness are denoted by dashed arrows. Path (A) is the structure for a standard multiple regression, which makes no assumptions about how the patterns of correlations among the traits arise. Paths (B) and (C) consider how developmental biology impacts the correlation structure among traits. In these structures there are also residuals associated with each of the downstream traits, as the influence of any upstream traits only determines part of their value. In both (B) and (C), the weight of the seed that founded the plant influences the other three traits being considered. Path model (B) assumes no hierarchical developmental structure from seed to the three traits, while path model (C) assumes that seed weight just influences height, and that height then influences the other two traits. Note that the covariance matrix of all three models is identical. Suppose we find that seed weight is correlated with fitness. Path diagrams B and C allow us to determine if S directly impacts fitness, only impacts it indirectly through its impact on downstream traits that in turn directly impact fitness, or has both direct and indirect effects (path analysis can quantify the importance of each).

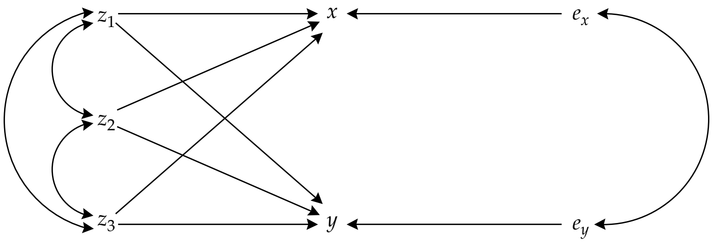


Figure A2.2 Path diagram for a system in which two dependent variables (x and y) are jointly influenced by three explanatory variables (z_1, z_2, and z_3) and residuals (e_x and e_y), which may be correlated. More complicated path diagrams could have one or more of the explanatory variables dependent on upstream explanatory variables, e.g., $z_1 \to z_2$ and/or $z_1 \to z_3$ (e.g., Example A2.1). Operationally, when a final path diagram is displayed, many of the proposed connections may not have statistical support, and these are thus often left out. Further, the strength of a path is often indicated by the thickness of the arrow, with thick arrows implying a strong connection and thin arrows a weak connection.

BIVARIATE ANALYSIS

An exceedingly useful property of path analysis is its ability to quantify the degree of association between two variables in terms of one or more mutually shared explanatory variables. Figure A2.2 is identical in form to Figure A2.1 except that z_1, z_2, and z_3 are now causal determinants of two characters, x and y. There are 10 distinct pathways connecting x and y: three direct paths $x \leftarrow z_1 \rightarrow y$, $x \leftarrow z_2 \rightarrow y$, and $x \leftarrow z_3 \rightarrow y$, and seven indirect paths $x \leftarrow z_1 \leftrightarrow z_2 \rightarrow y$, $x \leftarrow z_2 \leftrightarrow z_1 \rightarrow y$, $x \leftarrow z_1 \leftrightarrow z_3 \rightarrow y$, $x \leftarrow z_3 \leftrightarrow z_1 \rightarrow y$, $x \leftarrow z_2 \leftrightarrow z_3 \rightarrow y$, $x \leftarrow z_3 \leftrightarrow z_2 \rightarrow y$, and $x \leftarrow e_x \leftrightarrow e_y \rightarrow y$. The correlation between x and y is simply the sum of the products of path coefficients and correlation coefficients along these paths:

$$\begin{aligned}\rho_{xy} = {} & p_{x1}p_{y1} + p_{x2}p_{y2} + p_{x3}p_{y3} + p_{x1}\rho_{12}p_{y2} + p_{x2}\rho_{12}p_{y1} + p_{x1}\rho_{13}p_{y3} \\ & + p_{x3}\rho_{13}p_{y1} + p_{x2}\rho_{23}p_{y3} + p_{x3}\rho_{23}p_{y2} + p_{xe}e_{xy}p_{ye}\end{aligned}$$

where e_{xy} represents the correlation between the residual terms e_x and e_y. This expression

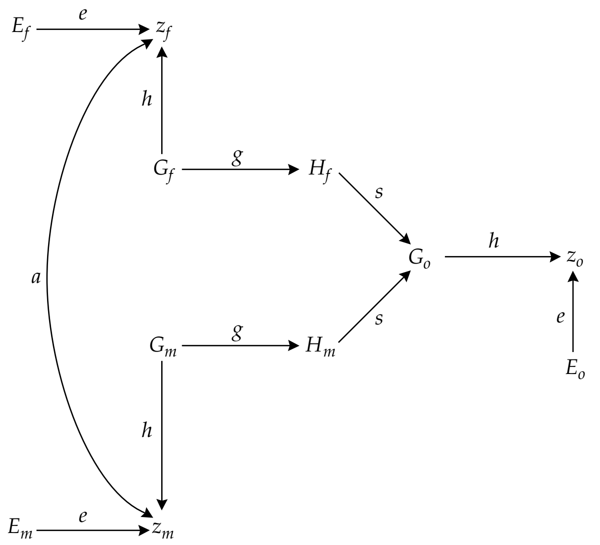


Figure A2.3 Wright's (1921a) path diagram, slightly modified, for the phenotypes of parents and offspring. Variables are defined in the text.

may be generalized to define the correlation between any two variables with n common causal sources of variation,

$$\rho_{xy} = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} p_{xi} \rho_{ij} p_{yj} \right) + p_{xe} e_{xy} p_{ye} \tag{A2.5}$$

For terms in which $i = j$, ρ_{ij} is set equal to one. Just as Equation A2.4 partitions a variance into components, Equation A2.5 partitions a correlation into a series of paths through shared explanatory variables.

APPLICATIONS

Path analysis is very useful in quantitative genetics because explicit statements can often be made about causality. Among the following examples, the first two illustrate how path analysis can be used to derive some fundamental relationships concerning phenotypic correlations.

Phenotypic Correlation Between Parents and Offspring

Much of the methodology of quantitative genetics relies on the comparison of phenotypic measures in close relatives. A common application involves the regression of offspring on parent, which was one of Wright's (1921a) earliest uses of path analysis. Defining an individual's phenotype (z) to be the sum of its genotypic value (G) and an environmental deviation (E), the phenotypes of a father (f) and mother (m) may be written

$$z_f = G_f + E_f$$
$$z_m = G_m + E_m$$

Provided the parents are not sibs, their environmental deviations may be treated as independent random variables with respect to G. If, however, mates select each other on the basis of phenotypes, a correlation may exist between z_f and z_m. We will denote this correlation by a. Under additive gene action, the genotypic value of an offspring is equal to the sum of the gametic contributions from its father (H_f) and mother (H_m),

$$G_o = H_f + H_m$$

An unambiguous path diagram can be constructed for such a familial structure (Figure A2.3). There are four path coefficients: h from genotypic to phenotypic values, e from environmental effects to phenotypic values, g from genotypic value to gametic value, and s from gametic value to genotypic value. These are assumed to be constant across generations.

We now consider the correlation between the phenotype of a parent and that of its offspring, ρ_{op}. Here we focus on the father-offspring correlation, although in this example, identical results arise for mother-offspring analysis. Regardless of which parent is considered, there are two paths connecting it to its offspring. The first results from the direct gametic contribution that a parent makes to its offspring; i.e., $z_f \leftarrow G_f \rightarrow H_f \rightarrow G_o \rightarrow z_o$. The contribution of this path to ρ_{of} is the product of four path coefficients, $hgsh$. The second path, which only exists under assortative mating, is an indirect route through a mate's gamete, i.e., $z_f \leftrightarrow z_m \leftarrow G_m \rightarrow H_m \rightarrow G_o \rightarrow z_o$. Its contribution to ρ_{of} is $ahgsh$. Summing up,

$$\rho_{of} = h^2 gs(1+a) \tag{A2.6}$$

A further simplification of this expression, which eliminates the coefficients g and s, is possible. The genotypic value G_o is determined by the two direct paths $H_f \rightarrow G_o$ and $H_m \rightarrow G_o$, each of which contributes a proportion s^2 to the variance of G_o, and by the indirect paths $G_o \leftarrow H_f \leftrightarrow H_m \rightarrow G_o$ and $G_o \leftarrow H_m \leftrightarrow H_f \rightarrow G_o$. The correlation between H_f and H_m is determined by the single path $H_f \leftarrow G_f \leftarrow z_f \leftrightarrow z_m \rightarrow G_m \rightarrow H_m$ and is equal to h^2g^2a. Therefore, each indirect path makes a proportional contribution of $(hgs)^2a$ to the variance of G_o. The equation of complete determination for G_o is then

$$1 = 2s^2 + 2(hgs)^2 a$$

which upon rearrangement yields

$$s = [2(1 + h^2g^2a)]^{-1/2} \tag{A2.7}$$

Wright (1921a) pointed out that, provided the path coefficients remain constant across generations, the correlation between a genotype and a gamete that *it produces* (G_f and H_f) will be the same as the correlation between a genotype and a gamete that *produced it* (G_o and H_f). It can be seen directly from the path diagram that the first of these correlations is simply g. There are two paths connecting G_o and H_f, ($H_f \rightarrow G_o$ and $H_f \leftarrow G_f \rightarrow z_f \leftrightarrow z_m \leftarrow G_m \rightarrow H_m \rightarrow G_o$), however, so their correlation is $s + ghahgs$. Equating this expression to g,

$$g = s(1 + h^2g^2a) \tag{A2.8}$$

Multiplying Equations A2.7 and A2.8 together, it can be seen that

$$gs = s^2(1 + h^2g^2a) = 0.5$$

Thus, Equation A2.6 simplifies to

$$\rho_{of} = h^2 \left(\frac{1+a}{2} \right) \tag{A2.9}$$

which in the absence of assortative mating ($a = 0$), reduces to

$$\rho_{of} = \frac{h^2}{2} \tag{A2.10}$$

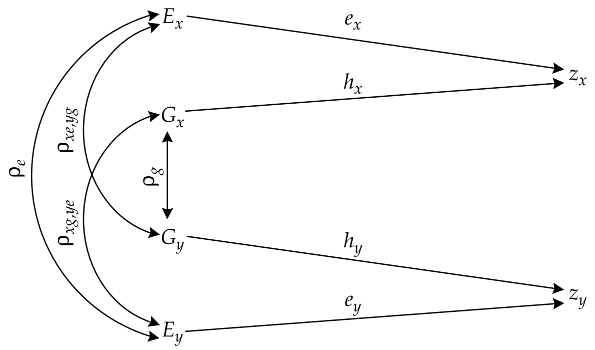


Figure A2.4 Path diagram for the phenotypic correlation between two characters (z_x and z_y) in terms of genetic values (G_x and G_y) and environmental deviations (E_x and E_y).

Mate selection on the basis of phenotypes influences the resemblance between parents and offspring in a particularly simple manner. Perfect disassortative mating ($a = -1$) completely eliminates the correlation between parent and offspring phenotypes, while perfect assortative mating ($a = +1$) doubles it.

Quantitative geneticists have long referred to the fraction of phenotypic variance that is additive genetic in basis as the narrow-sense heritability and abbreviated it as h^2. The use of this notation, particularly the square, may seem puzzling. Returning to Figure A2.3, the origin of h^2 can now be seen to be a historical tribute to Wright's (1921a) path diagram. Under the additive model of gene action, the equation of complete determination for an individual's phenotype is

$$h^2 + e^2 = 1$$

where h^2 is the proportion of the phenotypic variance due to the direct path from the genotypic value.

Correlations Between Characters

Path analysis can also be used to describe the correlation between two different characters, x and y, in the same individual. Here we denote the two phenotypes as

$$z_x = G_x + E_x$$
$$z_y = G_y + E_y$$

where, as usual, E_x has a mean of zero and is independent of G_x, and the same properties apply to E_y and G_y. The path diagram joining the two traits is drawn in its most general form in Figure A2.4. The path $G_x \leftrightarrow G_y$ indicates the possibility of a correlation between genotypic values of the two traits, owing to their expression being mutually determined by shared genes and/or linkage disequilibrium among alleles influencing the two traits (Chapter 26). Correlation between the environmental effects on the two traits is denoted by ρ_e, while those between the genotypic value of one trait and the environmental deviation of the other are indicated by $\rho_{xe,yg}$ and $\rho_{xg,ye}$. The phenotypic correlation between characters x and y, ρ_{xy}, derives from four possible paths: $z_x \leftarrow G_x \leftrightarrow G_y \rightarrow z_y$, $z_x \leftarrow E_x \leftrightarrow E_y \rightarrow z_y$, $z_x \leftarrow E_x \leftrightarrow G_y \rightarrow z_y$, and $z_x \leftarrow G_x \leftrightarrow E_y \rightarrow z_y$. Summing the appropriate products of path and correlation coefficients,

$$\rho_{xy} = h_x \rho_g h_y + e_x \rho_e e_y + e_x \rho_{xe,yg} h_y + e_y \rho_{xg,ye} h_x \tag{A2.11}$$

Note that there are only two arrows pointing to z_x and that these come from variables that are uncorrelated (as E_x and G_x are not connected by any paths). The same is true for z_y. Thus, by the equation of complete determination,

$$h_x^2 + e_x^2 = 1$$
$$h_y^2 + e_y^2 = 1$$

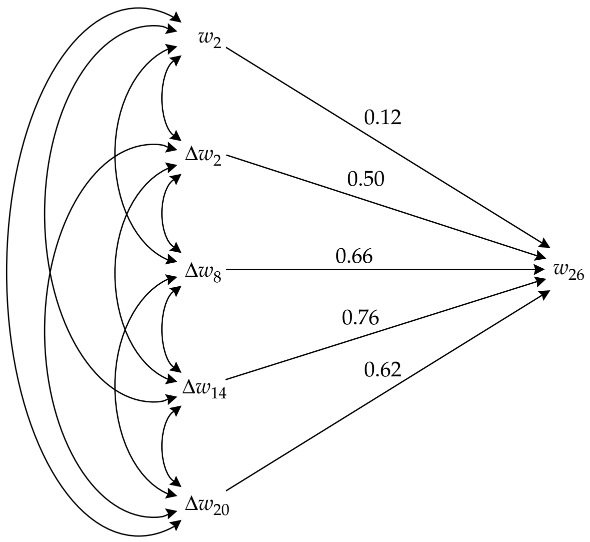


Figure A2.5 Path diagram for weight at day 26 as a function of five growth components for a population of feral pigeons. Numbers on the single-headed arrows are the path coefficients. The correlations between explanatory variables are given in Table A2.1.

Rearranging and substituting $e_x = \sqrt{1-h_x^2}$ and $e_y = \sqrt{1-h_y^2}$ into Equation A2.11,

$$\rho_{xy} = h_x h_y \rho_g + \rho_e \sqrt{(1-h_x^2)(1-h_y^2)} \\ + \rho_{xe,yg} h_y \sqrt{1-h_x^2} + \rho_{xg,ye} h_x \sqrt{1-h_y^2} \qquad \text{(A2.12)}$$

Thus, the phenotypic correlation between two traits is entirely described in terms of correlations between components of the traits and their heritabilities. Frequently in quantitative-genetic applications, the correlations between E_x and G_y and between E_y and G_x are zero. In this case, Equation A2.12 reduces to

$$\rho_{xy} = h_x h_y \rho_g + \rho_e \sqrt{(1-h_x^2)(1-h_y^2)} \qquad \text{(A2.13)}$$

Growth Analysis

Biological features in which a whole can be considered to be the sum of several individual parts are numerous. For example, the total diet of a predator often consists of several prey species, and the total seed set by a plant can be partitioned into contributions from various flowers. The problem considered here is the size of an individual (or character) at time t, z_t. This can be expressed simply as the sum of the initial size, z_1, and an arbitrary number (n) of subsequent growth increments, z_2 to z_n,

$$z_t = z_1 + z_2 + \cdots + z_n \qquad \text{(A2.14)}$$

With this type of model, all of the terms on the right necessarily sum to z_t, so there is no residual error term. Moreover, all of the partial regression coefficients (the β_i coefficients on the z_i in the multiple regression equation) are equal to one. Returning to Equations A2.1a and A2.2, it can be seen that the equation of complete determination for z_t reduces to

$$1 = \frac{1}{\sigma^2(z_t)} \left[\sum_{i=1}^{n} \sigma^2(z_i) + 2 \sum_{i=1}^{n} \sum_{j \geq i}^{n} \sigma(z_i, z_j) \right] \qquad \text{(A2.15)}$$

Table A2.1 Correlations (above diagonal) and path contributions (diagonal and below) of growth components to weight on day 26 for a sample of 100 feral pigeons. Here * and ** denote correlations that are significance at the 5% and 1% levels. The diagonal elements (direct contributions to the variance of w_{26}) are simply the squares of the respectively path coefficients given in Figure A2.5. The below-diagonal elements denote the contributions resulting from correlations between characters x and y, and are obtained as $2p_x\rho_{xy}p_y$. (*Source*: D. Droge, unpubl. data.)

	w_2	Δw_2	Δw_8	Δw_{14}	Δw_{20}
w_2	0.014	0.232*	0.100	−0.069	−0.186
Δw_2	0.027	0.255	−0.014	−0.316**	−0.045
Δw_8	0.015	−0.010	0.437	−0.157	−0.096
Δw_{14}	−0.012	−0.244	−0.159	0.584	−0.167
Δw_{20}	−0.027	−0.028	−0.079	−0.158	0.385

Thus, the elements of the variance-covariance matrix for the growth components in Equation A2.14 provide a complete description of the direct and indirect contributions to the variance of size at time t.

As an example of the application of Equation A2.15, the growth dynamics of a population of feral pigeons will be examined. One hundred birds were weighed at regular intervals from shortly after birth to fledging. The path analysis here will consider the weight at day 26 as a function of an initial weight at day 2 plus four subsequent six-day growth increments (days 2–8, 8–14, 14–20, and 20–26). Each growth increment is the difference between adjacent weighings, so letting w_t be the weight on day t, Equation A2.14 becomes

$$\begin{aligned} w_{26} &= w_2 + (w_8 - w_2) + (w_{14} - w_8) + (w_{20} - w_{14}) + (w_{26} - w_{20}) \\ &= w_2 + \Delta w_2 + \Delta w_8 + \Delta w_{14} + \Delta w_{20} \end{aligned}$$

The path diagram (Figure A2.5) illustrates that there are five direct paths and ten indirect paths to w_{26} (each of which must be counted twice). The proportional contributions of these paths are directly obtainable from the variances and covariances of the observed variables and are summarized in Table A2.1. Recalling that for this model all $\beta_i = 1$, the path coefficients can be computed using Equation A2.3a. For example, the contribution of direct path $w_2 \to w_{26}$ is simply $[p(w_2, w_{26})]^2 = \sigma^2(w_2)/\sigma^2(w_{26})$, while the total contribution from the indirect paths $w_2 \leftrightarrow \Delta w_2 \to w_{26}$ and $\Delta w_2 \leftrightarrow w_2 \to w_{26}$ is $2\sigma(w_2, \Delta w_2)/\sigma^2(w_{26})$.

Several aspects of the growth properties of this population are revealed by the path analysis. First, very little of the variation in size at age 26 is accounted for by the size at birth, i.e., $(p_{w_2,w_{26}})^2 = 0.014$. Most of it arises from variation in the post-natal growth rates. Second, all of the indirect paths make negative or negligibly positive contributions to w_{26}. The sum of the direct (diagonal elements of Table A2.1) and indirect (below-diagonal elements) paths are 1.675 and -0.675 respectively. The contribution from the path involving Δw_2 and Δw_{14} ($p = -0.244$) is particularly pronounced because of the highly significant negative correlation between Δw_2 and Δw_{14} ($\rho = -$ 0.316). On the other hand, while w_2 and Δw_2 are significantly positively correlated ($\rho = 0.232$), their indirect contribution to w_{26} is very small because of the small path coefficient from w_2 to w_{26} ($p_{w_2,w_{26}} = 0.12$; Figure A2.5). The preponderance of negative correlations between growth components is indicative of **compensatory growth**. Individuals that experience early periods of relatively rapid growth generally also experience subsequent periods of slowed growth. More details on this method of growth analysis, as well as estimators for the sampling variance of path coefficients, may be found in Lynch (1988d).

Causality Analysis in Biological Pathways

As introduced in Chapter 21, an important application of path analysis is in disentangling the roles of molecular features (such as QTL genotypes, mRNA and protein levels, chromatin features, methylation marks, etc.) in influencing a trait. The basic idea was introduced

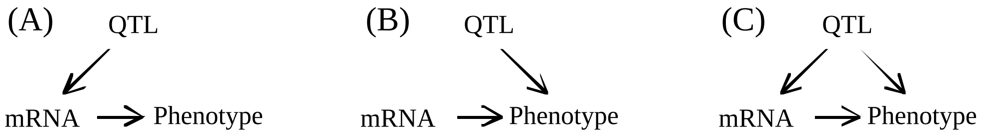
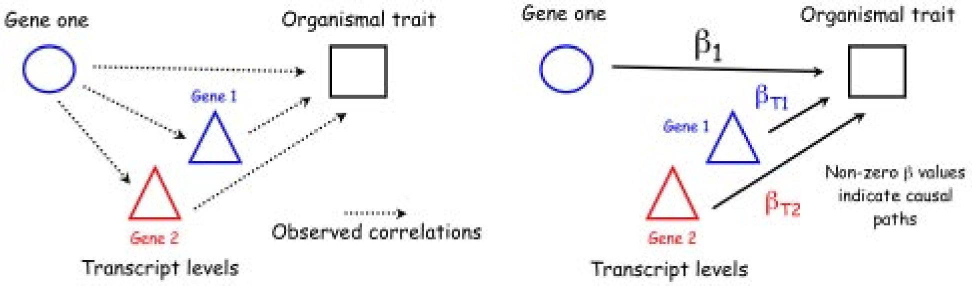


Figure A2.6 Suppose a detailed functional genomic analysis has shown that a focal trait is influenced by the genotype at a QTL and also by mRNA level (a quantitative trait transcript, QTT), potentially from a different locus than the focal QTL. **(A):** The impact of the QTL is as an eQTL that influences the trait only by influencing QTT levels. **(B):** The QTL has an impact on the trait independent of QTT levels. **(C):** The QTL acts on both the QTT and the trait.

by Schadt et al. (2005) and extended by a number of authors (Kulp and Jagalur 2006; Liu et al. 2008; Millstein et al. 2009; Wu et al. 2010). The term **endophenotype** refers to any intermediate molecular phenotype in the mapping of a genotype to a final trait value (such as mRNA, protein, or metabolite levels; histone acetylation/methylation status, etc.). With the rise of functional genomics, a very large number of such intermediate features can be scored (Chapter 21), and critical to our understanding of trait expression is determining the causal structure of how these components interact. Figure A2.6 shows a simple example wherein levels of mRNA (T) from a particular locus influences trait value z (it is quantitative trait transcript, QTT, for that trait; Chapter 21). The genotype (G) at a focal QTL also influences the trait phenotype, so that in a simple correlation analysis, both the QTL and QTT appear to influence the trait (both G and T are correlated with z). However, as Figure A2.6 shows, there are three different causal structures (we ignore the reverse causal structure, QTL $\rightarrow$ trait $\rightarrow$ QQT). Note that these options correspond to three different path diagrams. If (A) is correct, then, conditional on the QTT value, there should be no residual correlation between QTL genotype and trait value. If (B) is correct, conditioning on the QTT value will not change the correlation between QTL genotype and trait, while if (C) is correct, the residual correlation between the QTL and trait will change, but will remain nonzero.

Example A2.2 Consider the situation diagrammed below where a QTL (gene one; represented by a circle) has an impact on both its own transcript level and that at a second gene, represented by triangles (Mackay et al. 2009). Transcript levels of both genes are correlated with the trait value, as is the genotype at the QTL. How can we determine which factor (or factors) directly impact the trait?



Assuming no other features (confounders) that influence two (or more) of these components (e.g., Figure 21.3B), the general setting is shown on the left in the above figure. The figure on the right shows the direct impacts we wish to test: β_1 for the direct effect of the QTL on the trait, β_{T1} for the direct impact of transcript one level on the trait, and finally β_{T2} for the direct impact of transcript two level on the trait. If we have a set of individuals with data vectors (z, G, T_1, T_2), then these various β can be estimated from the multiple regression

$$z = \mu + \beta_1 G + \beta_{T1} T_1 + \beta_{T2} T_2 + e$$

We can also estimate the strength of the paths from gene one to both transcripts (β_{1Ti}), and thus compute the indirect contribution to the trait, namely $\beta_{1T1}\beta_{T1} + \beta_{1T2}\beta_{T2}$, corresponding to the paths gene one → T_1 → trait and gene one → T_2 → trait, respectively. See Example 21.3 (and Figure 21.3) for an application of this approach (often called **mediation analysis**) for determining the impact of one transcription on another, potentially downstream, transcript (*cis*-mediated *trans* effects).

Finally, some basic terminology for causality, Consider three factors, A, B, and C that have some causal association. The setting where $A \rightarrow B \rightarrow C$ is called a **chain**, and B is said to be a **mediator**. If this is the only path connecting A and C, then conditioning on B results in A and C being uncorrelated. The setting where $A \leftarrow B \rightarrow C$ is called a **fork**, and B is said to be a **confounder**. Conditioning on B again removes the correlation between A and C. Finally, the setting where $A \rightarrow B \leftarrow C$ is called an **inverted fork**, and B is said to be a **collider**. Here conditioning on B can result in a false correlation between A and C (Rohrer 2018).

MENDELIAN RANDOMIZATION

The method of **Mendelian randomization** (**MR**) combines concepts from path analysis, meta-analysis (Appendix 6), and GWAS (Chapter 20) to use genetics as a complement for randomized control trails (RCT; Appendix 9) when examining the causal impact of a potential risk factor on an outcome (e.g., disease). We can illustrate the basic idea as follows. Suppose a large population sample shows a correlation between hair loss and sugar intake (observational epidemiology). It could be that: (i) sugar → hair loss, (ii) hair loss → sugar, (iii) an unmeasured factor (a confounder, U) influences both, hair loss $\leftarrow U \rightarrow$ sugar, or (iv) some combination of these three scenarios. In an ideal setting of a RCT, one would start with a very young cohort (ideally at birth), randomly assign them into high and low sugar diet groups, and follow hair loss over the next fifty years. Of course, such RCTs for assessing the risk of long-term exposure to some factor are generally only possible in laboratory and domesticated species under controlled conditions. Even when such an RCT is feasible, its sample size will almost certainly be too small to detect modest effects (say an odds ratio, OR, $<$ 1.5). Now suppose a GWAS finds a particular allele that naturally generates high sugar levels relative to its alternate allele. One could now look at hair loss as a function of this allele in a random population sample. Further, the massive sample size of a modern GWAS routinely detects very small effects (ORs on the order of 1.05 or less; Chapter 20). The randomization of the treatment (sugar level) occurs via *Mendelian segregation* at conception by randomly assigning alleles to offspring, with any other environmental factors naturally randomized over the two groups. This leads to what Hingorani and Humphries (2005) called "Nature's randomised trials."

Strictly speaking, a truly randomized design would contrast variants *within* a family (e.g., between sibs), while MR is typically performed on a population sample. As with the TDT test (Chapter 17), within-family contrasts remove any complications from population stratification. In a population sample, structure could be present, where both the disease and one of the variants occur in a subpopulation at higher frequencies than the rest of the sample, creating an association between them (Chapter 20). Davies et al. (2019) discuss other sources of bias that can arise in a population-based MR, and how within-family MR accommodates these concerns.

Even in settings where a short-term RCT is at least conceptually possible, MR still has advantages. Differences in cost and effort is an obvious one, using a population sample as opposed to a smaller and more expensive longitudinal RCT. Thus, one can think of an MR as a *screening procedure* for candidate associations for a follow-up RCT (such as in drug discovery settings). While (as discussed below), false positives can arise in MR studies, because of

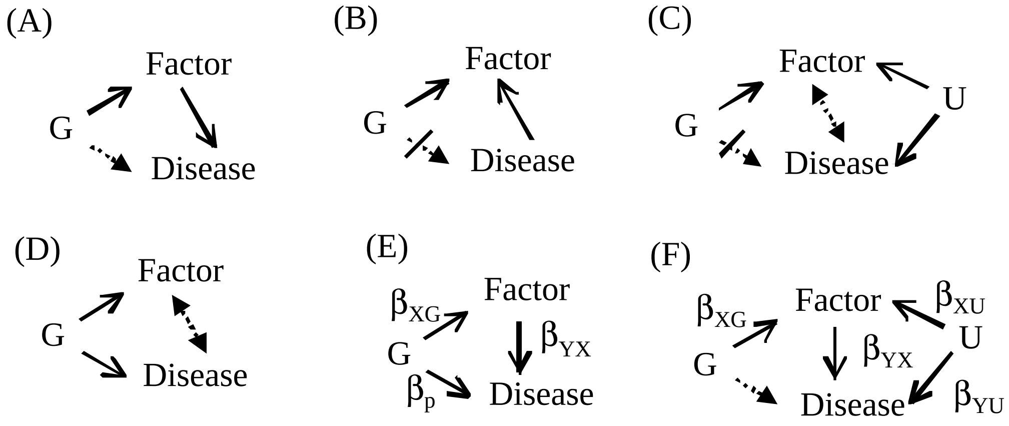


Figure A2.7 Path diagrams for settings where MR will work (A through C), and where it can fail (D), in testing the hypothesis that a factor (*X* in general, with *G* used when the factor is a specific genotype) influences disease risk (*Y*). Solid arrows indicate causal connections, dashed arrows indicate induced correlations. The **instrumental variable** (or **instrument**), *G*, is the genotype at a locus influencing the factor being tested. Factor is often denoted as **exposure** in the literature, and disease can be generalized to an **outcome** variable (or state). **(A)** When the SNP influences the factor *and* the factor influences the disease, we see a correlation between the SNP and the disease state. **(B)** MR controls for reverse causality (disease → factor). Because the disease cannot alter the genotype, there is no SNP–disease correlation. **(C)** MR controls for a confounder (*U*) that influences both the factor and disease state. While *U* results in a correlation between the factor and disease state, there is no correlation between the SNP and disease. **(D)** MR can fail when the SNP has a pleiotropic effect on both the factor and the disease. This induces a correlation between the factor and the disease in the presence of a SNP–disease correlation. Note that pleiotropy could either be strict (the single SNP influences both traits) or LD-induced, where a linked SNP influences a different trait from the target SNP. More generally, MR can fail when there is a path from *G* to *Y independent* of *X*. **(E)** In the presence of a pleiotropic SNP when the factor also influences the trait, the total association β_{YG} between *G* and *Y* is the sum of two different paths, one direct (SNP → disease, β_p) and one indirect (SNP → factor → disease; $\beta_{XG}\beta_{YX}$). **(F)** When the factor influences the disease, and both are influenced by a confounder, MR still correctly shows that factor → disease, but gives a biased estimate of the true effect, β_{YX}, when *G* is a weak instrument (Equation A2.17b).

their greater power relative to a RCT, a negative MR result (even if a false negative because of low power) will likely generate a negative RCT result. The caveat to this statement is that MR usually involves a rather small change in the exposure, while an RCT typically imparts a much larger perpetuation, but this may be offset by the fact that MR often represents life-long exposure whereas RCTs only over the course of the experiment. Further, because alleles have been allocated at birth, MR deals with the setting where the critical time of exposure is unknown (i.e., its impact could occur at a younger age than used in the RCT). MR also accommodates cases where withholding of a treatment might be unethical (e.g., Gray and Wheatley 1991) and, conversely, when *applying* a treatment would be unethical, such as examining the health risk from pesticide exposure. If genetic variation exists in the ability to detoxify a suspected agent, this provides an MR solution to test for exposure risk that is controlled for confounders (Davey Smith and Ebrahim 2003). Finally, one can use **intergenerational Mendelian randomization** to test for maternal effects, using the randomization in the mother of alleles influencing a factor suspected to be involved in a material risk effect that influences the phenotype of its offspring (Davey Smith and Ebrahim 2004; Lawlor, et al. 2017; Evans et al. 2019).

The basic MR concept was proposed by Katan (1986), with Gray and Wheatley (1991) later coining the term Mendelian randomization. Katan noted that low cholesterol levels correlated with increased cancer risk, but it might be that an early (and undetected) tumor decreases cholesterol levels long before the cancer is otherwise displayed (**reverse causality** with cancer → cholesterol). However, certain alleles at the *ApoE* gene naturally have reduced cholesterol levels, starting from birth. Katan reasoned that one could use control of the *ApoE* genotype as a surrogate for control of cholesterol level. If low cholesterol → cancer, then these low-cholesterol alleles should be over-represented in cancer patents. Conversely, if cancer → low cholesterol, then no such association is expected. See Figure A2.7 for the path diagrams for these, and other, settings.

While the logic of MR is straightforward, as with many issues in quantitative genetics, the devil is in the details. As noted by Burgess et al. (2019), MR "*is an important and valuable tool for learning about causal relationships using genetic data, but it is also a fallible one. ⋯ fundamentally the approach depends on being able to find genetic variants that are plausible proxies for intervention on the risk factor*." One should always perform sensitivity analysis (detailed below) before an MR analysis is considered complete, especially one based on multiple GWAS hits. As noted by Burgess et al. (2017), if no such analysis is done, then any results should "*be viewed as speculative and incomplete*" and "*treated with skepticism.*"

The development of MR methodologies is a rapidly growing field that we will only briefly examine. Reviews and discussion are given by Davey Smith and Ebrahim (2003, 2004, 2005), Thomas and Conti (2004), Nitsch et al. (2006), Didelez and Sheehan (2007), Ebrahim and Davey Smith (2008), Lawlor et al. (2008), Sheehan et al. (2008), Brion et al. (2014), Davey Smith and Hemani (2014), VanderWeele et al. (2014), Boef et al. (2015), Burgess and Thompson (2015), Burgess et al. (2015b, 2017a, 2018, 2019), Haycock et al. (2016), Zheng et al. (2017), Hemani et al. (2018), Slob and Burgess (2019), and Zhu (2020). Power calculations are discussed by Brion et al. (2013), Freeman et al. (2013), and Deng et al. (2020).

Example A2.3 Benn et al. (2017) used MR to examine whether a low level of LDL cholesterol is a risk factor for either Alzheimer's or Parkinson's disease. Cholesterol is a major component of myelin sheaths enclosing neurons in the brain, and observational studies suggest a correlation between lower LDL levels and increased risk for both Alzheimer's and Parkinson's. Indeed, Been et al. found such a significantly increased risk for Parkinson's in a population sample of 110,000 Danes. This is of serious concern given the widespread use of statins and other cholesterol-lowering drugs to reduce heart disease risk. The authors initially focused on two genes segregating alleles that naturally lower LDL levels, *PCSK9* and *HMGCR*. They first showed that variants at these loci were independent of known confounding risk factors of the LDL cholesterol-disease relationship (age, sex, hypertension, smoking, physical activity, alcohol consumption, educational level, menopausal status). An MR involving 26 variants at these loci found that LDL-lower alleles showed no increased risk for either neurological disorder. The authors then considered 380 variants at other LDL-lowering loci detected from a GWAS. Again, no association between LDL-lowering alleles and increased disease risk was found. Indeed, collectively these alleles showed a significant *reduction* of Alzheimer's risk (OR 0.64, CI 0.52 to 0.79).

Factors probed by MR are not just restricted to presumed molecular intermediates, they can also be modifiable behaviors. One such example is Chen et al. (2008), who used variants at *ALDH2* as surrogates for alcohol consumption. Individuals homozygous for null *ALDH2* alleles experience adverse effects to alcohol, and as a result drink considerably less than other genotypes. Variants at a second locus (*ADH*) similarly influence alcohol consumption, and Howe et al. (2019) used an *ADH* variant (rs1229984) to examine potential causes for the correlation in alcohol consumption among couples. One possibility for this correlation is that the behavior of one partner influences that of the other. A second is mate choice, seeking a partner with a similar alcohol consumption (assortative mating). A third possibility is an unknown confounder influencing both mate choice and alcohol consumption (such as socioeconomic status). Howe et al. found that the rs1229984 status of one partner predicts the phenotypic (consumption) status of the other. Further, it also predicts the rs1229984 status of their partner, suggesting that assortative mating drives this observed relationship. A number of other interesting human examples can be found in the reviews by Davey Smith and Ebrahim

(2004) and Sheehan et al. (2008).

There are obvious applications of MR beyond human studies. For example, in WL Chapter 20, we discussed a hypothetically observed association between seed level (as a proxy of fitness) in plants and alkaloid levels (plant secondary compounds). This might suggest a beneficial role of these compounds, for example by reducing insect damage. However, it could also represent a confounding effect, with (say) plants grown in nitrogen-rich soil having both higher fitness and higher alkaloid levels. If a GWAS found SNPs showing variation in alkaloid production, then if alkaloid → seeds, higher alkaloid alleles should also have more seeds, while if there is a confounding effect on both, no such association is expected.

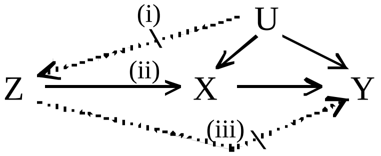


Figure A2.8 The three requirements for Z to be an instrumental variable (or instrument) for the impact of a factor X on an outcome Y. First, any confounder (U) that influences both the factor and the disease (Y) is *not* influenced by Z (lack of path i). We also assume that the confounder does not influence Z ($U \rightarrow Z$), which is very reasonable when the instrument is a genotype. Second, Z must directly influence the factor X (path ii is present). Finally, there is *no* direct path from Z to Y (path iii is absent).

The Concept of Instrumental Variables (IVs)

The underpinning of MR is the concept of an **instrumental variable** (**IV**), or **instrument**, a idea originally from economics. This concept first appeared in Appendix B of Philip G. Wright's (1928) *The tariff on animal and vegetable oils*. There was, however, much speculation that Appendix B was actually written by Philip's eldest son, Sewall (yes, the inventor of path analysis). After a detailed analysis, Stock and Trebbi (2003) concluded it was more likely than not that Philip *wrote* the Appendix, but there was strong evidence that the key *idea* came from Sewall. Brief reviews of the utility, and limitations, of IVs in epidemiological studies are given by Greenland (2000) and Hernán and Robins (2006).

To more formally define an IV, suppose the goal is to test whether X is causal to Y, and (if so) estimate the strength β_{YX} of this effect (how much a unit change in X changes Y). Let U denote the set of all variables that influence *both* X and Y (i.e., all the confounders). Then as shown in Figure A2.8, Z is an IV if it satisfies three conditions: (i) Z is independent of U, (ii) Z is associated with X, and (iii) Z is independent of Y conditional on X and U (this is often called the **exclusion restriction**), namely that there are no paths connecting Z and Y other than the one that runs through X. If any of these conditions fail, Z is said to be an **invalid instrument**. Condition (iii) is the assumption of no (horizontal) pleiotropy when the IV is a genotype, and is often problematic. Condition (i) can fail in the presence of population stratification (Chapter 20) or sampling following selective survival (Nitsch et al. 2006). A more detailed discussion of other routes by which one, or more, of these assumptions can fail is given by VanderWeele et al. (2014), who, along with Glymour et al. (2012), also review validation methods. In what follows, we use Z to denote a general instrumental variable, and G when the instrument is a genetic variant (i.e., the MR setting).

When the IV assumptions hold, then applying the rules of path analysis to Figure A2.7A, we have

$$\beta_{YZ} = \beta_{XZ}\,\beta_{YX}, \quad \text{implying} \quad \beta_{YX} = \frac{\beta_{YZ}}{\beta_{XZ}} \tag{A2.16a}$$

where β_{YZ} and β_{XZ} are the regression slopes of Y and X, respectively, on Z. This leads to the instrumental variable **ratio estimate** (or **Wald estimate**) for β_{YX},

$$\widehat{\beta}_{IV,R} = \frac{\widehat{\beta}_{YZ}}{\widehat{\beta}_{XZ}} \tag{A2.16b}$$

Thomas et al. (2007) approximated the standard error using the delta method (Equation A1.19b), whose leading term is

$$\sigma^2(\,\widehat{\beta}_{IV,R}\,) \simeq \frac{\sigma^2(\,\widehat{\beta}_{YZ}\,)}{(\,\widehat{\beta}_{XZ}\,)^2} \tag{A2.16c}$$

Equation A2.16c assumes that β_{XZ} is known without error ($\widehat{\beta}_{XZ} = \beta_{XZ} + e$, where $|e| \ll |\beta_{XZ}|$), commonly referred to in the MR literature as the **NOME** assumption (**NO Measurement Error**; Bowden et al. 2016a). As a result, Equation A2.16c is often called the **NOME weight** for the ratio. Improved approximations were given by Thomas et al. (2016), Thompson et al. (2016), and Bowden et al. (2019), while other approaches for obtaining the SE were discussed by Pierce and Burgess (2013).

When pleiotropy is present (a path from Z to Y without going through X), Equation A2.16b is a biased estimate of β_{YZ}. As outlined in Figure A2.7E, letting β_p be the path coefficient from pleiotropy, the total association between Z and Y becomes

$$\beta_{YZ} = \beta_p + \beta_{XZ}\,\beta_{YX} \tag{A2.17a}$$

yielding

$$\frac{\beta_{YZ}}{\beta_{XZ}} = \beta_{YX} + \frac{\beta_p}{\beta_{XZ}} \tag{A2.17b}$$

The careful reader might notice that even when $\beta_p = 0$, G technically still displays pleiotropy as it impacts both X and (indirectly) Y. The MR literature refers to the effect on Y from G that is fully mediated through X ($\beta_{XZ}\,\beta_{YX}$) as **vertical pleiotropy**, also called **mediated** (Solovieff et al. 2013), **secondary** (Hodgkin 1998), or **spurious pleiotropy** (Gruneberg 1938). Any pleiotropic effect of G on Y that does not pass through X (β_p) is called **horizontal** or **biological pleiotropy** (Solovieff et al. 2013). In our following discussion, any mention of pleiotropy refers to the latter.

Estimating a quantity by the ratio of two estimates can have undesirable statistical properties. An alternative approach that avoids a ratio is to use **two-stage least-squares** (**2SLS** or **TSLS**) regression. Assume the instrument is a genotype, G. Under 2SLS, one first predicts the value of X given G, $\widehat{X}|G$, and then regresses Y on $\widehat{X}|G$, namely,

$$Y_i = \alpha_{YX} + \beta_{IV,2S}(\widehat{X}_i|G_i) + e_i, \quad \text{where} \quad \widehat{X}_i|G_i = \alpha_{XG} + \beta_{XG}\,G_i, \tag{A2.18}$$

where the estimated slope, $\widehat{\beta}_{IV,2S}$, for the regression of Y on $X|G$ is taken as the estimate for β_{YX}. The 2SLS estimator easily generalizes to multiple instruments (predicting X using a multivariate regression on IVs). This is what a TWAS does, where X is the level of a focal transcript, which is predicted using the (*cis*) eQTL genotypes of an individual (Chapter 21).

A key design concern with any MR is the sampling strategy. In a **single-sample design**, $G, X,$ and Y would be measured on all the members of some sample. However, the intermediate factor X may be more challenging and/or expensive to measure. A **subsample** approach scores X on only some subset of the sample. Pierce and Burgess (2013) showed that, for a strong instrument (defined below), scoring only 20% of the samples for X results in almost the same power as scoring the entire sample. A **two-sample design** can also be used, wherein the association between X and G is measured in one sample and between Y and G in a second. Such a setting allows GWAS summary statistics (Chapter 20 and 21) to be used in place of (often unobtainable) individual-level data. Indeed, these two estimates could be based on a meta-analysis over a number of studies (Appendix 6). Obviously, bias can arise if these different samples have different features (e.g., differences in confounders, population or LD structure, etc.). Minelli et al. (2004), Haycock et al. (2016), and Burgess et al. (2017b) discussed how to use standard sensitivity analysis tools from meta-analysis (heterogeneity tests, forest and funnel plots; Appendix 6) to look for biases and inconsistencies in the data sets. A final issue is Beavis effects, in that a marker detected as being significant (under a low power setting) often has its effect overestimated (Chapters 18 and 20). Richmond et al. (2014) recommended using a subsampling approach in such cases, detecting a G–X association using some part of the sample, and then estimating its effect, β_{XG}, in the remanding part (e.g., Example 21.1).

When using a valid instrument, MR is a rather robust procedure for *demonstrating* that X is causal to Y. However, estimating the actual *strength* of such an association is a bit more challenging. One issue is that even when G is a valid IV, it may still be a **weak instrument**, having a small impact of X relative to the noise in the system. The term weak as slightly misleading, as a very small actual effect (G accounts for only a small fraction of the variance in X) is a weak instrument under small sample size, but becomes strong as the sample size increases. The rough rule in the literature is that the F statistic for the regression of X on G should exceed ten for an instrument to be considered strong (Staiger and Stock 1997; Lawlor et al. 2007). Equation A2.19b shows how this transition from weak to strong occurs as sample size increases.

The ratio estimate of β_{YX} is biased by finite sample size, which is generally only a problem when a weak instrument is used. The directional impact of **weak instrument bias** depends upon the sampling scheme. In a two-sample design, because the estimates of β_{XG} and β_{YG} are from independent data, they are independent and (as we show shortly), the *bias shrinks the estimate of* β_{YX} *towards zero* (in the direction of the null). Conversely, under a one-sample design, the estimate of β_{YX} is *skewed towards any confounder effects* (Burgess and Thompson 2011; Pierce and Burgess 2013).

To see this distinction, consider Figure A2.7F, and let X_i, Y_i, and U_i be the values of the factor, disease, and confounder in individual i. Then

$$X_i = \beta_{XG} G_i + \beta_{XU} U_i + e_{X_i}, \qquad Y_i = \beta_{YX} X_i + \beta_{YU} U_i + e_{Y_i} \tag{A2.19a}$$

Under a one-sample design, the value of the confounder U_i is shared by X_i and Y_i, while *independent* draws of U influence X and Y in a two-sample design. Under the simplifying assumption of two states of the instrumental variable (1 and 0) and ignoring any effects of e, Burgess and Thompson (2011) showed (for a one-sample design) that the ratio estimator implies

$$\beta_{IV,R} = \beta_{YX} + \beta_{YU} \left(\frac{\Delta \overline{U}}{\beta_{XG} + \beta_{XU} \Delta \overline{U}} \right) \tag{A2.19b}$$

where $\Delta\overline{U} = \overline{U}_1 - \overline{U}_0$ is the random sample differences in the mean confounder values between genotypes. With a weak instrument ($|\beta_{XG}| \ll |\beta_{XU}\Delta\overline{U}\,|$), then $\beta_{IV,R} \simeq \beta_{YX} + \beta_{YU}/\beta_{XU}$, showing a bias towards the confounder effect (β_{YU}). Assuming n of each of the genotypes in the sample and that $U \sim N(0, \sigma^2_U)$, then $\Delta\overline{U} \sim N(0, 2\sigma^2_U/n)$. Hence, $|\Delta\overline{U}\,|$ shrinks towards zero as n increases, changing a weak IV into a strong one for sufficiently largely n, with $|\beta_{XG}| \gg |\beta_{XU}\Delta\overline{U}\,|$ and thus $\beta_{IV,R} \simeq \beta_{YX}$.

An issue related to weak instruments is **canalization**, the buffering of fluctuations in intermediate steps in a pathway to limit downstream perturbations of the final end product. As noted by Davey Smith and Ebrahim (2003), canalization can result in a rather large change in X from a genetic variant imparting only a small change in Y. A similar issue could occur if the variant polymorphism impacts a product that is a few steps removed from the factor we are tying to test. Finally, the basic MR analysis assumes that Z (or G), X, and Y are all *scalars*. We next consider a *vector* **Z** (**G**) of instruments (genotypes) for predicting X. Extensions allowing for a vector **X** (a correlated set of exposure factors) and/or a vector **Y** (a correlated set of outcomes, such as closely related diseases) have been considered by Thompson et al. (2005), Burgess and Thompson (2015), Rees et al. (2016), and Sanderson, et al. (2019).

Mendelian Randomization Using Multiple Variants

MR can be a powerful tool if one has a valid strong single instrument (e.g., a variant with a large effect on X). However, most traits are impacted by a large number of loci of modest to small effect (Chapters 20 and 21). As a result, the early days when MR were mostly based on Mendelian single factors (e.g., Davey Smith and Ebrahim 2003) gave way to extensions to exploit polygenic GWAS signals. In this setting there is still a single exposure (X) and outcome (Y) variable, but now MR leverages information from multiple GWAS hits influencing X to test its causality on Y (e.g., Example A2.3).

There are obvious advantages, and pitfalls, to extending an analysis to multiple instruments. Even if each has a fairly weak effect on X, their cumulative effect could be considerable, offering a substantial increase in power. Countering this, as more variants are added, they bring with them the risk that at least some are invalid instruments (such as displaying horizontal pleiotropy). Finally, LD can induce correlations among multiple GWAS hits in the same gene, and one must adjust for these. A variety of solutions for treating these concerns have been suggested, many based on concepts from meta-analysis (Appendix 6).

We start with the strictest assumption, that *all* of the k tested variants are *valid instruments*. A second assumption is *functionally consistent effects* over all variants. While each variant may have quantitatively different impacts on X (β_{XG} can vary over G), β_{YX} is the same for all: no matter how a change ΔX is generated, the resulting average change in Y is $\beta_{YX}\Delta X$. A example where this assumption might fail is if one SNP is an eQTL, while another is both an eQLT and an sQTL (Table 21.1). Both SNPs might give the same total level of transcript, but differential splicing may have a greater impact, so that if X is simply total transcript amount, β_{YX} will likely vary over these SNPs.

Because most GWAS-based MRs use a two-sample design, we assume this structure in the rest of our discussion. Let β_{Yj} and β_{Xj} be the estimated regression slopes of Y and X, respectively, on the instrumental variable, variant j. Using Equation A2.16b, one constructs the ratio estimate, $\widehat{\beta}_{IV,R,j}$, for each variant. If the IV assumptions (Figure A2.8) hold, these should all estimate a common parameter (β_{YX}), subject to sampling noise. We can express this as a standard fixed-effects meta-analysis (Equation A6.30a),

$$\widehat{\beta}_{IV,R,j} = \beta_{YX} + e_j, \qquad \text{with} \quad e_j \sim N\left[0, \sigma^2(\widehat{\beta}_{IV,R,j})\right] \tag{A2.20a}$$

This is just a GLM (Chapter 10), with the BLUE estimate of β_{YX} given by the inverse-variance weighted average (Equation A6.30b),

$$\overline{\beta}_{YX} = \frac{\sum_{j=1}^{k} w_j\, \widehat{\beta}_{IV,R,j}}{\sum_{j=1}^{k} w_j}, \quad \text{where} \quad w_j = \frac{1}{\sigma^2(\widehat{\beta}_{IV,R,j})} \tag{A2.20b}$$

Equation A2.20b assumes the estimates over different variants are uncorrelated, but can be modifed to allow for LD (correlations among variants). Using the approximation for the standard error of the ratio estimate given by Equation A2.16c (the NOME assumption),

$$w_j \cdot \widehat{\beta}_{IV,R,j} \simeq \left(\frac{\beta_{Xj}^2}{\sigma^2(\beta_{Yj})}\right) \cdot \left(\frac{\beta_{Yj}}{\beta_{Xj}}\right) = \beta_{Yj}\beta_{Xj}/\sigma^2(\beta_{Yj}) \tag{A2.20c}$$

and Equation A2.20b reduces to the **inverse-variance weighted** (**IVW**) estimator of Burgess et al. (2013),

$$\overline{\beta}_{YX} = \frac{\sum_{j=1}^{k} \beta_{Yj}\beta_{Xj}/\sigma^2(\beta_{Yj})}{\sum_{j=1}^{k} \beta_{Xj}^2/\sigma^2(\beta_{Yj})} \tag{A2.20d}$$

whose standard error is given by Equation A6.30c.

Another approach follows from the concept of **polygenic risk scores** (Chapter 31). The idea is that one takes a candidate set of variants and constructs an index of these to serve as a single instrumental variable. This is the **allele scores** method of Burgess and Thompson

(2013). One could either construct an unweighted score (simply counting the number of risk alleles, namely those increasing X) or weight alleles by some scheme (such as by their estimated effects as in a polygenic risk score). The idea is to avoid weak instrument bias by using a single sum as opposed to averaging ratios based on a number of weak instruments. Simulations by Burgess and Thompson showed that this approach can be powerful when all of the variants are valid instruments, but can be severely biased when as few at 10% of the variants are invalid. Further, seriously biased estimates often occur with weighted sums, as the weights are often incorrectly estimated (e.g., Beavis effects). Burgess et al. (2015a) detailed how to use summary data and correlated variants in constructing allele score tests.

A number of more robust estimators allowing from some of the instruments (variants) to be invalid (e.g., show pleiotropy) have been proposed. The simplest two were suggested by Bowden et al. (2016b), using the notion that the median is often rather robust to outliers and model departures. Ordering the ratio estimates for the k variants from smallest to largest, the simple **median estimator** takes the median (that value which divides the upper and lower half of the k estimates) as the estimate. Their **weighted median estimator** is a modification of this idea, and proceeds as follows. First, a weight is assigned to each of the estimates (e.g., using Equation A2.20b), which are standardized to sum to one, and the weights ranked from smallest to largest. Again, the median is taken, but now using the cumulative sum of the weights, with the variant whose cumulative sum is 50% has its ratio taken as the group estimate. Their simulations showed that these estimators were consistent even when up to 50% of the information comes from invalid instrumental variables. It should be noted that consistency is a large-sample feature. Nonetheless, median estimates serve as a sensitivity check, in that if the median and IVW estimates are rather similar, one has more confidence in the MR estimate, while if they radically differ, further exploration into the data is likely needed.

A number of regression-based approaches have been proposed to handle the general case of an unknown number of invalid instruments among the variants. The assumption is that any invalid instrument raises entirely from pleiotropy. From Equation A2.17a, the resulting total effect of variant j on Y becomes

$$\beta_{Yj} = \beta_{p,j} + \beta_{YX}\,\beta_{Xj} \tag{A2.21}$$

where the goal is to estimate the shared β_{YX} effect. As noted by Bowden et al. (2017, 2018), one signal of pleiotropy (assuming $\sigma^2(\beta_{p,j}) \neq 0$) is excessive heterogeneity in the ratio estimates obtained from different SNPs, which can be tested using a number of heterogeneity tests from meta-analysis, such the Cochran's Q (Equation A6.31) or Higgins-Thompson I^2 (Equation A6.37c). Bowden et al. (2019) developed improved weights to yield better performance of the Q test than under the NOME assumption weights (Equation A2.16).

Even in the presence of pleiotropy, IVW gives a consistent estimate of β_{YX} when two conditions are jointly satisfied (Bowden 2017). The first is **balanced pleiotropy**, wherein $E[\beta_p] = 0$. The second is the **InSide assumption**, for **Instrument Strength independent of direct effects** (Bowden et al. 2015). This is essentially a weaker version of the exclusion restriction assumption (the lack of path (iii) in Figure A2.8), and assumes that the magnitude (and direction) of the pleiotropic effect ($\beta_{p,j}$) is uncorrelated with the direct effect $\beta_{X,j}$ on X. Note that under balanced pleiotropy while there is no *average* effect from pleiotropy, it will result in excess variance of the individual ratio estimates, and hence can be potentially detected by a heterogeneity test.

Several approaches have been suggested to deal with the presence of **directional pleiotropy**, $E[\beta_p] \neq 0$, which attempt to estimate the common β_{YX} in Equation A2.21 under different assumptions on the distribution of $\beta_{p,j}$ (summarized in Bowden et al. 2017, Hermani et al. 2018, Verbanck et al. 2018, Slob and Burgess 2020, Zhu 2020, and Cho et al. 2021). The simplest assumption is that $\beta_{p,j}$ is the same over all variants. In this setting, the β_{YX} estimates will not show excess heterogeneity, but the presence of pleiotropy can be detected by other tests. The most common is the **Mendelian randomization-Egger** (**MR-Egger regression**) approach of Bowden et al. (2015). It is based on the Egger regression test

(Equation A6.38) from meta-analysis, where a nonzero intercept indicates the presence of a publication bias (e.g., negative test results are not published). The model becomes

$$\widehat{\beta}_{Yj} = \beta_p + \beta_{YX}\,\widehat{\beta}_{Xj} + e_j, \qquad \text{with} \quad e_j \sim N\left[0, \sigma^2(\widehat{\beta}_{Yj})\right] \tag{A2.22a}$$

The intercept term, β_p, should be zero in the presence of balanced pleiotropy (this is the **MR-Egger intercept test**). This regression accomplishes three goals: testing for the presence of directional pleiotropy ($\beta_p \neq 0$), test causality of $X \to Y$ ($\beta_{YX} \neq 0$), and (most problematic) estimating the effect β_{YX}. More generally, Equation A2.22a can still give a consistent estimate of β_{YX} when $\beta_{p,j}$ vary, provided that the InSide assumption holds (Bowden et al. 2017). In this setting, β_p is average amount of directional pleiotropy.

Turning to estimation, Bowden et al. (2015) showed that the MR-Egger estimate of β_{YX} is slightly biased, with

$$\beta_{YX,MR-Eg} = \beta_{YS}\,\left(\frac{\sigma^2(\beta_{Xj})}{\sigma^2(\widehat{\beta}_{Xj})}\right) \tag{A2.22b}$$

Because $\sigma^2(\widehat{\beta}_{Xj}) > \sigma^2(\beta_{Xj})$, the estimate of β_{YX} is shrunk towards zero (i.e., back towards the null), a phenomena known as **regression dilution bias**. Writing $\widehat{\beta}_{Xj} = \beta_{Xj} + e_k$, so that $\sigma^2(\widehat{\beta}_{Xj}) = \sigma^2(\beta_{Xj}) + \sigma^2(e_j)$, we see that if the β_{Xj} are known without error ($\sigma^2(e_j) = 0$, the NOME assumption), then the $\beta_{YX,MR-Eg}$ estimate is consistent. Bowden et al. (2015) noted that this assumption can be tested by quantifying the heterogeneity of the β_{Xj} estimates (the strength of variant j on the exposure factor X) using the Higgins-Thompson I^2 measure of heterogeneity (Equation A6.37c). As discussed in Appendix 6, I^2 is the fraction of variation due differences in effects (and hence ranges from zero to one). Bowden et al. suggested that if the I^2 statistic for the β_{Xj}, which they denote at I^2_{GX}, in the analysis exceeds 0.9, then the NOME assumption is reasonable, and the shrinkage of the β_{YX} estimate is at most 10%. An I^2_{GX} value close to one implies the variance among the true β_{Xj} values is much larger than their individual sampling errors, $\sigma^2(\beta_{Xj}) \gg \sigma^2(e_j)$. A number of outlier detection approaches have been proposed to fine-tune regression approaches, such as **HEIDI-outlier** (Zhu et al. 2018), **MR-PRESSO** (Verbanck et al. (2018), and **MR-TRYX**(Cho et al. 2020).

As stressed by several authors, both MR-Egger and median estimators should be viewed as *sensitivity analysis methods* for probing assumptions about the MR. Additional statistics have been proposed, such as the Q_R statistic of Bowden et al. (2017). This is just the ratio of the heterogeneity of data around the MR-Egger fitted slope divided by the heterogeneity of data around the IVW slope. A value of Q_R much less than one indicates that the MR-Egger is a better and should be given serious consideration. Bowden (2017) stressed the joint use of both Q_R and I^2_{GX} to determine in which cases the MR-Egger should be used, while Burgess and Thompson (2017) provide further discussion on how to interpret various MR-Egger findings.

An entirely different approach for dealing with pleiotropy is the use of penalized regressions (Example 20.4). Rees et al. (2019) proposed to allow $\beta_{p,j}$ to vary, but to penalize their total absolute sum. Under the LASSO (which shrinks most $\beta_{p,j}$ to zero, generating variable selection), the estimate is given by

$$\widehat{\beta}_{YX,L\lambda} = \min_{\beta_{p,j},\beta_{YX}} \left(\sum_{j=1}^{k} \frac{(\widehat{\beta}_{Yj} - \beta_{p,j} - \beta_{YX}\widehat{\beta}_{Xj})^2}{\sigma^2(\widehat{\beta}_{Yj})} + \lambda \sum_{j=1}^{k} |\,\beta_{p,j}\,| \right) \tag{A2.23}$$

This returns both the common factor β_{YX} as well as the pleiotropic effects of any retained variants (following model selection). One could also use the LASSO for strict model selection, performing an MR using only those variants whose $\beta_{p,j}$ have been shrunk to zero.

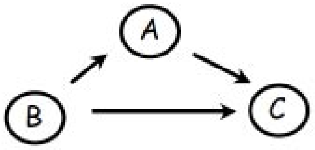


Figure A2.9 A simple directed acyclic graph (DAG). The nodes (A, B, C) indicate elements in a regulatory network, while the arrows (edges) indicate the direction of interaction.

BASIC GRAPH THEORY FOR BIOLOGICAL NETWORKS

One of the rich harvests from functional genomics are elaborate **network structures**, interactions between molecular elements, such as metabolic pathways, regulatory networks, and protein-protein maps (showing which proteins physically interact with each other). Given the rapid accumulation of such information, and its developing importance in quantitative genetics, we conclude by presenting a few basic concepts from graph theory that appear in the network literature. Barabási and Oltvai (2004), Vidal et al. (2011), and Hu et al (2016) review basic concepts with a focus on bioloigcal networks.

Consider Figure A2.9. The items in the network (A, B, and C; for example, proteins) are referred to as **nodes**, and the connections (interactions) between them are called **edges**. An arrow for an edge indicates a directional interaction, while a line indicates an interaction with no assumed directionality. In an **acyclic graph**, no pathways connect a node back to itself through a series of intermediate nodes (e.g., there are no **feedback loops**). Most path diagrams are thus **directed acyclic graphs** (or **DAGs**). We can represent the basic **topology** (shape) of a graph by a matrix **M** containing zeros and ones. In an **undirected graph** (only lines, no arrows), $M_{ij} = 1$ if nodes i and j are connected by an edge, resulting in a square, symmetric matrix (as $M_{ij} = M_{ji}$). In a directed graph, $M_{ij} = 1$ indicates that node i influences node j. In this case, $M_{ij} = M_{ji}$ only if both nodes influence each other, so that the matrix describing the graph is typically not symmetric. For the directed graph in Figure A2.9, the resulting matrix (order the columns as A, B, C) is

$$\mathbf{M} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

More generally, we can replace the elements of **M** by measures of the strengths (and signs) of interactions between nodal elements. Note that the object **M** is a multi-dimensional quantitative trait, and thus, provided we can measure it in a set of relatives, we can perform standard quantitative-genetics tasks (such as separating genetic from environmental variation). Similarly, with marker data we can also map regQTLs (Table 21.1) that impact various aspects of **M**, such as its leading eigenvalue (if we wish to extract a univariate trait), or more general features in a multivariate setting.

In order to probe the structure of a biological network, we first need to consider the features of a **random graph**, as this provides one natural null model. Random graphs, where edges are randomly connected to nodes, are often called **Erdõs-Rényi** (1959), **ER, graphs** after the two mathematicians who first considered their basic properties. Formally, we start with n nodes, with p the probability that any two randomly chosen nodes are connected. We consider several metrics for the structure of a graph and compare how the ER values differ from those seen in actual biological networks.

The first descriptor of a graph is its **degree distribution,** $P(k)$, the probability that a randomly chosen node is linked to exactly k other nodes. Under ER graphs, $P(k)$ follows a Poisson distribution (Equation 2.21a), with success parameter $z = np$,

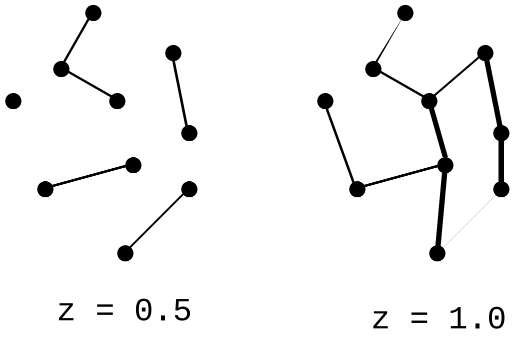


Figure A2.10 The phase transition in ER graphs at z = 1, where a giant component forms, in which a large fraction of the nodes in the graph are interconnected.

$$\Pr(X = k) = \frac{(np)^k \exp(-np)}{k!}, \quad \text{for } k = 0, 1, \cdots \tag{A2.24a}$$

Note that the probability a random node is not connected to any other ($k = 0$) is simply $\exp(-np)$, and that the number of expected connections falls off exponentially.

A second descriptor is the **path length distribution**, d, a measure of how well connected all of the nodes in the graph are to each other. Formally, this is the distribution of smallest number of steps that it takes to go from one node in the graph to any other. For an ER graph, the average path length between any two nodes is roughly given by

$$d_{ER} \sim \ln(n)/\ln(z), \quad z = np \tag{A2.24b}$$

ER graphs display a very striking **phase transition** at z = 1 (on average, one connection between any two randomly chosen nodes), forming a **giant component**, wherein a large fraction of the nodes in the graph become interconnected (Figure A2.10). The biological implication of a giant component is that most of the elements in the graph influence (or at least interact with) most other elements. Thus, the network moves from a series of discrete, non-overlapping modules into a single integrated structure. In a classic paper, Kaufmann (1969) modeled gene circuits as **random Boolean networks**. In Kaufmann's model, genes were randomly connected (formed an ER graph) and simple on/off (Boolean) rules applied (such as if your neighbor is off, you are on). Kaufmann found that such random networks showed giant regulatory components, features that appear to show *complex order* and give all the appearance of being *highly evolved*, yet they were *entirely random.* This raises the key point that ***random graphs can show considerable structure.*** Hence, ***one must be careful in inferring evolution simply because a network shows a complex, and highly integrated, structure.***

A final feature about ER graphs is that they are **small world**, wherein that the average path distance between any two nodes is short. The classic example of a small-world graph is the **Kevin Bacon conjecture**. The assertion is that every actor can be traced back to an actor who appeared in a film with Ken Bacon. The **Bacon number** for an actor is the number of actors (nodes) one must go through to find this connection. Mathematicians have a similar concept, an Erdõs number. Erdõs was very prolific, and your **Erdõs number** is how many mathematicians you must go back through to make a connection with someone who wrote a paper with Erdõs. Formally, in a small-world graph, the average path distance is close to that predicted for an ER graph, $d_{ER} \sim \ln(n)/\ln(z)$. The critical feature of small-world graphs is that they ***propagate information very efficiently.*** One can transform a highly regular graph into a small world graph by randomly rewiring a small fraction of the nodes.

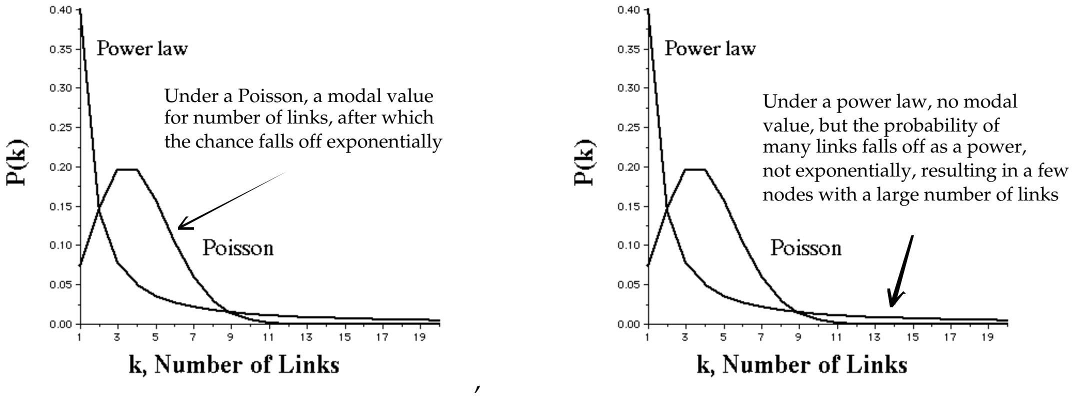


Figure A2.11 Poisson versus power law (scale-free) degree distributions, *P(k)*. Under a Poisson (left; Equation A2.24a), there is some modal number of links, after which the number of links falls off exponentially. In contrast, under a power law distribution (right; Equation A2.24c), there are a few hubs with very many connections.

How do cellular networks depart from random graphs? Most importantly, their ***degree distribution often follows a power law,***

$$P(k) \sim k^{\gamma} \tag{A2.24c}$$

This is in contrast to the Poisson distribution for k expected under an ER graphs (Figure A2.11). Graphs with a power distribution of links are called **scale-free**. In a scale-free graph, a few of the nodes will have ***very many connections***. Such nodes are often called **hubs**, where just a few locations influence a large number of nodes (such as a single eQTL influencing a large number of transcripts; Chapter 21). The important feature of scale-free graphs is that they are ***fairly robust to perturbations***: most randomly chosen nodes can be removed with little effect on the system. While removing a hub has a critical effect, the chance that a randomly removed node is a hub is small. A second feature of scale-free networks is that they are **ultra small**, meaning that their mean path length is actually *shorter* than under a random graph.

Because a scale-free structure gives a network inherent stability, biological homeostasis may just be ***a simply consequence of this structure, rather than a highly evolved feature***. How might such scale-free graphs evolve? The answer turns out to be rather simple: when we add new nodes, the scale-free feature arises when they have a ***slight preference to attach to already established nodes***. This feature is exactly what is thought to happen as gene duplication adds more nodes (elements) to a network (e.g., *Arabidopsis* Interactome Mapping Consortium 2011).

The initial view that most cellular networks appeared to be scale free has been modified somewhat (Vidal et al. 2011). While metabolic and protein-protein interaction networks tend to be scale-free, genetic regulatory networks often show a mixed behavior. With a directional graph, we can distinguish between the **incoming** and **outgoing** degree distribution, namely the average number of inputs to a node (the number of incoming arrows to a node: how many other nodes directly influence it) and the average of outputs from a node (the number of arrows leaving a node: how many other nodes it directly influences). Regulatory networks often show a scale-free outgoing distribution, meaning that there are a few hubs that can influence numerous nodes. However, the incoming degree distribution is better fitted with an exponential, implying that genes directly regulated by the input from a number of nodes are rare. Han et al (2004) observed two rather different types of hubs in biological networks: **date hubs**, whose interactions with their partners varies over time, and **party hubs**, that are highly coregulated with their partners. Vidal et al. (2011) suggested that date hubs connect different functional modules of a network with each other, whereas party hubs largely act inside functional modules, but also see Agarwal et al. (2010).

Example A2.4 Apparently well-buffered networks can often be disrupted by single-gene mutations, resulting in a dramatic increase in phenotypic variance. Rutherford and Lindquist (1998) called genes (where such mutations can arise) **phenotypic capacitors**, with the classic example being heat-shock protein Hsp90, a molecular chaperone (a protein that helps other proteins fold properly). Hsp90 knockouts in both *Drosophila* and *Arabidopsis* show increased phenotypic variance over a number of traits. One hypothesis is that capacitors are often associated with hubs in critical networks. To more broadly search for capacitors, Levy and Siegal (2008) examined phenotypic variances over a suite of morphological features in a collection of 4700 single-gene knockout lines in yeast (*Saccharomyces cerevisiae*). Roughly 300 ($\sim$ 5%) of these were capacitors, as their knockout lines showed a significant (within-line) increase in the phenotypic variance over a number of traits. Levy and Siegal found that such genes were likely to be at hubs in either networks involving physical interactions between their products (hubs in the protein-protein interaction network) or involving genetic interactions between loci (hubs in the synthetic lethal interaction network). Masel and Siegal (2009) review general issues on robustness.