UCLA College | Physical Sciences Statistics

STATS 10: Introduction to Statistical Reasoning
Chapter 4

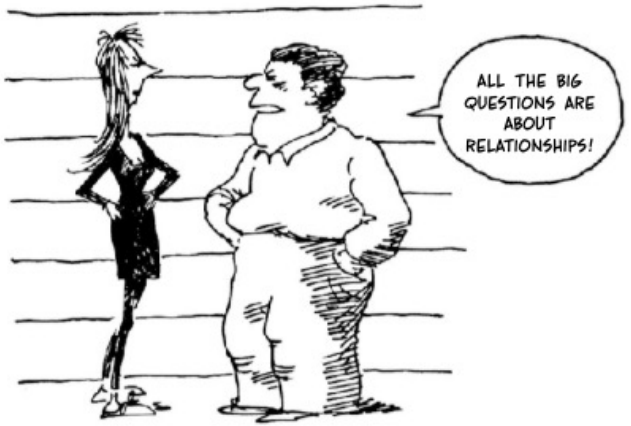
Bingling Wang
Department of Statistics

Exploring the Relationship between Two Variables

In most studies involving two variables, each of the variables has a role.

- The explanatory/independent variable
- The response/dependent variable

1. How is the treatment effect related to the dose?
2. How well can we predict students' freshman year GPA from their SAT score?
3. How does real estate website estimate the value of a home based on certain characteristic?
4. Can we predict life expectancy from blood pressure level?
5. How is the number of calories in a hot dog related to the type of hot dog (beef or poultry)?



Scatterplot

Two Numerical Variables Example

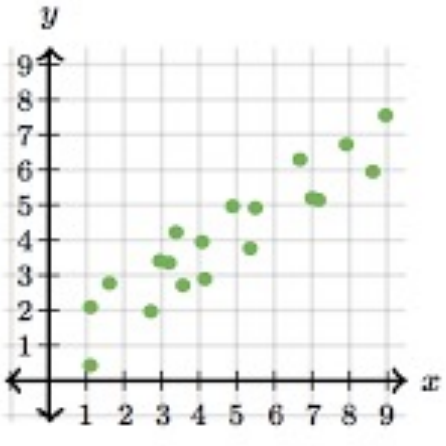
Highway Signs

Data on 30 drivers (of ages 18 to 82 years old) were collected to explore the relationship between a driver's age and the maximum distance at which signs were legible, and then use the study's findings to improve safety for older drivers.

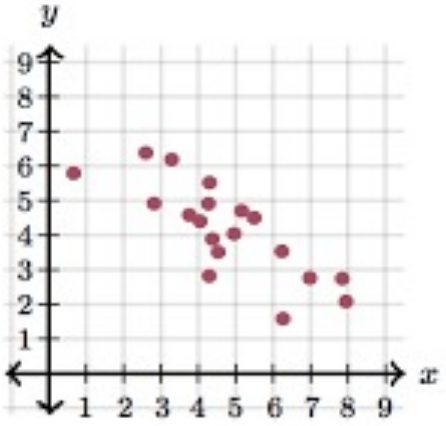
	Age	Distance
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.....		
Driver 30	82	360

Trend / Direction

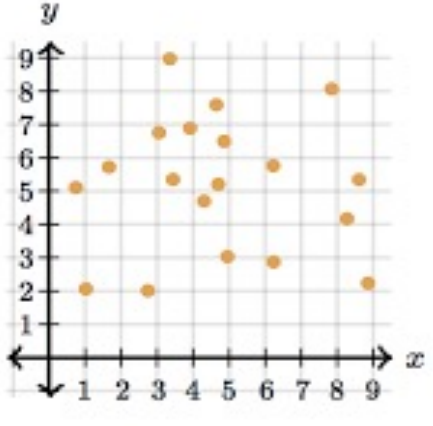
The **Trend** is the general tendency of the scatterplot as you scan from left to right.



Positive/increasing trend



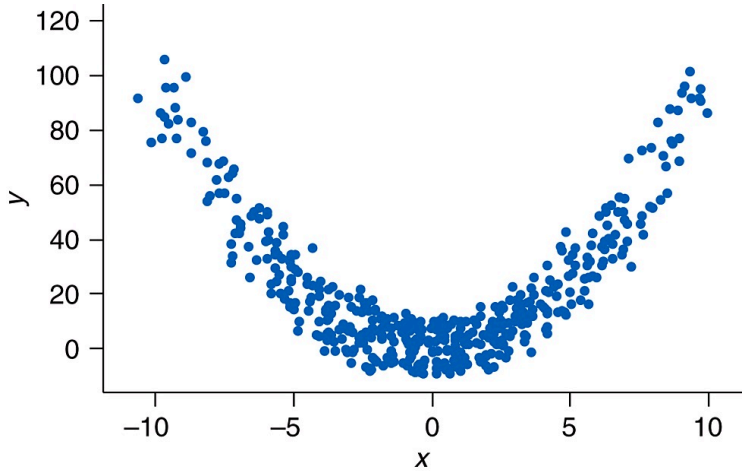
Negative/decreasing trend



Neither positive nor negative

Changing Trend

The trend is not always easily categorized as positive or negative.



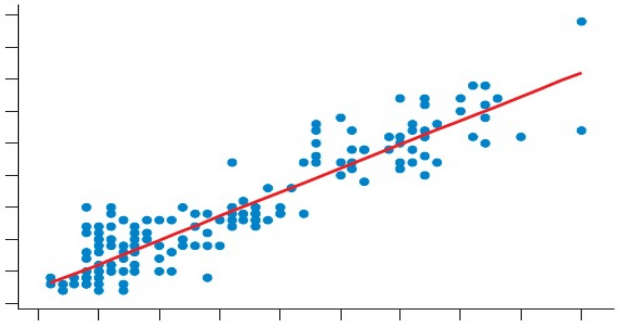
The trend changes depending on the x values.

Example: the trend is negative for lower values of x and positive for higher values of x.

Shape

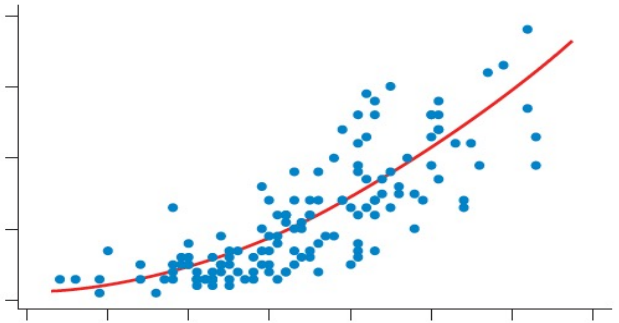
Linear:

- Data points appear as a cloud of points stretched out in a generally consistent, straight form.
- The scatterplot clusters around a line.
- Linear trends always increase/decrease at the same rate.



Non-linear:

- Patterns that are more complex and can't be modeled by a straight line.
- The rate of increase/decrease in the trend changes depending on the values of the variables.

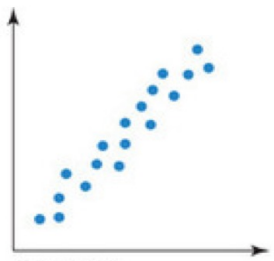


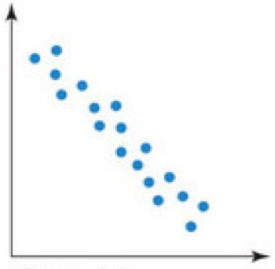
Strength

The **Strength** is described by the amount of scatter in the scatterplot.

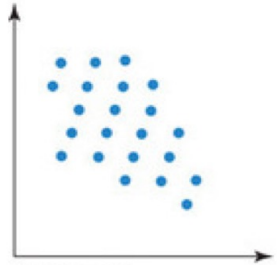
Scatter refers to the spread of the points in the vertical direction.

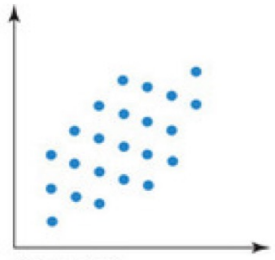
- Small amount of scatter, or little vertical variation indicates a ***strong*** association





- Larger amount of scatter, or vertical variation indicates a ***weak*** association





Description of Relationships

The description should include:

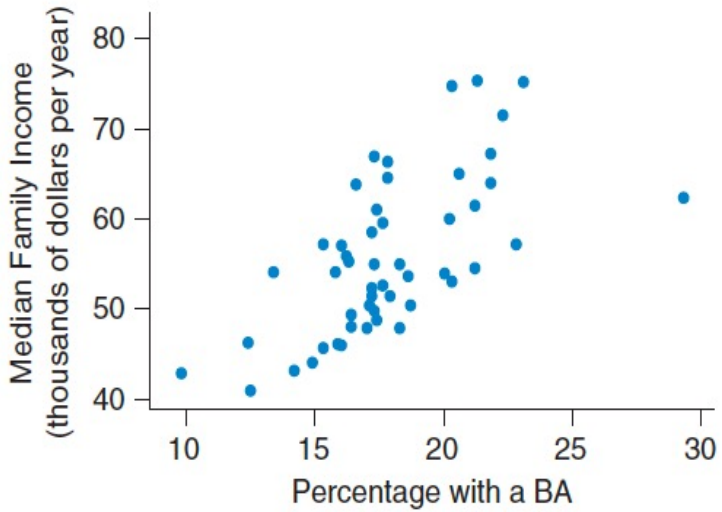
- **Trend:** Is there an association?
- **Shape:** Is the trend linear or nonlinear?
- **Strength:** Is the trend strong, weak or moderate?
- **Unusual points:** Are there outliers, cluster of points, or anything that does not fit the general pattern?
- **Context of the data:** Explain what these mean for the variables.

Use phrase like “tends to” when describing an association

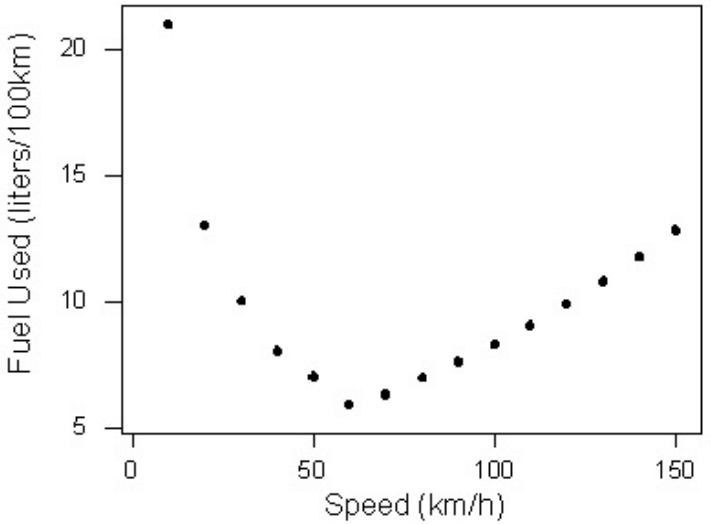
Avoid using absolute terms when interpreting trends.

Examples

1. Data on 50 states taken from the U.S. Census shows how the median family income is related to the population (25 years or older) with a college degree or higher.

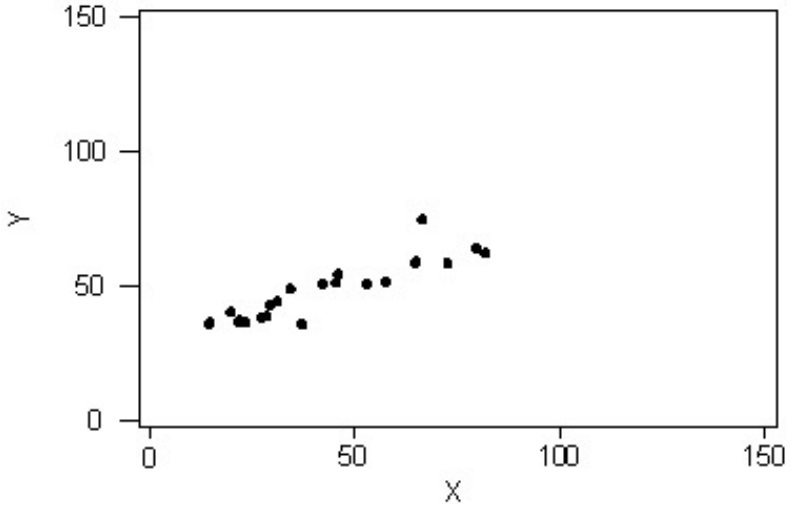


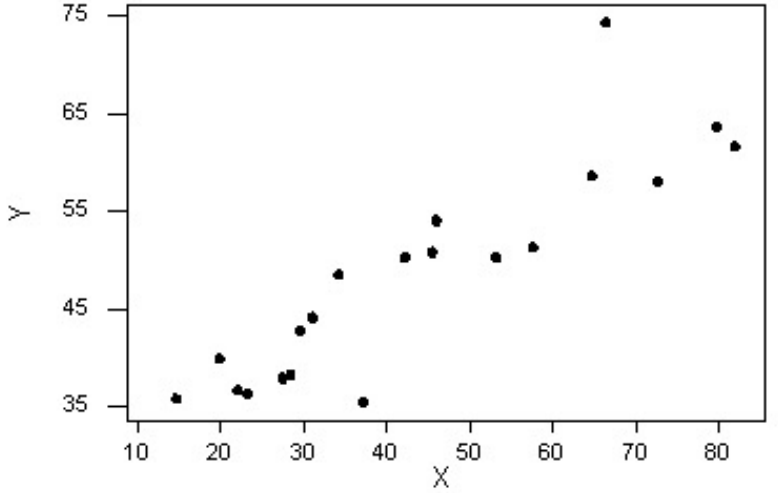
2. Consider the relationship between the average amount of fuel used (in liters) to drive a fixed distance in a car (100 kilometers), and the speed at which the car is driven (in kilometers per hour).



Measuring Strength of Association

Which scatterplot shows a stronger association?





Correlation Coefficient

Pearson's Correlation Coefficient (r)

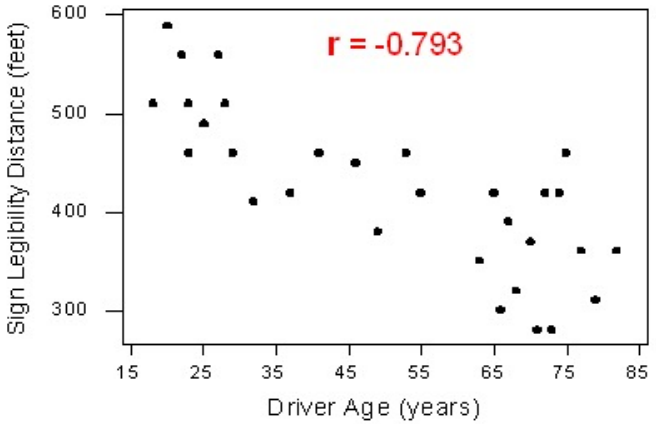
-- A numerical measurement of the strength of a linear relationship

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \, (y_i - \bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y} = \frac{1}{n-1}\sum z_x \, z_y$$

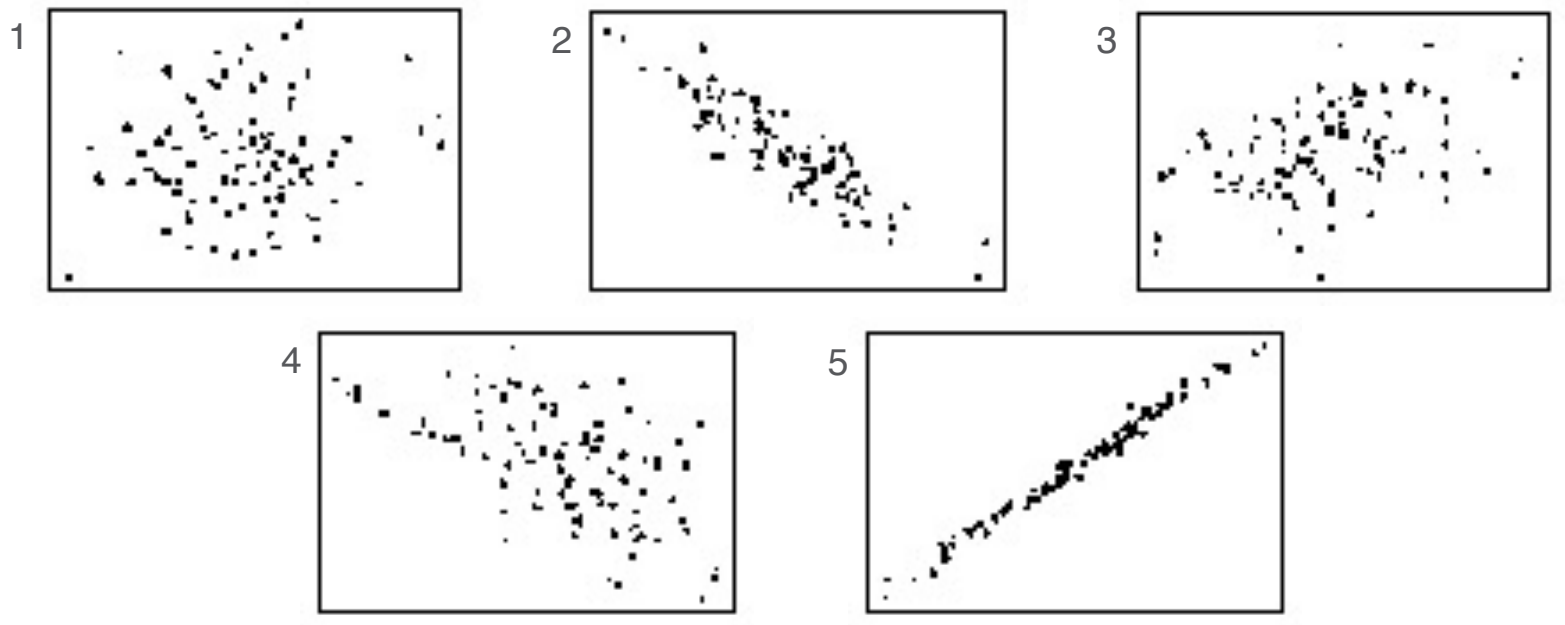
- Symbol: r
- Range: [-1, 1]
- The sign gives the trend of the relationship.

The correlation coefficient makes sense only if the trend is **linear** and both variables are **numerical**

Highway Sign Visibility

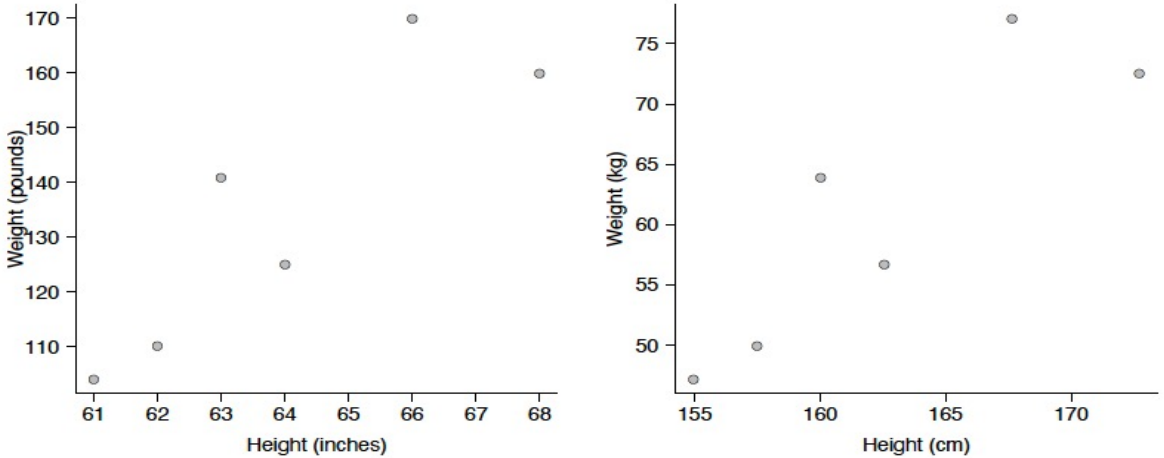


Visualizing Correlation Coefficient

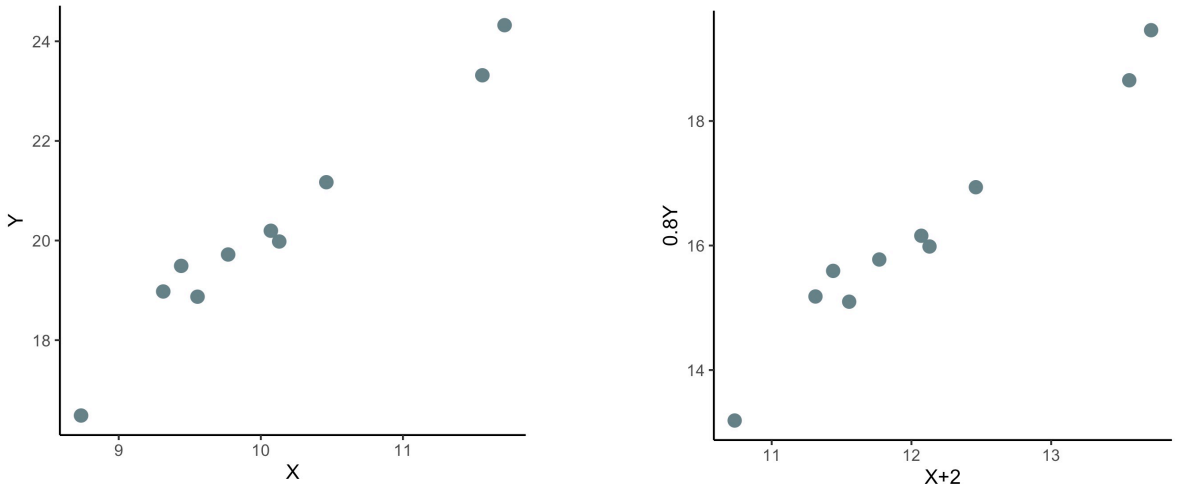


A. 0.436 B. 0.100 C. -0.897 D. 0.995 E. -0.575

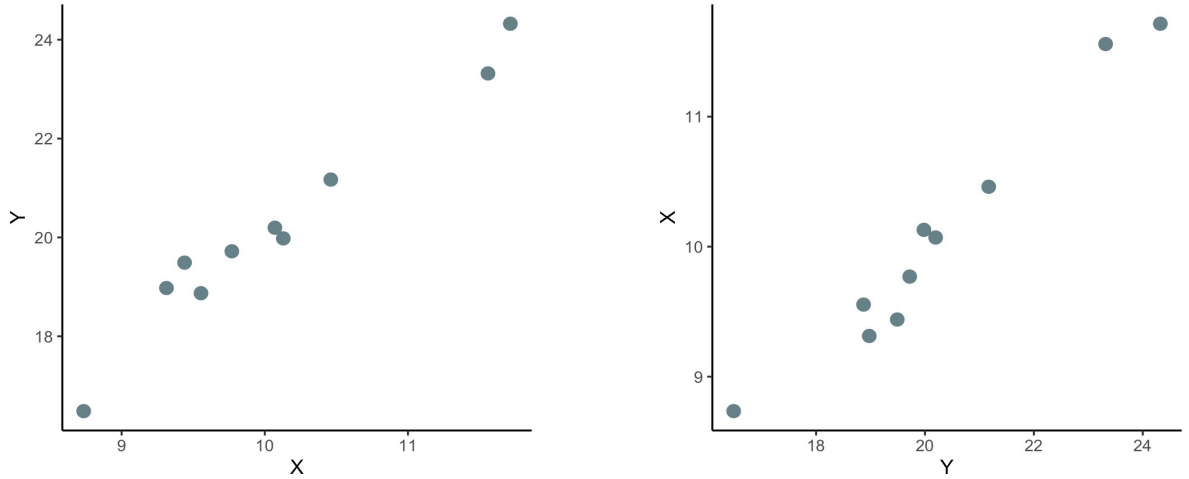
Compare the Correlation Coefficient



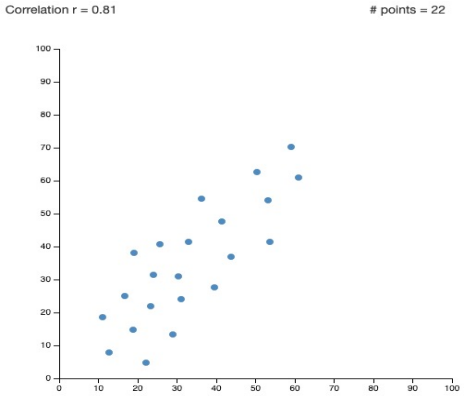
Compare the Correlation Coefficient

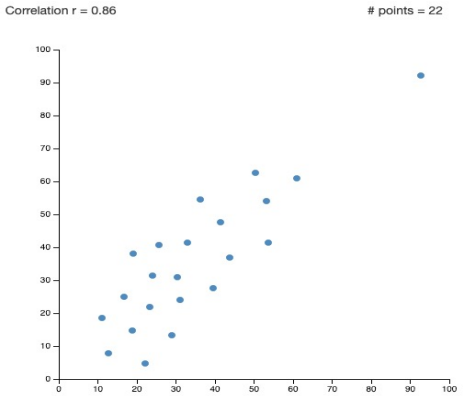


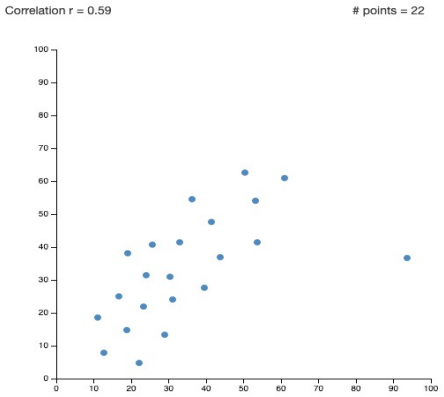
Compare the Correlation Coefficient

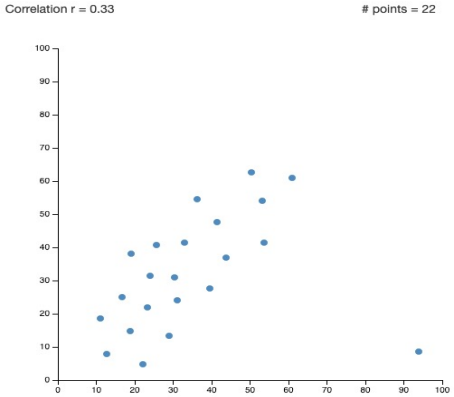


Compare the Correlation Coefficient









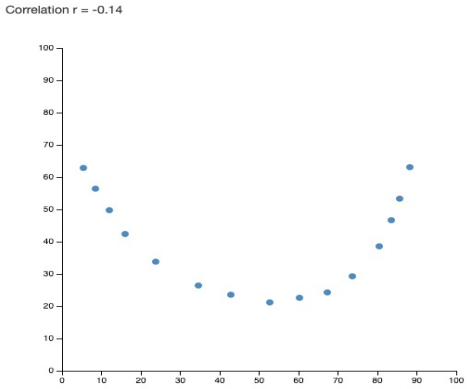
Properties of the Correlation Coefficient

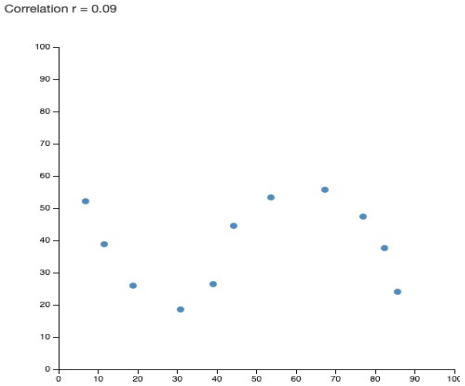
1. Correlation coefficient is unitless.
2. Adding a constant or multiplying by a positive constant does NOT affect the value of the correlation coefficient.
3. The order of the variables does not matter.
4. Correlation coefficient is sensitive to outliers.

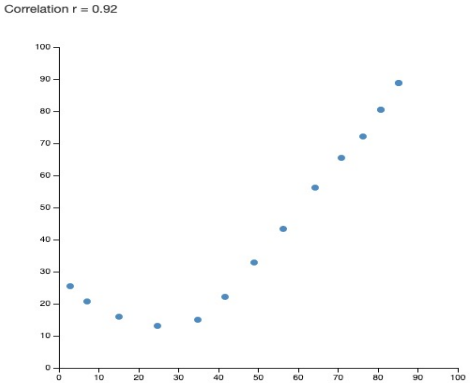
Caution!

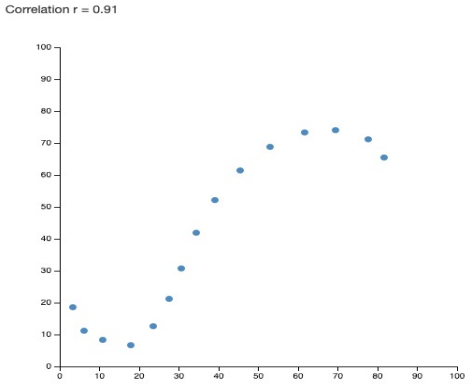
Linear! Correlation coefficient measures the strength of linear association.

- The correlation does not tell you the shape of a trend
- If you know that the association is linear, then the correlation coefficient is a measure of its strength.









Always check the linearity of the data before interpreting the correlation coefficient!

Modeling Linear Trends

Modeling Linear Trends

How much more do people tend to weigh for each additional inch in height?

How much value do cars lose each year as they age?

In order to use relationships/trends between variables to make predictions, we need a way to summarize the linear relationship.

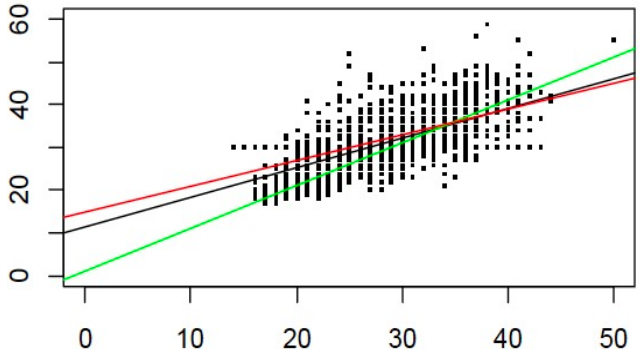
Statistical Model -- consists of an equation and a set of conditions that describe when the model will be appropriate.

- We assume that the trend can be summarized by a mathematical equation
- We use observed data to estimate the mathematical equation

Modeling linear trends with an equation

For linear trends, we assume that the trend can be summarized by the equation of a line.

The **regression line** is a statistical model that summarizes the linear trend of the observations. It also represents our prediction for any new or future observations.

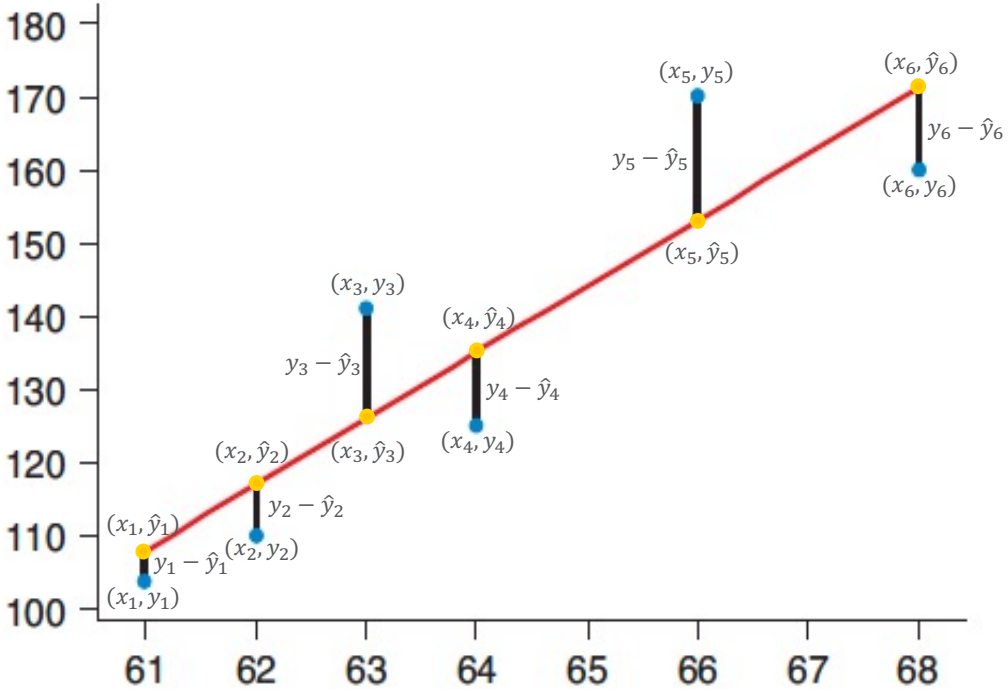


The regression line is given by an equation for a straight line with the form:

$$y = a + bx$$

y – the response variable; x – the explanatory variable; a – the intercept; b – the slope

Finding the Regression Line of "Best" fit



Blue dots: observed (x, y) values

Red line: fitted regression line

Orange dots: predicted values from the model, denoted by $(x, \hat{y})$, $\hat{y}$ -- predicted value of y.

Residual: the vertical distance between each observation and the line, $y_i - \hat{y}_i$.

The regression line of best fit is the line for which the sum of squared residuals is the smallest.

-- least squares line

Interpreting the Regression Line

The mathematical expression of the regression line:

$$\hat{y} = a + bx$$

1. The slope: $b = r\frac{s_y}{s_x}$

- **Calculation**: using the correlation coefficient, r, and the standard deviations of the explanatory variable (s_x) and the response variable (s_y)
- **Interpretation**: how much do we expect y to change by, on average, when x is increased by one additional unit
  - When the slope b is positive, y is expected to increase as x increases
  - When the slope b is negative, y is expected to decrease as x increases

2. The intercept: $a = \bar{y} - b\bar{x}$

- **Calculation**: using the mean of y ($\bar{y}$), and the the mean of x ($\bar{x}$), and the slope.
- **Interpretation**: the predicted value of y when x is 0.
  - The y-intercept is meaningful only if it makes sense for x to equal 0.

Regression Line Example

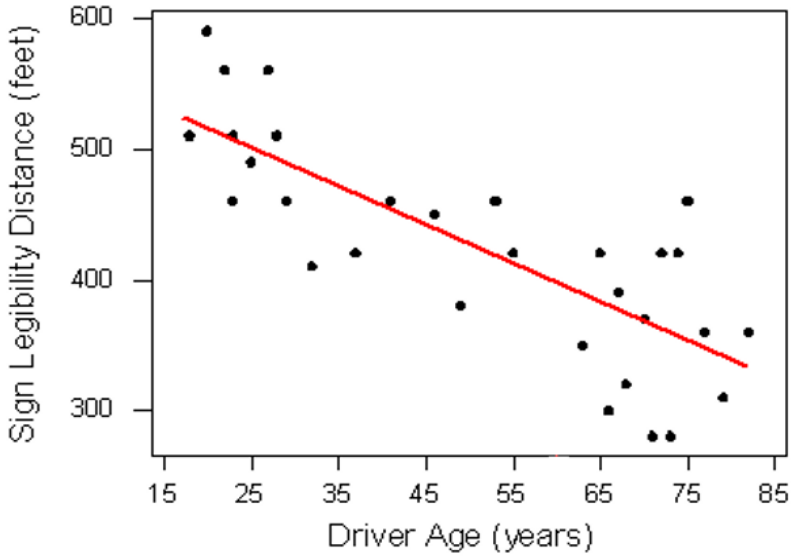
Suppose the fitted least-squares regression equation for relationship between height and weight is:

$$\widehat{\text{weight}} = -442.882 + 9.029 \times \text{height}$$

- The slope is 9.029.
  The following statements are equivalent:
  - For one inch increase in height, we would **expect** the weight to increase on average by 9.029 pounds.
  - For every additional inch in height, weight **tends** to increase by 9.029 pounds.
  - Every increase of 1 inch in height is **associated** with an increase in weight of 9.029 pounds.

- The intercept is -442.882.
  - It is the predicted value for weight if height is 0. This is obviously impossible. Therefore, it doesn't make sense to interpret the intercept in this case.

Regression Line Example



	Age (X)	Distance (Y)
Mean	51	423
SD	21.78	82.8
Correlation	-0.793	

Evaluating the Linear Model

Measure Goodness of Fit

The Coefficient of Determination

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

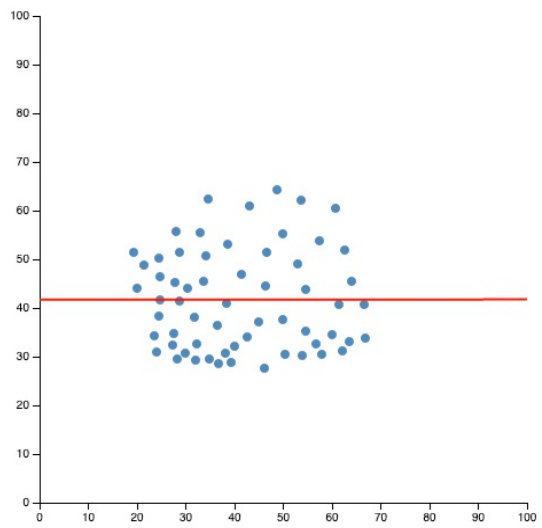
- The correlation coefficient squared (R^2, r^2, r-squared).
- Range: $0 \leq r^2 \leq 1$
- Measures how much the variation in response variable y is explained by the predictor x.
- Often converted to a percentage ($0\% - 100\%$).

Example: the correlation between X and Y is $r = -0.778$

$$r^2 = (-0.778)^2 = 0.605 = 60.5\%$$

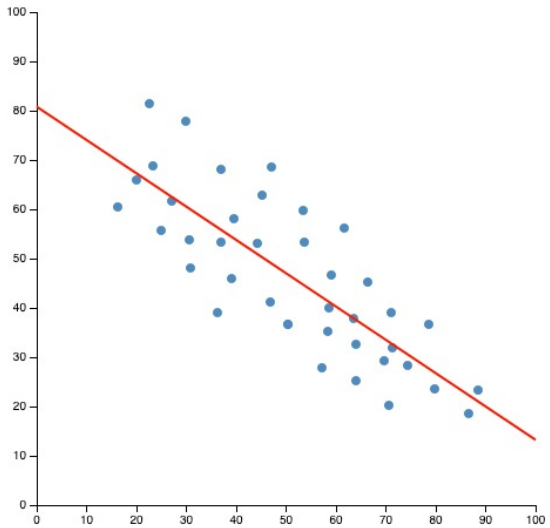
60.5% of the variation in Y is explained by X.

Coefficient of Determination



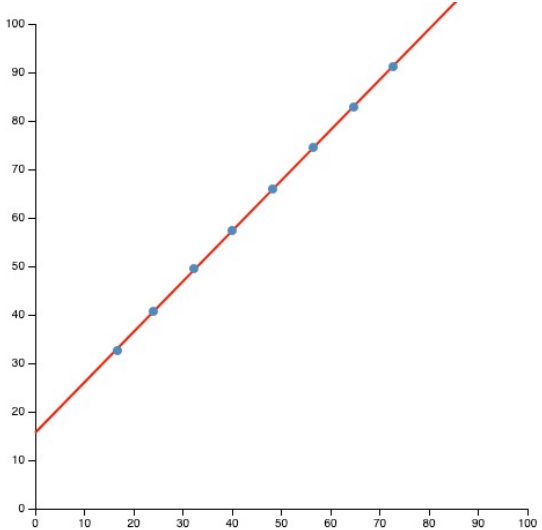
$r^2 = 0\%$:

None of the variation in y is explained by x.



$r^2 = 64\%$

Some portion (64%) of the variation in y is explained by x.



$r^2 = 100\%$

The variation in y is perfectly explained by x.

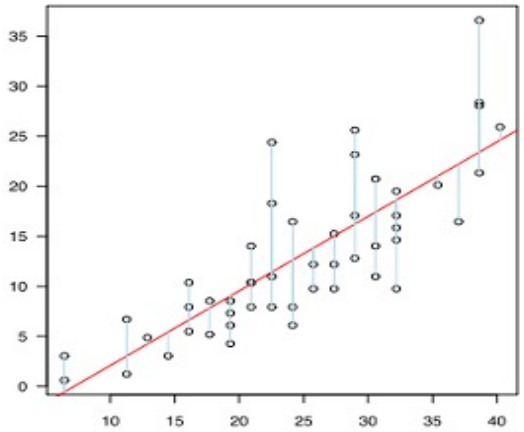
Residual Plot

Residual = Observed value – Predicted value

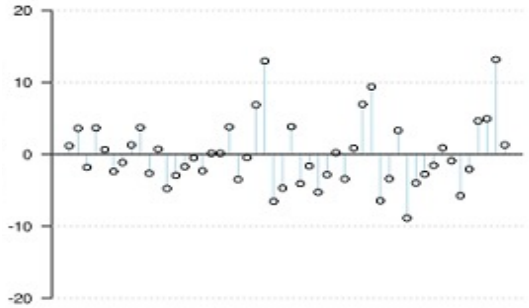
A residual plot shows how close each data point is vertically from the regression line.

- The horizontal axis -- the explanatory variable.
- The vertical axis -- the residuals.

Scatterplot with fitted regression line



Residual plot



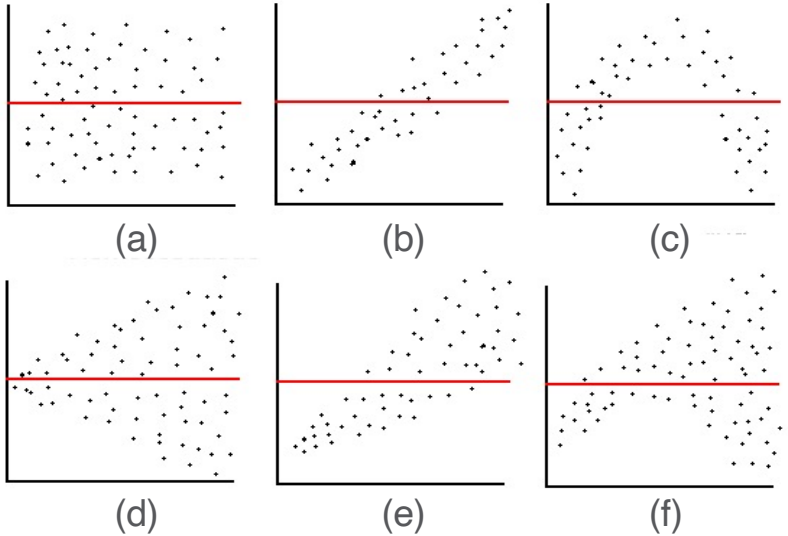
Goodness of fit -- Residual Plot

Good fit:

- The points are scattered randomly around 0.
- There is no apparent pattern in the plot

Bad fit:

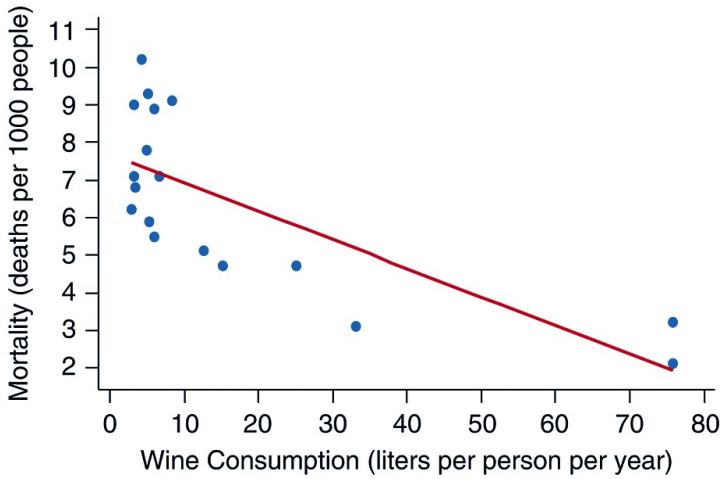
- The points are NOT randomly scattered around 0.
- There are apparent patterns in the plot



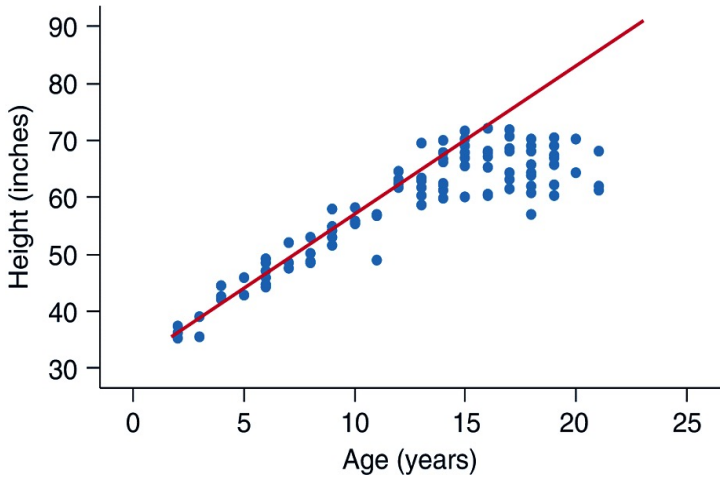
Cautionary Notes

- Do not fit linear models to nonlinear relationships.
  - Linear regression models are useful only for linear associations
- Correlation is not causation!
  - A strong correlation or a good-fitting regression line is not sufficient evidence of a cause-and-effect relationship.
- Beware of outliers!
- Do not extrapolate!
  - The linear trend may not continue to hold beyond the range of the data.

Examples



Fitting linear model to nonlinear relationship



Do not extrapolate