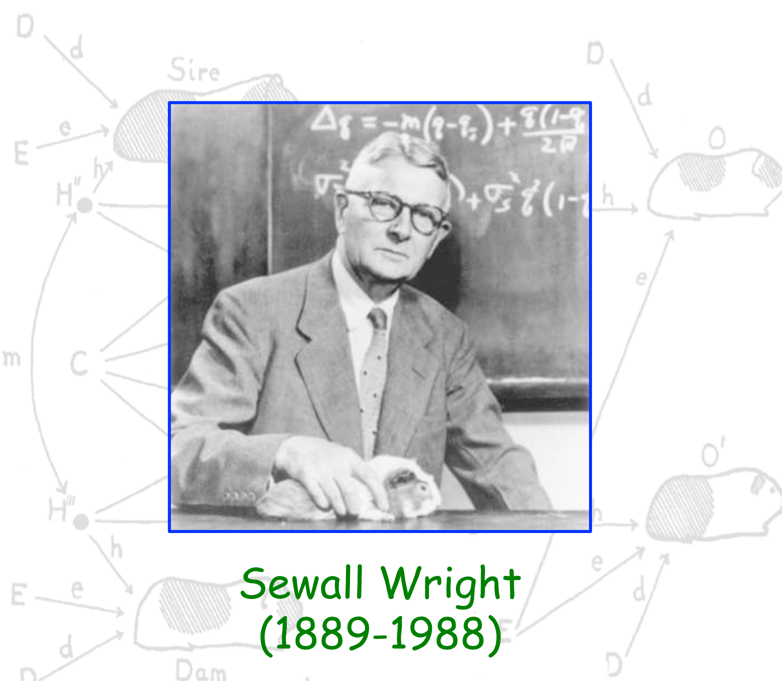
Path Analysis

Guilherme J. M. Rosa
University of Wisconsin-Madison

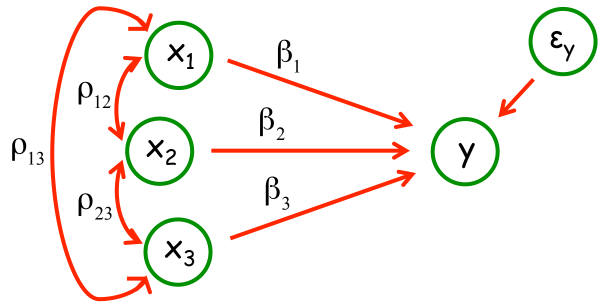


Sewall Wright
(1889-1988)

Path Analysis

(Wright, 1921)

- Interpretation of correlation between two variables in terms of hypothetical paths of causation between them
- Quantification of the relative contribution of causal sources of variance and covariance in a system of interrelated variables



Sewall Wright's Path Coefficients

→ Linear Model: $y = \mu_y + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon_y$

$$\begin{cases} x_i \sim (0, \sigma^2_{x_i}), \;\; \varepsilon_y \sim (0, \sigma^2_{\varepsilon_y}) \\ \mathrm{Cov}[x_i, x_j] = \rho_{ij}\sigma_{x_i}\sigma_{x_j} \;\; \text{and} \;\; \mathrm{Cov}[x_i, \varepsilon_y] = 0 \end{cases}$$

$$E[y] = \mu_y$$

$$\mathrm{Var}[y] = \sigma^2_y = \sum_{i=1}^{n} \beta_i^2 \sigma^2_{x_i} + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \beta_i \beta_j \rho_{ij} \sigma_{x_i} \sigma_{x_j} + \sigma^2_{\varepsilon_y}$$

$$\Rightarrow 1 = \sum_{i=1}^{n} \left(\beta_i \frac{\sigma_{x_i}}{\sigma_y}\right)^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \left(\beta_i \frac{\sigma_{x_i}}{\sigma_y}\right)\left(\beta_j \frac{\sigma_{x_j}}{\sigma_y}\right)\rho_{ij} + \frac{\sigma^2_{\varepsilon_y}}{\sigma^2_y}$$

Let: $p_{yi} = \beta_i \frac{\sigma_{x_i}}{\sigma_y}$ and $e_y = \frac{\sigma_{\varepsilon_y}}{\sigma_y}$, then:

$$1 = \sum_{i=1}^{n} p_{yi}^2 + 2\sum_{i=1}^{n}\sum_{i>j}^{n} p_{yi} p_{yj} \rho_{ij} + e_y^2$$

- Equation for "complete determination of y"

p_{yi}^2 : Fraction of variance accounted for by variation in x_j when all other variables held constant (without affecting variation in x_j)

e_y^2 : Fraction of variance not explained by model

$2p_{yi}p_{y_j}\rho_{ij}$: Fraction of variance "due to" joint determination by x_j, $x_{j'}$

$p_{yi} = \beta_i \sigma_{x_i} / \sigma_y$: is called "Path Coefficient"

Example 1. Basic Genetic Model of Quantitative Traits:

$$P = \mu + G + E \quad \begin{cases} E[P] = \mu, \ \ G \sim (0, \sigma_G^2), \ \ E \sim (0, \sigma_E^2) \\ Cov[G, E] = \sigma_{GE} = \rho_{GE}\sigma_G\sigma_E \end{cases}$$

$$Var[P] = \sigma_P^2 = \sigma_G^2 + \sigma_E^2 + 2\rho_{GE}\sigma_G\sigma_E$$

$$1 = \frac{\sigma_G^2}{\sigma_P^2} + \frac{\sigma_E^2}{\sigma_P^2} + 2\rho_{GE}\frac{\sigma_G}{\sigma_P}\frac{\sigma_E}{\sigma_P}$$

Let: $p_{PG} = \frac{\sigma_G}{\sigma_P} = h$ and $p_{PE} = \frac{\sigma_E}{\sigma_P} = e$, then:

$$1 = h^2 + e^2 + 2 \times h \times e \times \rho_{GE}$$

Example 2. Growth Analysis:

$$y_t = y_1 + y_2 + \ldots + y_n$$

Body size at time t

Initial size

Arbitrary number of subsequent growth increments

Other examples: total diet of a predator split by prey species; total seed set by a plant partitioned into contributions from various flowers

Notice: There is no residual error term; all partial regression coefficients are equal to one

$$1 = \frac{1}{\sigma_t^2}\left[\sum_{i=1}^{n}\sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n}\sigma_{ij}\right]$$

Pigeons Growth

(Lynch and Walsh, 1998, p.831-832)

Growth dynamics of a population of feral pigeons:

Weight at day 26 described as a function of initial weight (day 2) plus 4 subsequent six-day growth increments (days 2-8, 8-14, 14-20 and 20-26). Let w_j be the weight on day j, then:

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

Correlations (above diagonal) and path contributions (diagonal and below) of growth components to weight on day 26 for a sample of 100 feral pigeons (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

Some aspects of the growth properties of this population:

Very little of the w_{26} variation is accounted for by the size at birth, i.e. $(p^2_{w_2,w_{26}} = 0.014)$ - later growth components seem more important

The joint determination of pairs of growth increments are either negative of negligibly positive – compensatory growth

→ The Standardized Linear Model:

$$y = \mu_y + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon_y$$

$$\begin{cases} x_i \sim (0, \sigma^2_{x_i}), \;\; \varepsilon_y \sim (0, \sigma^2_{\varepsilon_y}) \rightarrow y \sim (\mu_y, \sigma^2_y) \\ \mathrm{Cov}[x_i, x_j] = \rho_{ij}\sigma_{x_i}\sigma_{x_j} \;\; \text{and} \;\; \mathrm{Cov}[x_i, \varepsilon_y] = 0 \end{cases}$$

$$\frac{y-\mu_y}{\sigma_y} = \beta_1 \frac{x_1}{\sigma_y} + \beta_2 \frac{x_2}{\sigma_y} + \ldots + \beta_n \frac{x_n}{\sigma_y} + \frac{\varepsilon_y}{\sigma_y}$$

$$= \left(\beta_1 \frac{\sigma_{x_1}}{\sigma_y}\right)\frac{x_1}{\sigma_{x_1}} + \ldots + \left(\beta_n \frac{\sigma_{x_n}}{\sigma_y}\right)\frac{x_n}{\sigma_{x_n}} + \left(\frac{\sigma_{\varepsilon_y}}{\sigma_y}\right)\frac{\varepsilon_y}{\sigma_{\varepsilon_y}}$$

$$y^* = p_{y1}x_1^* + p_{y2}x_2^* + \ldots + p_{yn}x_n^* + p_{y\varepsilon}\varepsilon^*$$

(Basic Linear Model in Path Analysis)

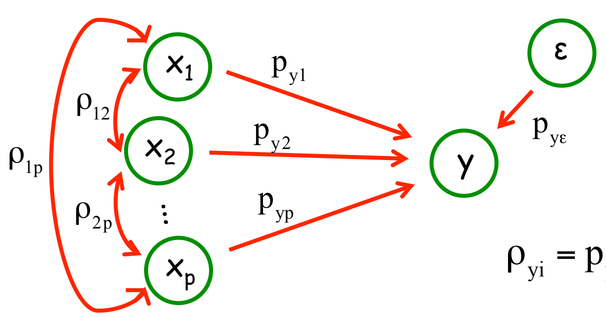
- $y^* = (y - \mu_y)/\sigma_y \rightarrow y \sim (0,1)$
- $p_{yi} = \beta_i \sigma_{x_i} / \sigma_y$ and $p_{y\varepsilon} = \sigma_{\varepsilon_y} / \sigma_y = e$
- $x_i^* = x_i / \sigma_{x_i} \rightarrow x_i^* \sim (0,1)$
- $\varepsilon^* = \varepsilon_y / \sigma_{\varepsilon_y} \rightarrow \varepsilon^* \sim (0,1)$
- $\mathrm{Cov}[x_j^*, x_{j'}^*] = \mathrm{Cov}[x_j / \sigma_{x_j}, x_{j'} / \sigma_{x_{j'}}] = \rho_{jj'}$

$$\mathrm{Var}(y^*) = 1 = p_{y1}^2 + \ldots + p_{yn}^2 + 2p_{y1}p_{y2}\rho_{12} + \ldots + 2p_{y(n-1)}p_{yn}\rho_{(n-1)n} + e^2$$

→ Path Diagrams:

The standardized linear model can be represented pictorially by a "path diagram" with cause-effect relationships depicted by straight directed arrows, and correlations by double headed curved arrows.

$$y = p_{y1}x_1 + p_{y2}x_2 + \ldots + p_{yp}x_p + p_{y\varepsilon}\varepsilon$$



$$\rho_{yi} = p_{yi} + \sum_{j \neq i} p_{yj}\rho_{ij}$$

→ Correlations Between Variables:

Simple System:

$$\begin{cases} y = \mu_y + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_y \\ z = \mu_z + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon_z \end{cases}$$

Standardized Form: $\begin{cases} y^* = p_{y1} x_1^* + p_{y2} x_2^* + e\varepsilon_y^* \\ z^* = p_{z3} x_3^* + p_{z4} x_4^* + p_{z5} x_5^* + d\varepsilon_z^* \end{cases}$

Note:

$$\mathrm{Cov}[y^*, z^*] = \mathrm{Cov}\left(\frac{y-\mu_y}{\sigma_y}, \frac{z-\mu_z}{\sigma_z}\right) = \frac{\mathrm{Cov}(y,z)}{\sigma_y \sigma_z} = \rho_{yz}$$

i.e., the covariance between two standardized variables is the correlation between them.

$$\rho_{yz} = \mathrm{Cov}[y^*, z^*]$$

$$= \mathrm{Cov}[p_{y1} x_1^* + p_{y2} x_2^* + e\varepsilon_y^*, p_{z3} x_3^* + p_{z4} x_4^* + p_{z5} x_5^* + d\varepsilon_z^*]$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{3} p_{yi} p_{zj} \mathrm{Cov}[x_i^*, x_j^*] + d \sum_{i=1}^{2} \mathrm{Cov}[x_i^*, \varepsilon_z^*]$$

$$+ e \sum_{i=3}^{5} \mathrm{Cov}[\varepsilon_y^*, x_j^*] + ed\mathrm{Cov}[\varepsilon_y^*, \varepsilon_z^*]$$

- Assuming that covariances between residuals and predictor variables are null:

$$\rho_{yz} = \sum_{i=1}^{2} \sum_{j=1}^{3} p_{yi} p_{zj} \rho_{ij} + ed \times \rho_\varepsilon$$

→ A system of correlations can be described in terms of path coefficients and of correlations between "explanatory" variables

Recursive System:

- Suppose we have the system (in standardized form; we drop the "star" from now on):

$$\begin{cases} y = p_{y1}x_1 + \varepsilon_y \\ z = p_{z1}x_1 + p_{z2}x_2 + \varepsilon_z \\ w = p_{w3}x_3 + p_{w4}x_4 + \varepsilon_w \\ x_3 = p_{31}x_1 + \varepsilon_3 \end{cases}$$

There are 6 path coefficients and 7 observables variables, with the following correlation matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{1y} & \rho_{1z} & \rho_{1w} \\ & 1 & \rho_{23} & \rho_{24} & \rho_{2y} & \rho_{2z} & \rho_{2w} \\ & & 1 & \rho_{34} & \rho_{3y} & \rho_{3z} & \rho_{3w} \\ & & & 1 & \rho_{4y} & \rho_{4z} & \rho_{4w} \\ & & & & 1 & \rho_{yz} & \rho_{yw} \\ & \text{Symm.} & & & & 1 & \rho_{zw} \\ & & & & & & 1 \end{bmatrix}$$

- Correlations can be described in terms of path coefficients and residual correlations:

$$1 = p_{y1}^2 + e_y^2$$

$$1 = p_{z1}^2 + p_{z2}^2 + 2p_{z1}p_{z2}\rho_{12} + e_z^2$$

$$\vdots$$

$$\rho_{12} = \rho_{12}$$

$$\rho_{13} = p_{31}$$

Complete determination of y, z, w and x_3

$$\rho_{2z} = p_{z1}\rho_{12} + p_{z2}$$

$$\rho_{2w} = p_{w3}\rho_{23} + p_{w4}\rho_{24}$$

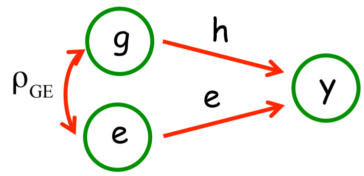
$$\vdots$$

$$\rho_{zw} = p_{z1}p_{w3}\rho_{13} + p_{z1}p_{w4}\rho_{14} + p_{z2}p_{w3}\rho_{23} + p_{z2}p_{w4}\rho_{24} + \rho_{\varepsilon_{zw}}$$

Note: We assume residuals are uncorrelated with x variables but perhaps correlated themselves.

Example 3

Relation between phenotype, genotype and environment



$$Corr(g,y) = \rho_{gy} = h + \rho_{GE} e$$

$$Corr(e,y) = \rho_{ey} = e + \rho_{GE} h$$

Example 4

Genetic, environmental and phenotypic correlations

$$X = \mu_x + G_x + E_x \rightarrow x = h_x g_x + e_x \varepsilon_x$$

$$Y = \mu_y + G_y + E_y \rightarrow y = h_y g_y + e_y \varepsilon_y$$

$$Cov(X,Y) = Cov(G_x,G_y) + Cov(G_x,E_y) + Cov(E_x,G_y) + Cov(E_x,E_y)$$

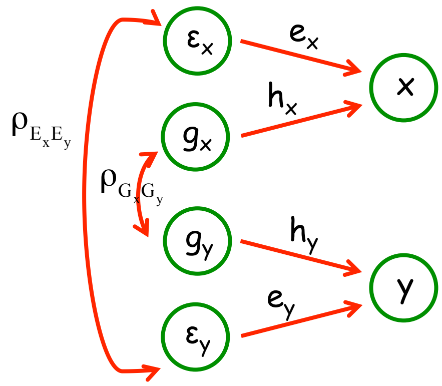
Usually assumed to be null

$$Cov(X,Y) = \rho_{G_xG_y} \sigma_{G_x} \sigma_{G_y} + \rho_{E_xE_y} \sigma_{E_x} \sigma_{E_y}$$

$$Corr(X,Y) = \rho_{G_xG_y} \frac{\sigma_{G_x}}{\sigma_X} \frac{\sigma_{G_y}}{\sigma_Y} + \rho_{E_xE_y} \frac{\sigma_{E_x}}{\sigma_X} \frac{\sigma_{E_y}}{\sigma_Y}$$

$$= \rho_{G_xG_y} h_x h_y + \rho_{E_xE_y} \sqrt{(1-h_x^2)(1-h_y^2)}$$

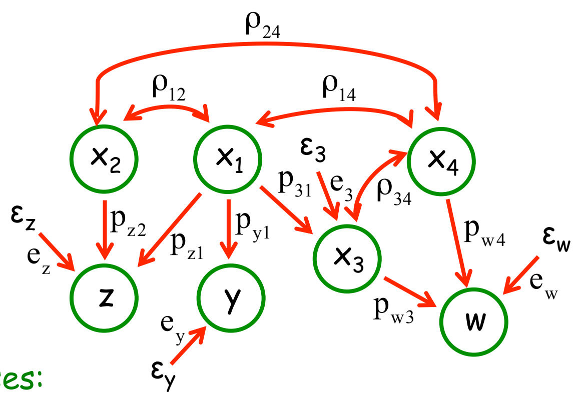
Using path analysis:



$$\mathrm{Cov}(X,Y) = e_x \rho_{E_xE_y} e_y + h_x \rho_{G_xG_y} h_y$$
$$= \rho_{G_xG_y} h_x h_y + \rho_{E_xE_y} e_x e_y$$

Example 5

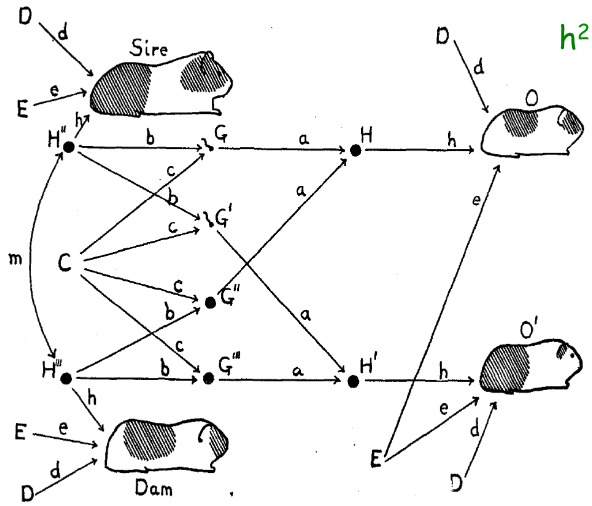
Recursive system, assuming no residual correlations



Some cases:

$$\rho_{zy} = p_{z2}\rho_{12}p_{y1} + p_{z1}p_{y1} \qquad \rho_{z2} = p_{z2} + \rho_{12}p_{z1}$$

$$\rho_{z3} = p_{z2}\rho_{12}p_{31} + p_{z1}p_{31} \qquad \rho_{zw} = p_{z2}\rho_{12}p_{31}p_{w3} + p_{z2}\rho_{24}p_{w4}$$
$$+ p_{z1}p_{31}p_{w3} + p_{z1}\rho_{14}p_{w4}$$



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

A diagram illustrating the relations between two mated individuals and their progeny. *H*, *H'*, *H''* and *H'''* are the genetic constitutions of the four individuals. *G*, *G'*, *G''* and *G'''* are four germ-cells. *E* and *D* represent tangible external conditions and chance irregularities as factors in development. *C* represents chance at segregation as a factor in determining the composition of the germ-cells. Path coefficients are represented by small letters. (Wright, 1921)