UCLA Statistics

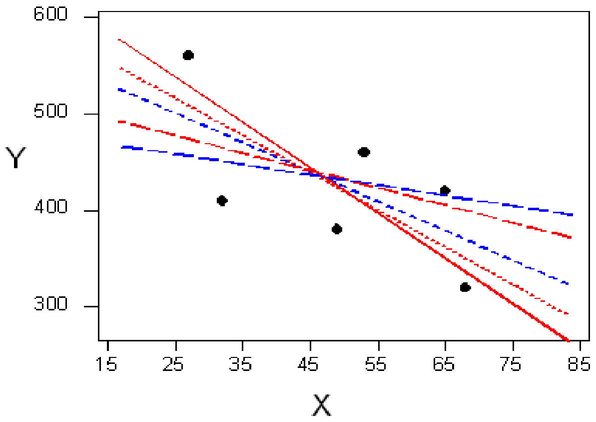
STATISTICS 10
Introduction to Statistical Reasoning

Bingling Wang
Ph.D., Lecturer, Department of Statistics

LINEAR REGRESSION

Modeling Linear Relationship with 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.



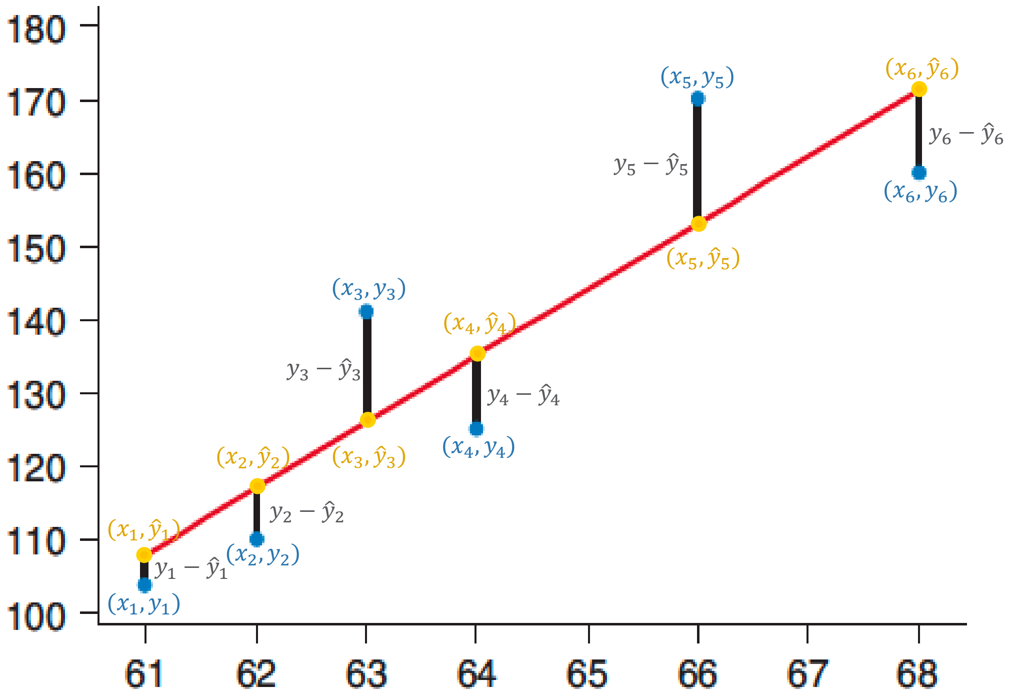
Which line best fits the data?

The equation for a straight line has 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 $(x, \hat{y})$

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

Criterion

The line with the smallest sum of squared residuals minimize $\sum(y_i - \hat{y}_i)^2$.

-- **least squares regression line**

Interpreting the Regression Line

The mathematical expression of the regression line:

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

Statistics needed for calculating a and b:

- $\bar{x}$, s_x -- the mean and the standard deviation of the explanatory variable
- $\bar{y}$, s_y -- the mean and the standard deviation of the response variable
- r -- the correlation coefficient

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

Interpretation: the average change in y when x increases by 1 unit

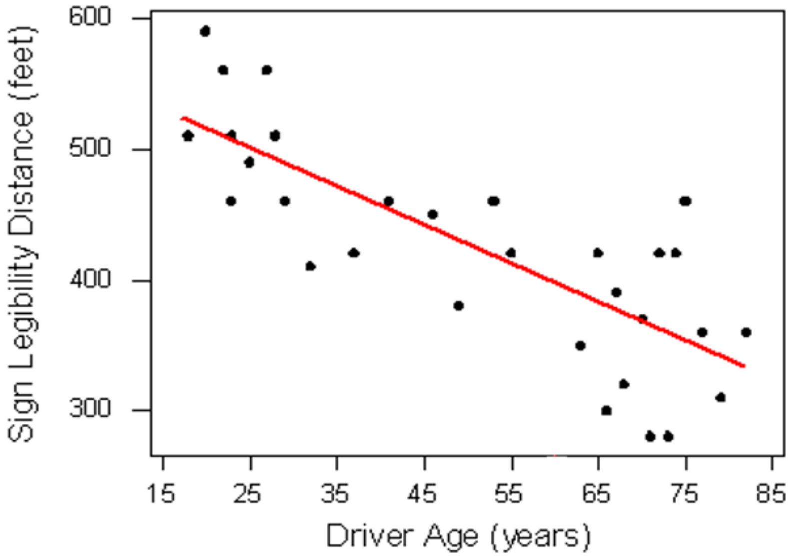
- When b is positive, y is expected to increase as x increases
- When b is negative, y is expected to decrease as x increases

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

Interpretation: the predicted value of y when x is 0

- The y-intercept is meaningful only if it makes sense for x to be 0.

Regression Line Example



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

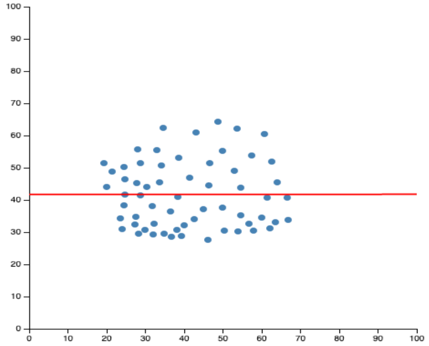
1. Find the least squares regression line.
2. Predict the maximum distance at which a sign is legible for a 60-year-old.

MODEL EVALUATION

Measure the Goodness of Fit -- R^2

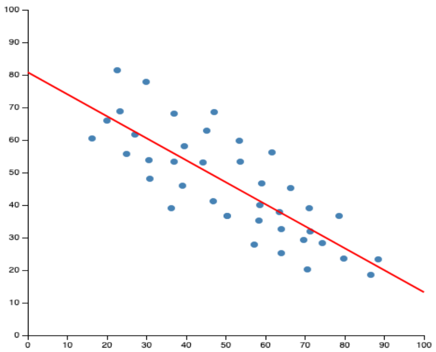
R^2 -- The Coefficient of Determination

- Range: $0 \leq r^2 \leq 1$, Often converted to a percentage $(0\% \ - 100\%)$.
- Measures how much the variation in response variable y is explained by the predictor x.
- The larger r^2, the smaller the amount of variation about the regression line.



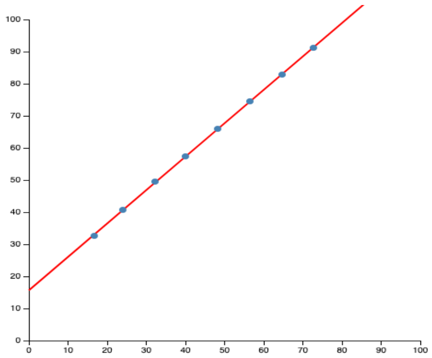
$r^2 = 0$:

None of the variation in y is explained by x.



$r^2 = 0.64$

64% of the variation in y is explained by x.



$r^2 = 1$

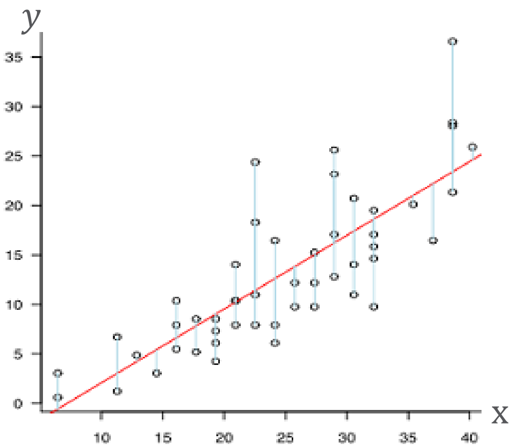
The variation in y is perfectly explained by x.

Measure the Goodness of Fit -- Residual Plot

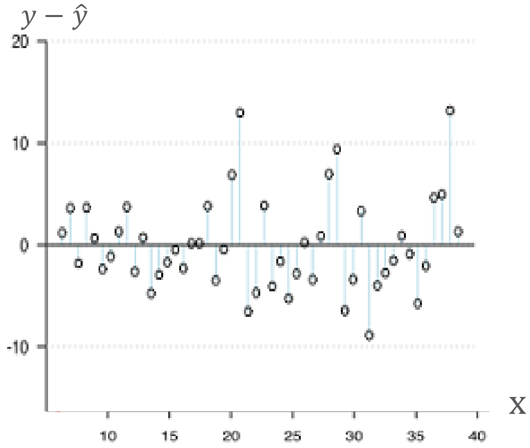
A residual plot shows how close each data point is vertically from the regression line.

- The horizontal axis -- the explanatory variable.
- The vertical axis -- the residuals [observed value y – predicted value $\hat{y}$]

Scatterplot of data with fitted regression line



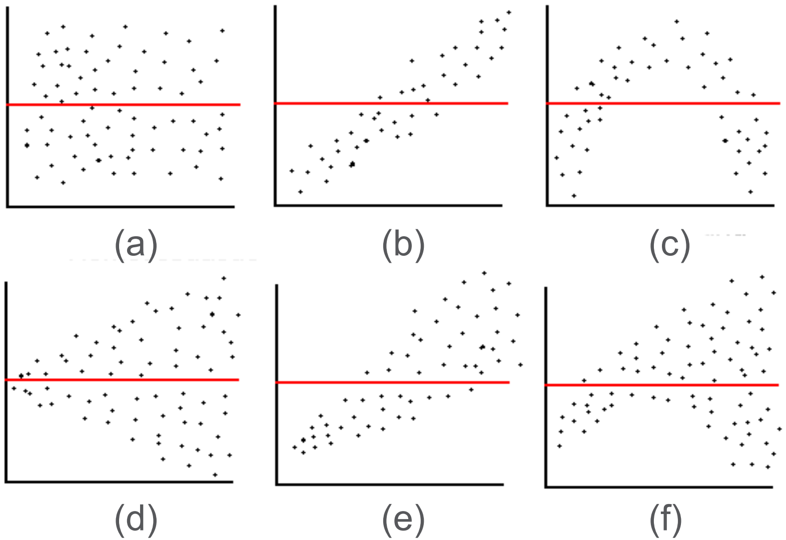
Residual plot



Good fit:

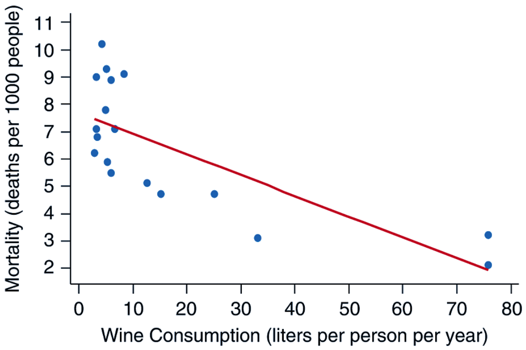
- The points are randomly scattered around 0.
- There is no apparent pattern in the plot.

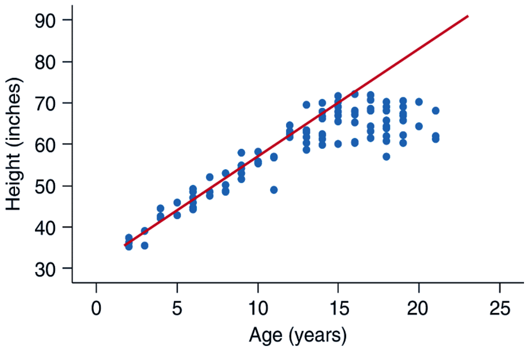
Goodness of fit -- Residual Plot

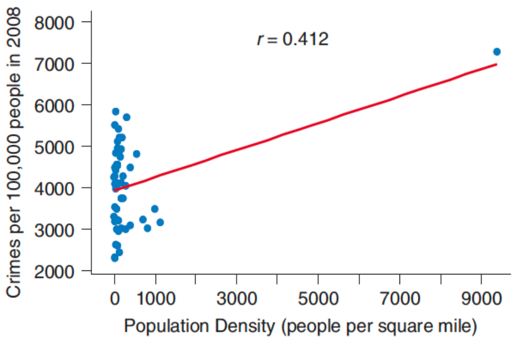


Cautionary Notes

- Do not fit linear models to nonlinear relationships.
- Do not extrapolate! -- The linear trend may not continue to hold beyond the range of the data.
- Beware of outliers!
- Correlation is not causation!







ASSOCIATION VS. CAUSATION

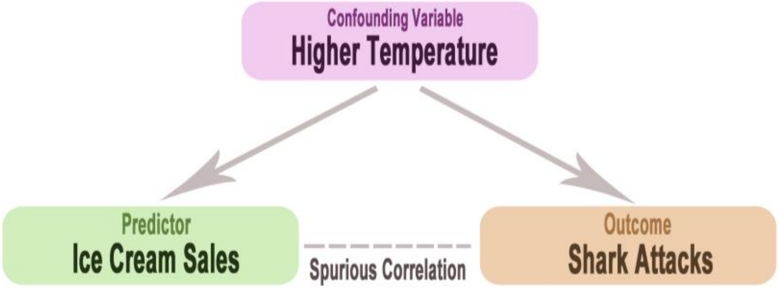
Association and Causation

Association: one variable provides information about another.

Two variables are associated if there is a relationship between them.

Caution! Association does NOT mean Causation!

Confounding Variable

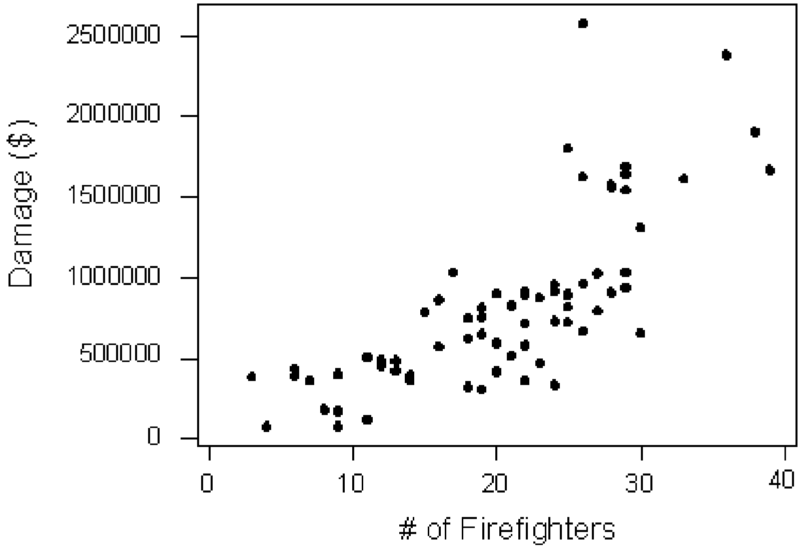


A third variable that influences the variables of interest.

- Causes a difference between the two groups
- Causes the two variables of interest to falsely appear to be causal related

Example -- Fire Damage

The scatterplot illustrates how the number of firefighters sent to fires (X) is related to the amount of damage caused by fires (Y) in a certain city.



Example - Hospital Death Rates

The following two-way table summarizes the data about the status of patients who were admitted to two hospitals in a certain city (Hospital A and Hospital B).

The purpose of the study is to examine whether there is a hospital effect on patients' status.

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

Example - Hospital Death Rates

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

Patients severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	57	1443	1500
	Hospital B	8	192	200
	Total	65	1635	1700

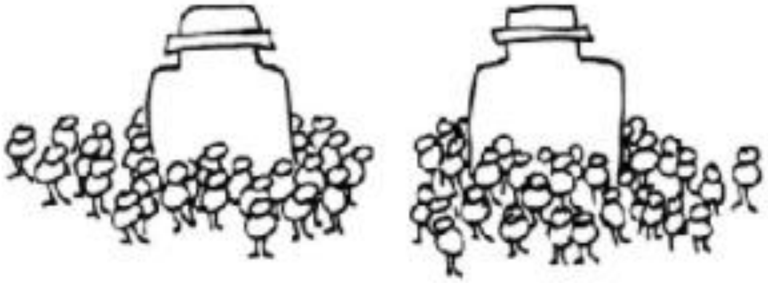
Patients *not* severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	6	594	600
	Hospital B	8	592	600
	Total	14	1186	1200

Establishing Causality

We want to answer whether the treatment variable **causes** the changes in the outcome variable

- **Treatment group**: subjects who receive the treatment of interest
- **Control group**: subjects who do not receive the treatment



In order to conclude causality from a study, it is important to have both treatment and control groups, and for subjects in both groups to be identical in every way except for the treatment.