Regression

• Bivariate linear regression:
  – Estimation of the *linear function* (straight line) describing the *linear component* of the *joint relationship* between two variables $X$ and $Y$.
    • Generally describe $Y$ as a function of $X$.
  – Covariance and *correlation* measure *strength* of the relationship between $X$ and $Y$.
    • Covariance depends on units of $X$ and $Y$.
    • Correlation is dimensionless.
  – Thus regression line and correlation (or covariance) provide *complementary information* about joint distribution.
• Linear components of relationships:
  – Represented by linear equations (involving $X^I$):
    • Two parameters (pieces of information) needed to uniquely specify a line:
      – Two points
      – One point and slope

  – Two comparable formulations:
    • Geometric: $Y = mX + b$
    • Statistical: $Y = b_0 + b_1X$
(1) **Geometric formulation:** \( Y = mX + b \)

- \( m = \text{slope of line (change in } Y \text{ relative to change in } X \). \)
  - Can be determined from any two points \((X_i, Y_i)\) and \((X_j, Y_j)\) lying on the line:
    \[
    m = \frac{\Delta Y}{\Delta X} = \frac{Y_j - Y_i}{X_j - X_i} = \frac{Y_i - Y_j}{X_i - X_j}
    \]

- \( b = \text{“}Y\text{-intercept”}, \text{ value of } Y \text{ when } X = 0. \)
  - If know slope \((m)\), can find intercept.
  - Substitute it, and the coordinates of any point on the line, into the equation, solve for \( b \):
    \[
    m = \frac{\Delta Y}{\Delta X} = \frac{Y_i - 0}{X_i - b}
    \]
  - In particular: \( b = \bar{Y} - m\bar{X} \)
  - In particular: \( b = Y_i - mX_i \)
(2) **Statistical formulation:** \( Y = b_0 + b_1 X \)
- \( b_0 \) = \( Y \)-intercept
- \( b_1 \) = slope
- This form extrapolates to polynomials of higher order:
  \[ Y = b_0 + b_1 X + b_2 X^2 + \ldots + b_k X^k \]
- Statistics \( b_0 \) and \( b_1 \) are assumed to be sample estimates of the population parameters \( \beta_0 \) and \( \beta_1 \).
  - Estimated with error,
  - From a sample of individuals that don’t lie exactly on a straight line:
    - Express individual “residual” variation.
      - Doesn’t fit the linear model.
    - Comprises two sources of variation:
      - (a) Individual biological variation.
      - (b) Measurement error.
Response (dependent) variable

‘True’ slope

Residual variation

Mean response

‘True’ linear relationship

Mean residual

‘True’ intercept

Predictor (independent) variable
• There are an infinite number of different regression models.
  – Correspond to differing assumptions made about the ‘direction’ of residual variation.

• Two most common models:
  (1) Predictive (type I).
  (2) Major axis (type II).
Predictive regression

(1a) **Predictive regression of Y on X:**

- The ‘true’ population relationship is assumed to be:
  \[
  Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \\
  Y_i - \varepsilon_i = \beta_0 + \beta_1 X_i
  \]

- All *residual variation* is assumed to be concentrated in Y, and none in X (vertical residuals).

- Arises from analysis of variance (ANOVA):
  - Purpose: test whether means for two or more groups (k groups) differ “significantly”.
  - Groups (X) assumed to be known, with no error.
  - Within-group residual variation (in Y) assumed to be normally distributed.
  - All groups assumed to have the same amount of variation.
Step from *ANOVA* to *regression*:

- Assume that groups can be represented by the (exact) value of a *predictor (independent)* variable, $X$.
- Ask whether there is a *linear trend in response* ($Y$) among groups.
- Appropriate in experimental studies in which predictor variable is set by the investigator, with no error.
Thus $X$ and $Y$ have different roles in the regression analysis:

- $X$ is the \textit{predictor} or \textit{independent} variable.
  - Assigned values independently of the model, with no error.
- $Y$ is the \textit{response} or \textit{dependent} variable.
  - Assumed to be dependent on the value of $X$, with some individual (residual) variation in response.

Estimation: $b_1 = \frac{S_{XY}}{S_X^2}$, \quad $b_0 = \bar{Y} - b_1 \bar{X}$
(1b) **Predictive regression of $X$ on $Y$:**

- Reverse role of $X$ and $Y$.
- ‘True’ population relationship is then:

\[
X_i = \delta_0 + \delta_1 Y_i + \eta_i \\
X_i - \eta_i = \delta_0 + \delta_1 Y_i
\]

- All *residual variation* is assumed to be concentrated in $X$, and none in $Y$ (=horizontal residuals).
- Estimation: 
  \[
d_1 = \frac{S_{XY}}{S_Y^2}, \quad d_0 = \bar{X} - d_1 \bar{Y}
\]
Major-axis regression

(2) **Major axis regression:**

- The two predictive-regression models above are extreme cases:
  
  - Predictive regression of $Y$ on $X$ assumes that $\text{var}(\eta_i)=0$; i.e., that all residual variation is in $Y$.
  
  - Predictive regression of $X$ on $Y$ assumes that $\text{var}(\epsilon_i)=0$; i.e., that all residual variation is in $X$.

- Instead of assuming that all residual variation (‘error’) is in either $X$ or $Y$, assume that both variables have some residual variation.

- If knew amounts of residual variation, could express their *relative errors* as a ratio:

$$\lambda = \frac{\text{var}(\epsilon_i)}{\text{var}(\eta_i)}$$
• So: \( \lambda = \frac{\text{var}(\varepsilon_i)}{\text{var}(\eta_i)} \)

• For any value of \( \lambda \), we can estimate the corresponding slope of a unique regression line:

\[
m = \frac{s_Y^2 - \lambda s_X^2 + \sqrt{(s_Y^2 - \lambda s_X^2)^2 + 4\lambda s_{XY}^2}}{2s_{XY}}
\]

\[
b = \bar{Y} - m \bar{X}
\]

• This general case is called \textit{functional regression}. 
• There are an infinite number of functional-regression lines, corresponding to values of \( \lambda \).

• In general, don’t know how much residual variation is in \( X \) versus \( Y \).
  – If treating \( X \) and \( Y \) equivalently, reasonable assumption: they have *approximately equal* amounts of residual variation: \( \text{var}(\varepsilon_i) = \text{var}(\eta_i) \)

\[
\lambda = \frac{\text{var}(\varepsilon_i)}{\text{var}(\eta_i)} = 1
\]

  – Then slope reduces to: \( m = \frac{s_Y^2 - s_X^2 + \sqrt{(s_Y^2 - s_X^2)^2 + 4s_{XY}^2}}{2s_{XY}} \)

  – This special case is *major-axis regression*:
    • Regression line equivalent to *major axis* of the *confidence ellipse* for the data.
• **Major-axis regression line** is always intermediate in slope between predictive regression of $Y$ on $X$ and that of $X$ on $Y$.
  – More precisely: major axis slope is *geometric mean* of the two predictive-regression slopes:

$$m = \sqrt[1]{b_1 / d_1}$$

($d_1$ is converted to reciprocal so that all three slopes describe change in $Y$ relative to change in $X$.)

Predictive and major-axis regressions

\[ r = 0.99 \]

\[ r = 0.95 \]

\[ r = 0.90 \]

\[ r = 0.70 \]
Predictive and major-axis regressions

$r = 0.50$

$r = 0.30$

$r = 0.20$

$r = 0.00$

X on Y

MA

Y on X
• Major-axis regression is the \textit{bivariate} case of \textit{principal component analysis}.  
  – Corresponds to case in which \textit{residuals} are assumed to be \textit{orthogonal} to the regression line.  
  – Can be extended to >2 variables.