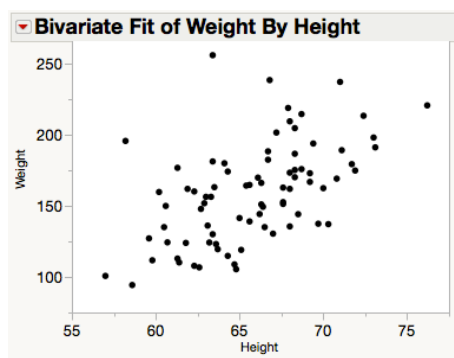
MTH 445/545
Linear Regression and Correlation

- Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others.
- Some Examples:
  - Height and weight of people
  - Income and expenses of people
  - Production size and production time
  - Soil pH and the rate of growth of plants

1

Correlation

- An easy way to determine if two quantitative variables are linearly related is by looking at their scatterplot.
- Another way is to calculate the correlation coefficient, denoted usually by r.
- The *Linear Correlation* measures the strength of the linear relationship between explanatory variable (x) and the response variable (y). An estimate of this correlation parameter is provided by the Pearson sample correlation coefficient, r.

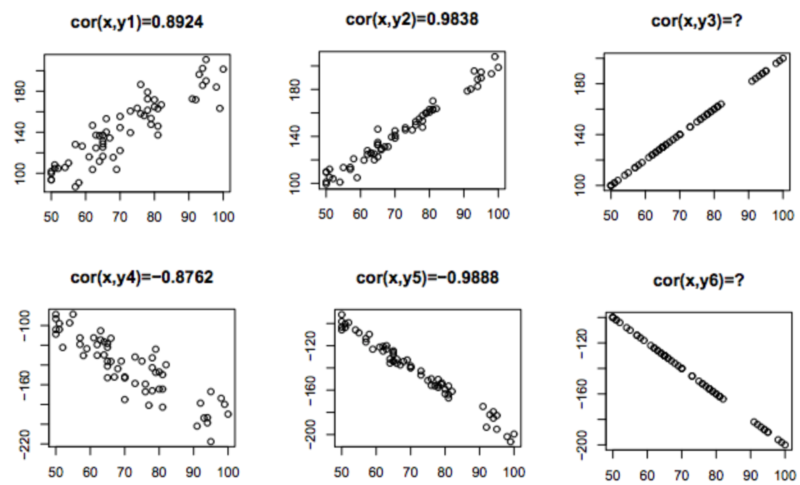


$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Note: $-1 \le r \le 1$.

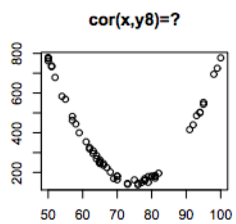
2

Example Scatterplots with Correlations



If X and Y are independent, then their correlation is 0.

Correlation



If the correlation between X and Y is 0, it doesn't mean they are independent. It only means that they are not linearly related.

One complain about the correlation is that it can be subjective when interpreting its value. Some people are very happy with r≈0.6, while others are not.

Note: Correlation does not necessarily imply Causation!

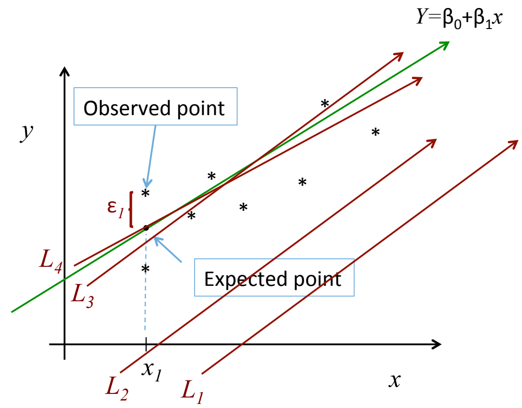
- Some Guidelines in Interpreting r.

Value of |r|	Strength of linear relationship
If |r|≥ 0.95	Very Strong
If 0.85 ≤ |r|< 0.95	Strong
If 0.65 ≤ |r|< 0.85	Moderate to Strong
If 0.45 ≤ |r|< 0.65	Moderate
If 0.25 ≤ |r|< 0.45	Weak
If|r|< 0.25	Very weak/Close to none

Simple Linear Regression

- Model: $Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$ ← *Random Error*

  where,
  - Y_i is the i^{th} value of the response variable.
  - x_i is the i^{th} value of the explanatory variable.
  - ε_i's are uncorrelated with a mean of 0 and constant variance σ^2.
- How do we determine the underlying linear relationship?
- Well, since the points are following this linear trend, why don't we look for a line that "best" fit the points.
- But what do we mean by "best" fit? We need a criterion to help us determine which between 2 competing candidate lines is better.



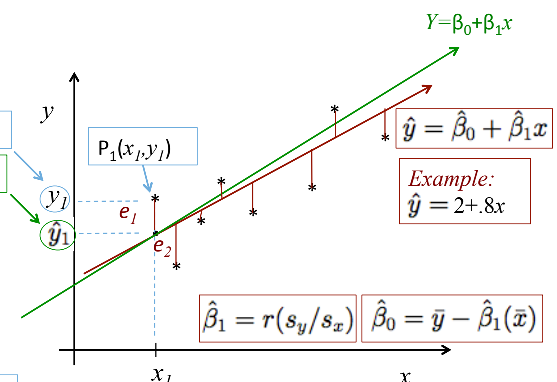
Method of Least Squares

- Model: $Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$

  where,
  - Y_i is the i^{th} value of the response variable.
  - x_i is the i^{th} value of the explanatory variable.
  - ε_i's are uncorrelated with a mean of 0 and constant variance σ^2.
    - Residual = (Observed y-value) – (Predicted y-value)

      → $e_1 = y_1 - \hat{y}_1$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad = f(\hat{\beta}_0, \hat{\beta}_1)$$



Method of Least Squares: *Choose the line that minimizes the SSE as the "best" line. This line is unknown as the Least-Squares Regression Line.*

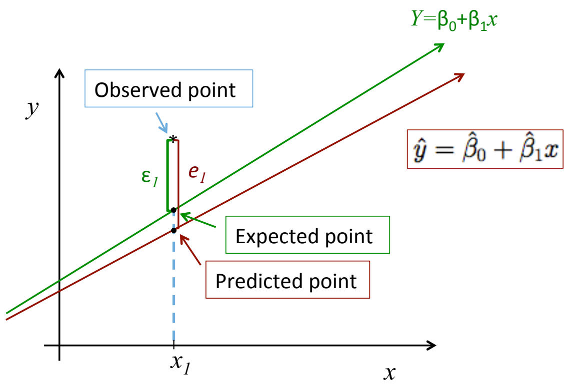
Question: *But there are infinite possible candidate lines, how can we find the one that minimizes the SSE?*

Answer: *Since SSE is a continuous function of 2 variables, we can use methods from calculus to minimize the SSE.*

Model Assumptions

- Model: $Y_i=(\beta_0+\beta_1 x_i) + \varepsilon_i$

  where,
  - ε_i's are uncorrelated with a mean of 0 and constant variance σ^2_ε.
  - ε_i's are normally distributed. (*This is needed in the test for the slope.*)



Since the underlying (green) line is unknown to us, we can't calculate the values of the error terms (ε_i). The best that we can do is study the residuals (e_i).

7