

# Correlation and Simple Linear Regression

A Evans

## Correlation: dual meaning.

“Are X and Y correlated?”

### Normal meaning:

*Is there a relationship between X and Y?*

### Statistical meaning:

*Is there a **straight-line** relationship between X and Y?*

## Correlation: assumptions

1. Symmetry between X and Y (neither dependent nor independent)
2. X and Y are normally distributed. (If not true, then use Spearman correlation, if few ties, or Kendall’s tau, if frequent ties.)
3. Each subject is independently selected (selection of one subject doesn’t influence the selection of another; eg, friends or family member).
4. Subjects are not from heterogeneous populations (since correlation gives the *average* relationship among the variables; averages *not* appropriate if heterogeneous subgroups).
5. Investigator doesn’t have control over the value of X.
6. Best description of relationship is a straight line.

## Formula:

$$r = \frac{\text{Sum of: } (x - \text{mean } x)(y - \text{mean } y)}{(N)(SD_x)(SD_y)}$$

$$r = (\text{slope of line})(SD_x/SD_y)$$

$$r = \frac{\Delta Y/SD_y}{\Delta X/SD_x}$$

## Interpretation of Correlation Coefficient (r):

**r** measures how closely the points cluster around a straight line.

**r** does **NOT** measure the slope of the line; however ...

**r** does measure the “standardized slope”: for every one SD change in X, there is an r SD change in Y.

$r^2$  = fraction of the variance that is shared by the two variables.

$r^2$  = the proportion of variance of Y explained by a straight-line relationship with X; and the proportion of variance of X explained by straight-line relationship with Y.

### **Outliers are important:**

r heavily influenced by outliers, just like the mean and SD.

One point can make a big difference, especially in small datasets.

Outliers can either increase or decrease the correlation coefficient (greatly).

Outliers might be due to measurement or recording or transcribing errors, but some outliers are legitimate.

## **Linear Regression**

*“What’s the formula of the straight-line relationship?”*

$$Y = a + bX$$

$$\text{Slope} = b = r (SD_y/SD_x)$$

(r = correlation coefficient)

### **Linear Regression Model:**

Computer fits the best straight line through the data, such that the distance between the points and the line are the smallest possible.

If the slope is zero,  $b=0$ , then there is no relationship between Y and X (the line is flat).

### **Assumptions of Linear Regression:**

1. **Linearity:** the best fitting line through the data is a straight line.
2. **Independence:** observations are independent (eg, only one observation per patient).
3. **Normality:** for each value of X, the values of Y are normally distributed.
4. **Equality of variance:** for each value of X, the variance of Y is the same.

### **Testing Assumptions: Eyeball Test**

*Look at a scatter plot to see if:*

1. Straight line is best, rather than curved line;
2. Spread of Ys are roughly normal at each level of X (no outliers);
3. Spread of Ys are roughly equal at each level of X.

**Note:** Just as the mean and SD can be greatly influenced by outliers, so too can the regression line.

**Look at a scatter plot of the residuals:** easier on the eyes than original data.

1. Plot the residuals on the vertical axis.
2. All the different types of residuals (Standardized, Studentized, Deleted, etc.) will give similar picture.
3. Use *Studentized deleted* for clearest picture (SPSS: *SDRESID*).
4. Plot the predicted value of Y on the horizontal axis (*ZPRED*).
5. **Look:** Are the points spread equally and symmetrically around a horizontal line?

#### **What To Do When Assumptions Violated:**

1. Check for (and fix) measurement error, especially for outliers. Then recheck assumptions.
2. Delete outliers **only if** appropriate. Recheck assumptions.
3. Transform X variable. Recheck assumptions. (*avoid transformations if possible*)
4. Use a nonlinear model.

#### **Linear Regression: Making Predictions**

***If assumptions satisfied, then:***

1. make predictions, and
2. estimate confidence in predictions:
  - confidence in predicting **average** response (Y) for a given X.
  - confidence in predicting **individual** response for a given X.
  - **SPSS: Graphs: Scatter: Simple:**  
enter Y and X: then click OK.  
double click on the graph: **Chart: Options: Fit line: Fit options:** Linear:  
Regression prediction lines: mean and individual; show  $R^2$ .