

# Correlation and Regression

Math 130: Introductory Statistics

## 1

### Goals of this section

- Learn about how to detect linear relationships between variables.
- Learn about how to estimate the relationship between many variables.

## 2 Correlation

### 2.1 Correlation coefficient

#### Correlation

- A **correlation** exists between two variables when one of them is related to the other in some way.
- The **linear correlation coefficient** is a measure of the strength of the linear relationship between two variables.

$$\text{Population: } \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

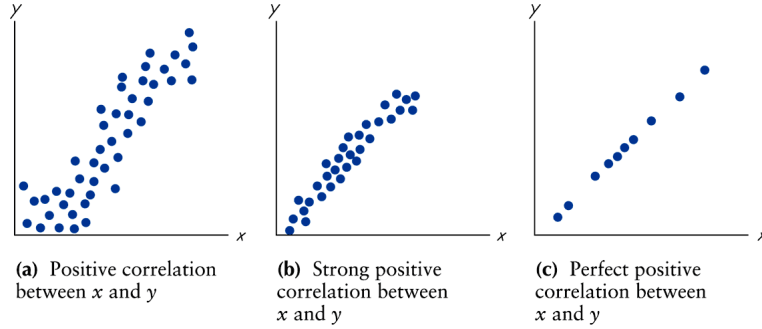
$$\text{Sample: } r = \frac{s_{xy}}{s_x s_y}$$

- Covariance:

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{N}}{N}$$

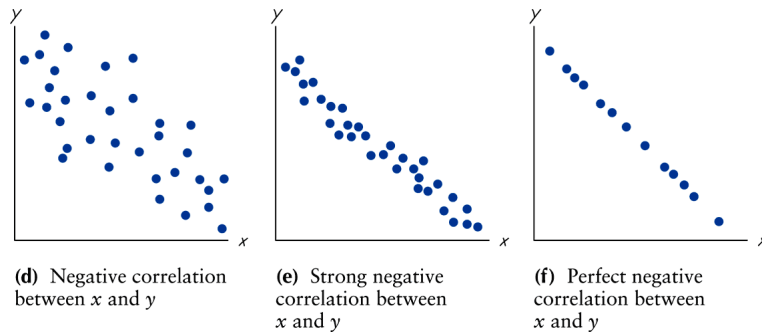
$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n - 1}$$

### Positive linear correlation



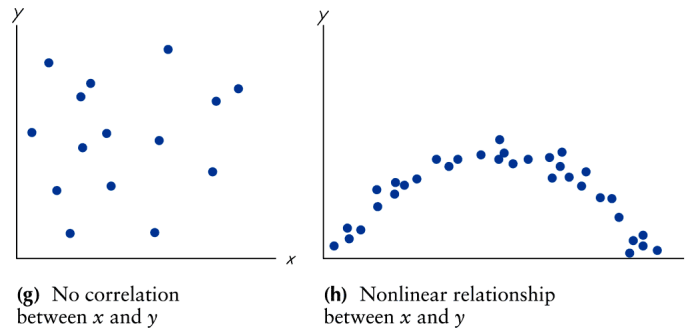
- Positive correlation: two variables move in the same direction.
- Stronger the correlation: closer the correlation coefficient is to 1.
- Perfect positive correlation:  $\rho = 1$

### Negative linear correlation



- Negative correlation: two variables move in opposite directions.
- Stronger the correlation: closer the correlation coefficient is to -1.
- Perfect negative correlation:  $\rho = -1$

### No linear correlation



- Panel (g): no relationship at all.
- Panel (h): strong relationship, but not a *linear* relationship.
  - Cannot use regular correlation to detect this.

## 2.2 Inferences about $\rho$

### Inferences about $\rho$

- What is the only thing you need to know to be able to do hypothesis testing and compute confidence intervals about correlations? [Standard deviation of the sampling distribution of  \$r\$](#)

$$\sigma_r = \sqrt{\frac{1 - \rho^2}{n - 2}}$$

- Hypothesis testing. Degrees of freedom =  $n - 2$ , test statistic:

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

- Confidence interval:

$$(r - E, r + E) \quad E = t_{\alpha/2, n-2} \sqrt{\frac{1 - r^2}{n - 2}}$$

## 2.3 Example

### Example

Suppose you are interested in determining whether lung capacity and smoking are correlated.

- You collect a sample of 109 people including smokers and non-smokers.
- For each individual you ask them how many cigarettes they have per day, then measure their lung capacity (in liters).

- The standard deviation of cigarettes per day is 7.1.
- The standard deviation of lung capacity is 0.27.
- Covariance is equal to -1.2.

Test the hypothesis that lung capacity and smoking are correlated.

## 3 Regression

### 3.1 Regression line

#### Regression

- Regression line: equation of the line that describes the linear relationship between variable  $x$  and variable  $y$ .
- Need to assume that one variable causes another.
  - $x$ : *independent* or *explanatory* variable.
  - $y$ : *dependent* variable.
  - Variable  $x$  can influence the value for variable  $y$ , but not vice versa.
- Example: Suppose one want to estimate how much smoking affects lung capacity.
  - $x_i$ : quantity of cigarettes smoked per day by individual  $i$  (independent).
  - $y_i$ : lung capacity of individual  $i$  (dependent).

#### Regression line

- Population regression line:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- The actual coefficients  $\beta_0$  and  $\beta_1$  describing the relationship between  $x$  and  $y$  are unknown.
- Use sample data to come up with an estimate of the regression line:

$$y_i = b_0 + b_1 x_i + e_i$$

- Since  $x$  and  $y$  are not perfectly correlated, still need to have an error term.

### Interpreting the slope

- Interpreting the slope,  $\beta_1$ : amount the  $y$  is predicted to increase when increasing  $x$  by one unit.
- When  $\beta_1 < 0$  there is a negative linear relationship. That is increasing  $x$  causes  $y$  to decrease.
- When  $\beta_1 > 0$  there is a positive linear relationship. That is increasing  $x$  causes  $y$  to increase.
- When  $\beta_1 = 0$  there is no linear relationship between  $x$  and  $y$ .

### Predicted values and residuals

- Given a value for  $x_i$ , can come up with a **predicted value** for  $y_i$ .

$$\hat{y}_i = b_0 + b_1 x_i$$

- This is not likely be the actual value for  $y_i$ .
- **Residual** is the difference between the actual value of  $y_i$  and the predicted value,  $\hat{y}$ .

$$e_i = y_i - \hat{y} = y_i - b_0 - b_1 x_i$$

## 3.2 Estimating regression line

### Least Squares Estimate

- How should we obtain the “best fitting line”.
- Ordinary least squares (OLS) method.
- Choose sample estimates for the regression coefficients that minimizes:

$$\sum_{i=0}^n (y_i - \hat{y}_i)^2$$

### Estimating the regression coefficients for a sample

- Formulas for the estimates of  $b_0$  and  $b_1$ .

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- To conduct hypothesis tests and confidence intervals, need... [Standard deviation of the sampling distribution of  \$b\_0\$  and  \$b\_1\$](#) .
- Unfortunately, formulas are too complicated for this class.
- Use a computer.
- The degrees of freedom for hypothesis testing =  $n - 2$ .