Lecture 8: Correlation and Intro to Linear Regression

Sandy Eckel
seckel@jhsph.edu

5 May 2008

Quantifying association

Goal: Express the strength of the relationship between two variables

- Metric depends on the nature of the variables
- For now, we'll focus on continuous variables
  (e.g. height, weight)
- Important! **association does not imply causation**

To describe the relationship between two continuous variables, use:

- Correlation analysis
  - Measures *strength* and *direction* of the linear relationship between two variables
- Regression analysis
  - Concerns prediction or estimation of outcome variable, based on value of another variable (or variables)

Correlation analysis

- Plot the data (or have a computer to do so)
- Visually inspect the relationship between two continous variables
- Is there a linear relationship (correlation)?
- Are there outliers?
- Are the distributions skewed?

Correlation Coefficient I

- Measures the strength and direction of the **linear** relationship between two variables X and Y
- Population correlation coefficient:

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2] \cdot \text{E}[(Y - \mu_Y)^2]}}$$

- Sample correlation coefficient:
  (obtained by plugging in sample estimates)

$$r = \frac{\text{sample cov}(X, Y)}{\sqrt{s_x^2 \cdot s_Y^2}} = \frac{\sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \cdot \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}}}$$

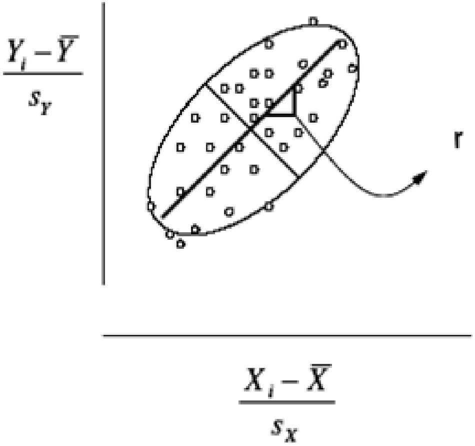
Correlation Coefficient II

The correlation coefficient, ρ, takes values between -1 and $+1$

- -1: Perfect negative linear relationship
- 0: No linear relationship
- $+1$: Perfect positive relationship

Correlation Coefficient III

- Plot standardized Y versus standardized X
- Observe an ellipse (elongated circle)
- Correlation is the slope of the major axis

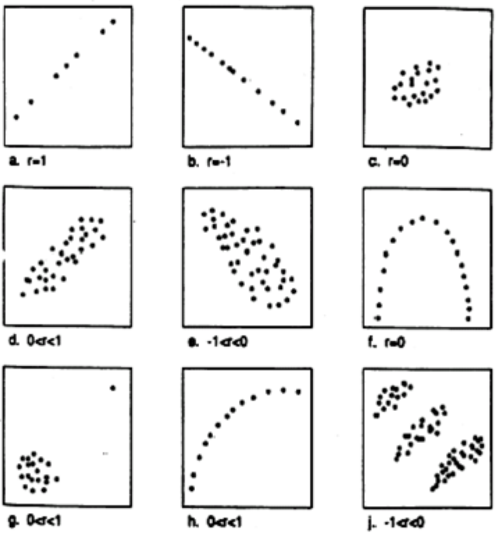


Correlation Notes

- Other names for r
  - Pearson correlation coefficient
  - Product moment of correlation

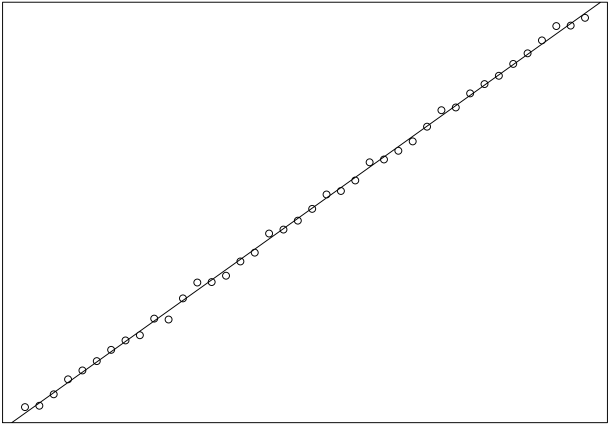
- Characteristics of r
  - Measures *linear* association
  - The value of r is independent of units used to measure the variables
  - The value of r is sensitive to outliers
  - r^2 tells us what proportion of variation in Y is explained by linear relationship with X

Several levels of correlation



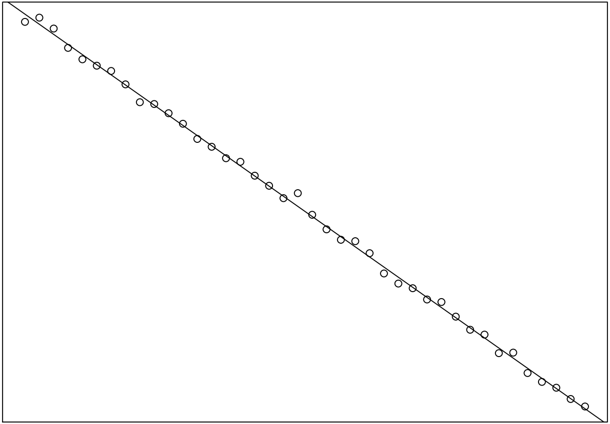
Examples of the Correlation Coefficient I

Perfect positive correlation, r $\approx$ 1



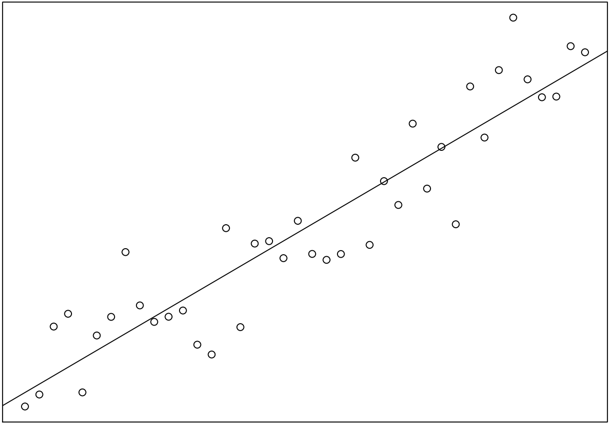
Examples of the Correlation Coefficient II

Perfect negative correlation, r ≈ -1



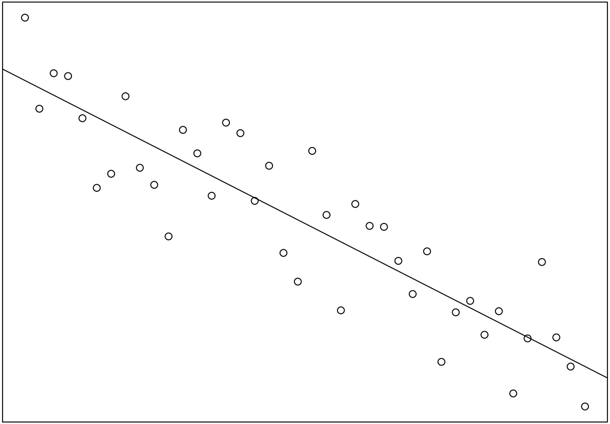
Examples of the Correlation Coefficient III

Imperfect positive correlation, $0 < r < 1$



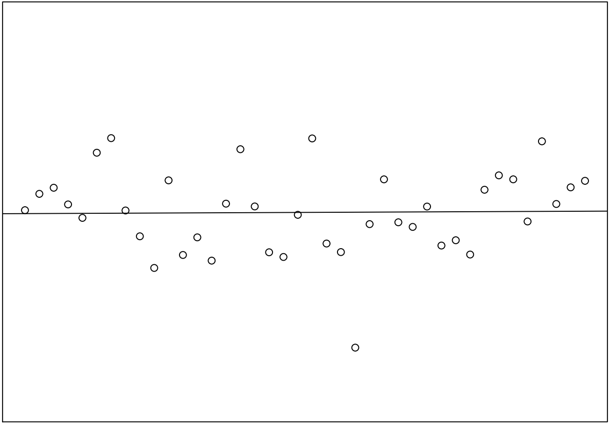
Examples of the Correlation Coefficient IV

Imperfect negative correlation, $-1 < r < 0$



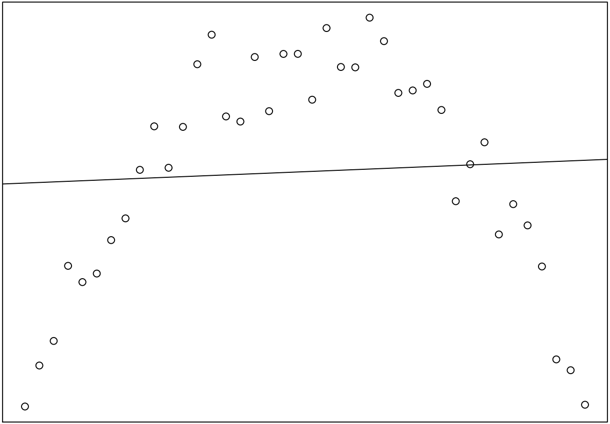
Examples of the Correlation Coefficient V

No relation, r $\approx$ 0



Examples of the Correlation Coefficient VI

Some relation but little *linear* relationship, r $\approx$ 0



Association and Causality

- In general, association between two variables means there is some form of relationship between them
  - The relationship is not necessarily causal
  - Association does not imply causation, no matter how much we would like it to
- Example: Hot days, ice cream, drowning

Sir Bradford Hill's Criteria for Causality

- Strength: magnitude of association
- Consistency of association: repeated observation of the association in different situations
- Specificity: uniqueness of the association
- Temporality: cause precedes effect
- Biologic gradient: dose-response relationship
- Biologic plausibility: known mechanisms
- Coherence: makes sense based on other known facts
- Experimental evidence: from designed (randomized) experiments
- Analogy: with other known associations

Simple Linear Regression (SLR): Main idea

Linear regression can be used to study a continuous outcome variable as a linear function of a predictor variable

Example: 60 cities in the US were evaluated for numerous characteristics, including:

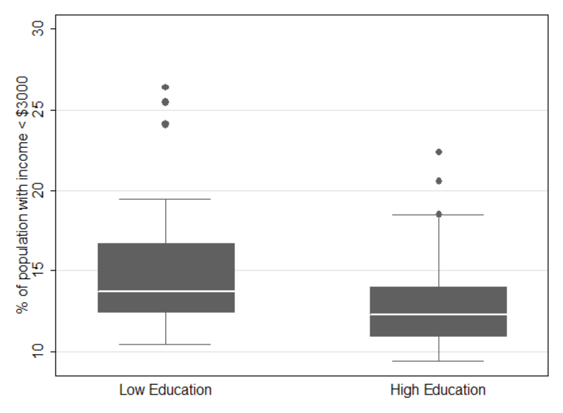
Outcome variable (y) the % of the population with low income

Predictor variable (x) median education level

Linear regression can help us to model the association between median education and % of the population with low income

Example: Boxplot of % low income by education level

Education level is coded as a binary variable with values 'low' and 'high'



Simple linear regression, t-tests and ANOVA

- Mean in low education group: 15.7%
- Mean in high education group: 13.2%

The two means could be compared by a t-test or ANOVA, but regression provides a unified equation:

$$
\begin{aligned}
\hat{y}_i &= \beta_0 + \beta_1 x_i \\
\hat{y}_i &= 15.7 - 2.5 x_i
\end{aligned}
$$

where

- $x_i = 1$ for high education and 0 for low education (x is called a dummy variable or indicator variable that designates group)
- $\hat{y}_i$ is our estimate of the mean % low income for the given the value of education
- what about the β's?

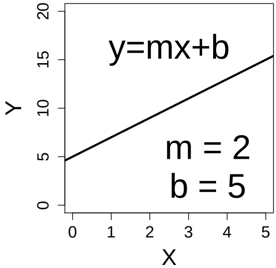
Review: equation for a line

Recall that back in geometry class, you learned that a line could be represented by the equation

$$y = mx + b$$

where

$$\begin{aligned} m &= \text{slope of the line (rise/run)} \\ b &= \text{y-intercept (value y when x=0)} \end{aligned}$$



Regression analysis represented by equation for a line

In simple linear regression, we use the equation for a line

$$y = mx + b$$

but we write it slightly differently:

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\begin{aligned} \beta_0 &= \text{y-intercept (value y when x=0)} \\ \beta_1 &= \text{slope of the line (rise/run)} \end{aligned}$$

Example: the model components

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$
$$\hat{y}_i = 15.7 + (-2.5) x_i$$

- $\hat{y}_i$ is the predicted mean of the outcome y_i for x_i
- β_0 is the intercept, or the value of $\hat{y}_i$ when $x_i = 0$
- β_1 is the slope, or the change in $\hat{y}_i$ for a 1 unit increase in $x_i = 0$
- x_i is the indicator variable of low or high education for observation i

Example: fill in covariate value to help interpretation

$$\begin{aligned}\hat{y}_i &= \beta_0 + \beta_1 x_i \\ \hat{y}_i &= 15.7 - 2.5 x_i\end{aligned}$$

- $x_i = 0$ (low education)

$$\begin{aligned}\hat{y}_i &= 15.7 - 2.5 \times 0 \\ &= 15.7 = \beta_0\end{aligned}$$

- $x_i = 1$ (high education)

$$\begin{aligned}\hat{y}_i &= 15.7 - 2.5 \times 1 \\ &= 13.2 = \beta_0 + \beta_1\end{aligned}$$

Interpretations

Intercept

- β_0 is the mean outcome for the **reference group**, or the group for which $x_i = 0$.
- Here, β_0 is the average percent of the population that is low income for cities with low education.

Slope

- β_1 is the **difference** in the mean outcome between the two groups (when $x_i = 1$ vs. when $x_i = 0$)
- Here, β_1 is **difference** in the average percent of the population that is low income for cities with high education compared to cities with low education.

Why use linear regression?

Linear regression is very powerful. It can be used for many things:

- Binary X
- Continuous X
- Categorical X
- Adjustment for confounding
- Interaction
- Curved relationships between X and Y

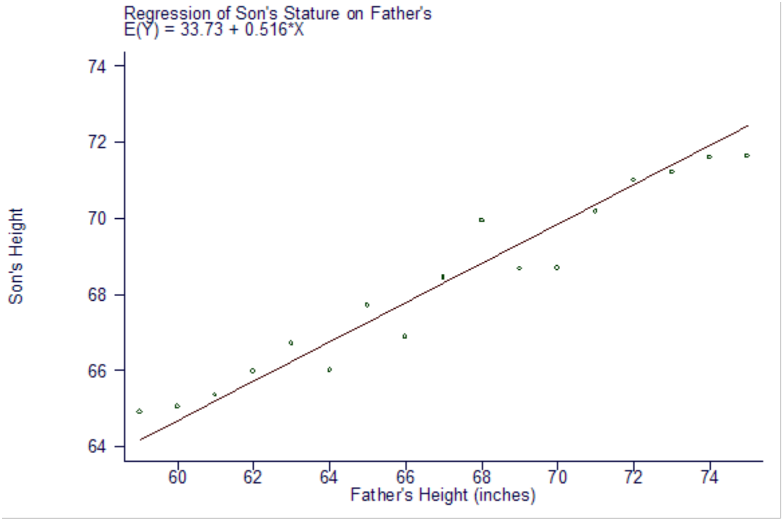
Regression analysis

- A regression is a description of a response measure, **Y**, the **dependent variable**, as a function of an explanatory variable, **X**, the **independent variable**.
- **Goal:** prediction or estimation of the value of one variable, Y, based on the value of the other variable, X.
- A simple relationship between the two variables is a **linear relationship** (straight line relationship)
- Other names: linear, simple linear, least squares regression

Foundational example: Galton's study on height

- 1000 records of heights of family groups
- Really tall fathers tend on average to have tall sons but not quite as tall as the really tall fathers
- Really short fathers tend on average to have short sons but not quite as short as the really short fathers
- There is a regression of a sons height toward the mean height for sons

Example: Galton's data and resulting regression



Regression analysis: population model

- Probability model: Independent responses $y_1, y_2, \ldots, y_n$ are sampled from

$$y_i \sim N(\mu_i, \sigma^2)$$

- Systematic model: $\mu_i = E(y_i|x_i) = \beta_0 + \beta_1 x_i$
  where

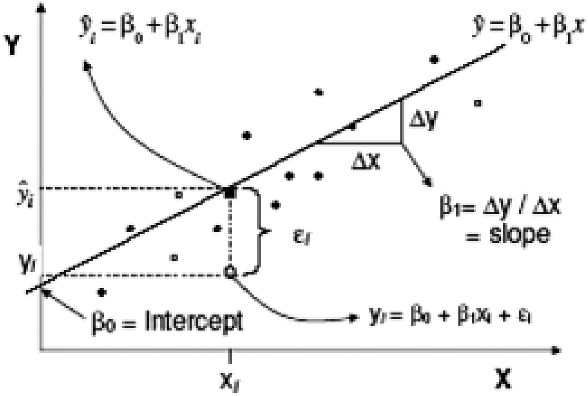
$$\begin{aligned} \beta_0 &= \mathit{intercept} \\ \beta_1 &= \mathit{slope} \end{aligned}$$

Another way to write the model

- Systematic: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Probability (random): $\epsilon_i \sim N(0, \sigma^2)$

- The response y_i is a linear function of x_i plus some random, normally distributed error, ϵ_i
- data $=$ signal $+$ noise

Geometric interpretation



Remember: two (equivalent) ways to write the model

- Probability: $y_i \sim N(\mu_i, \sigma^2)$
- Systematic: $\mu_i = E(y_i|x_i) = \beta_0 + \beta_1 x_i$
  where

$$\begin{aligned} \beta_0 &= \textit{intercept} \\ \beta_1 &= \textit{slope} \end{aligned}$$

OR

- Systematic: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Probability: $\epsilon_i \sim N(0, \sigma^2)$

The response, y_i, is a linear function of x_i plus some random, normally distributed error, ϵ_i

Interpretation of coefficients

- Intercept (β_0)
  Mean model: $\mu = E(y|x) = \beta_0 + \beta_1 x$
  - $\beta_0 =$ expected response when $x = 0$
  - Since $E(y|x=0) = \beta_0 + \beta_1(0) = \beta_0$

- Slope (β_1)
  $\beta_1 =$ change in expected response per 1 unit increase in x

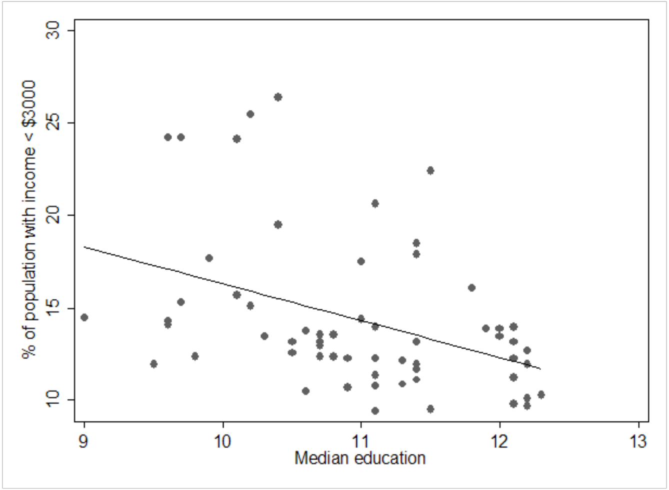
$$
\begin{array}{lll}
\text{Since:} & E(y|x+1) & = \beta_0 + \beta_1(x+1) \\
\text{And:} & \underline{E(y|x)} & = \underline{\beta_0 + \beta_1 x} \\
\multicolumn{2}{r}{\Delta E(y) \text{ from x to x+1} =} & \beta_1
\end{array}
$$

From Galton's example

- $E(y|x) = \beta_0 + \beta_1 x$
- $E(y|x) = 33.7 + 0.52x$
  where: y = son's height (inches)
  x = father's height (inches)
- Expected son's height = 33.7 inches when father's height is 0 inches
- Expected difference in heights for sons whose fathers' heights differ by one inch = 0.52 inches

City education/income example

When education is a continuous variable (not binary)



City education/income model

Using the continuous variable for median education in city i (x_i):

$$\begin{aligned} E(y_i|x_i) &= \beta_0 + \beta_1 x_i \\ E(y_i|x_i) &= 36.2 - 2.0x_i \end{aligned}$$

When $x_i = 0$

$$\begin{aligned} E(y_i|x_i) &= 36.2 - 2.0(0) \\ &= 36.2 = \beta_0 \end{aligned}$$

When $x_i = 1$

$$\begin{aligned} E(y_i|x_i) &= 36.2 - 2.0(1) \\ &= 34.2 = \beta_0 + \beta_1 \end{aligned}$$

When $x_i = 2$

$$\begin{aligned} E(y_i|x_i) &= 36.2 - 2.0(2) \\ &= 32.2 = \beta_0 + \beta_1 \times 2 \end{aligned}$$

City education/income model interpretation

Intercept (β_0)

- β_0 is the mean outcome for the reference group, or the group for which $x_i = 0$.
- Here, β_0 is the average percent of the population that is low income for cities with median education level of 0.

Slope (β_1)

- β_1 is the difference in the mean outcome for a one unit change in x.
- Here, β_1 is difference in the average percent of the population that is low income between two cities, when the first city has 1 unit higher median education level than the second city.

Finding β's from the graph

- β_0 is the y-intercept of the line, or the average value of y when $x = 0$.
- β_1 is the slope of the line, or the average change in y per unit change in x.

$$y = mx + b$$

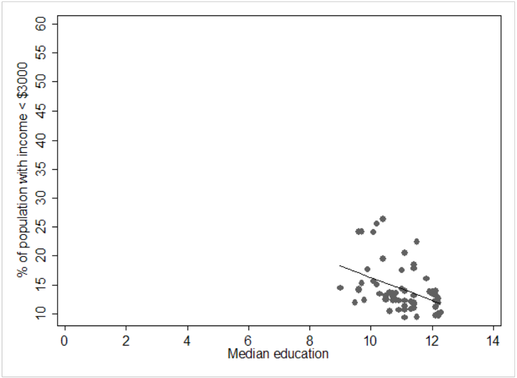
$$b = \beta_0, \quad m = \beta_1$$

$$\hat{\beta}_1 = \frac{rise}{run} = \frac{y_1 - y_2}{x_1 - x_2}$$

Note on notation:

- β_1 represents the **true** slope (in the population)
- $\hat{\beta}_1$ (or b_1) is the sample estimate of the slope

Where is our intercept?



The intercept isn't in the range of our observed data. This means:

- The intercept isn't very interpretable since the average of y when $x = 0$ was never observed
- Possible solution: we might want to *center* our x variable

Summary

Today we've discussed

- Correlation
- Linear regression with continuous y variables
- Simple linear regression (just one x variable)
  - Binary x ('dummy' or 'indicator' variable for group)
    β_1: mean difference in outcome between groups
  - Continuous x
    β_1: mean difference in outcome corresponding to a 1-unit increase in x
- Interpretation of regression coefficients (intercept and slope)
- How to write the regression model (2 ways)

Next time we'll discuss multiple linear regression
(more than one x variable) and confounding