2. Bivariate analysis

I. Background to bivariate analysis

A. Strength of correlation vs. form of relationship

B. Statistical significance vs. effect size

C. Causal relationship vs. joint dependence on additional factor(s)

II. Bivariate measures of correlation or association

A. Correlation coefficient, $r$

1. $r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$

2. $r^2$ rather than $r$ measures scatter in data
   - Consider bivariate normal ellipse of variation and circle with diameter equal to major axis of ellipse. $(1 - r^2)$ gives the ratio of the area of the ellipse to the area of the circle.
   - Ratio of major to minor axes is square root of ratio of eigenvalues, which will make sense when we get to principal components, etc.

3. Assumptions for $r$ to be a "good" measure of association
   a. relationship linear
   b. distribution bivariate normal (or at least symmetric)
   c. $r$ not pulled by outliers
   d. (tacit in the foregoing): residuals uncorrelated with $X$, $Y$, and each other

B. See Siegel and Castellan for analogues on nominal or ordinal data

III. Purposes of line fitting (and associated assumptions)

A. Bivariate linear model: $Y = aX + b + E$

1. $a =$ slope
2. $b =$ intercept
3. $E =$ error
4. Line passes through bivariate mean $(\bar{X}, \bar{Y})$
5. $\sum E = 0$
6. Once $a$ is known (see below), $b = \bar{Y} - a\bar{X}$

B. Prediction of $Y$ from $X$: Least squares linear regression (LSLR)

1. Assume $X$ known without error; minimize squared residuals ($E$) in $Y$
2. Writing expression for $\sum E^2$, taking first derivative with respect to $a$ and with respect to $b$, and setting first derivatives to zero to find minimum leads to:
   
   $a = \frac{\text{cov}(X,Y)}{\text{var}(X)} = r \cdot \frac{s_Y}{s_X}$
   
   • We'll cover more general problem of least squares minimization when we treat simultaneous linear equations later.

3. $a$ underestimates true slope if there is error in $X$
4. NB: $a_{yx} \neq a_{xy}^{-1}$

C. Description of mutual dependence: Reduced major axis (RMA)

1. Assume both $X$ and $Y$ known with error; minimize squared diagonal residuals. This leads to:
   
   $a = \text{sign}(r) \cdot \frac{s_Y}{s_X}$
   
   NB: $a_{yx} = a_{xy}^{-1}$

D. Poor (linear) correlation: Noise in data vs. inappropriate model
IV. Partial Correlation
A. Rationale: Relationship between $X$ and $Y$ may be obscured or exaggerated by mutual dependence on additional variables.
B. Goal: Determine relationship between $X$ and $Y$ with other variables held (statistically) constant. (I.e. substitute statistical control for experimental control.)
C. Calculation:

$$r_{xy} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

- Equivalent to correlation between LSLR residuals of $y$ on $z$ and those of $x$ on $z$
D. Like raw correlations, this measure of partial correlation assumes linear model.