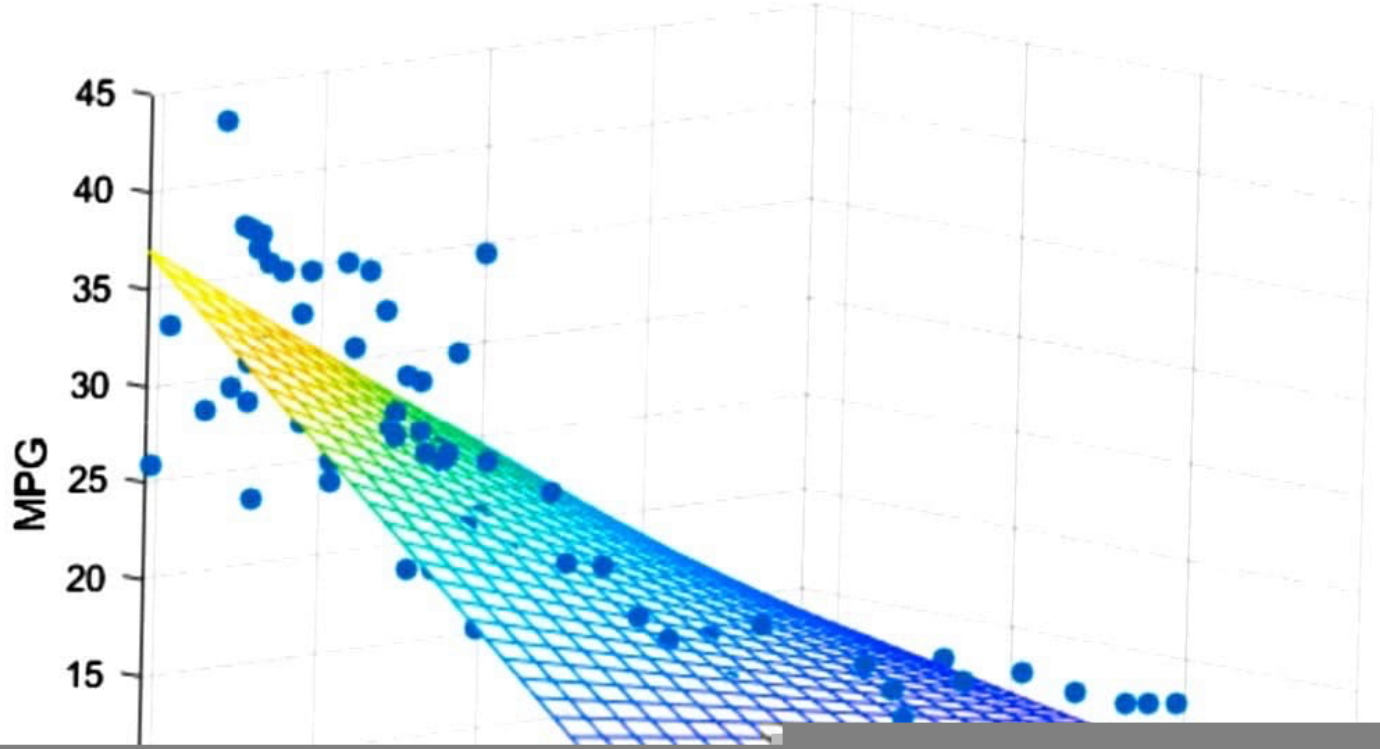
Linear Regression



Learning Goals

- What is regression?
- Why regression?
- Scatter plot
- Measures of association
  - Correlation coefficient
- Simple linear regression
  - Fitting a regression line

Regression Analysis

- ✓ Regression Analysis + Correlation = Predict future performance using past results

- ✓ While Correlation explains the degree of linear relationship that exists between two variables, Regression defines the relationship more precisely
- ✓ Regression analysis is a tool that uses data on relevant variables to develop a prediction equation, or model
- ✓ It generates an equation to describe the statistical relationship between one or more predictors and the response variable and to predict new observations

Simple Linear Regression

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable.

It looks for statistical relationship but not deterministic relationship.

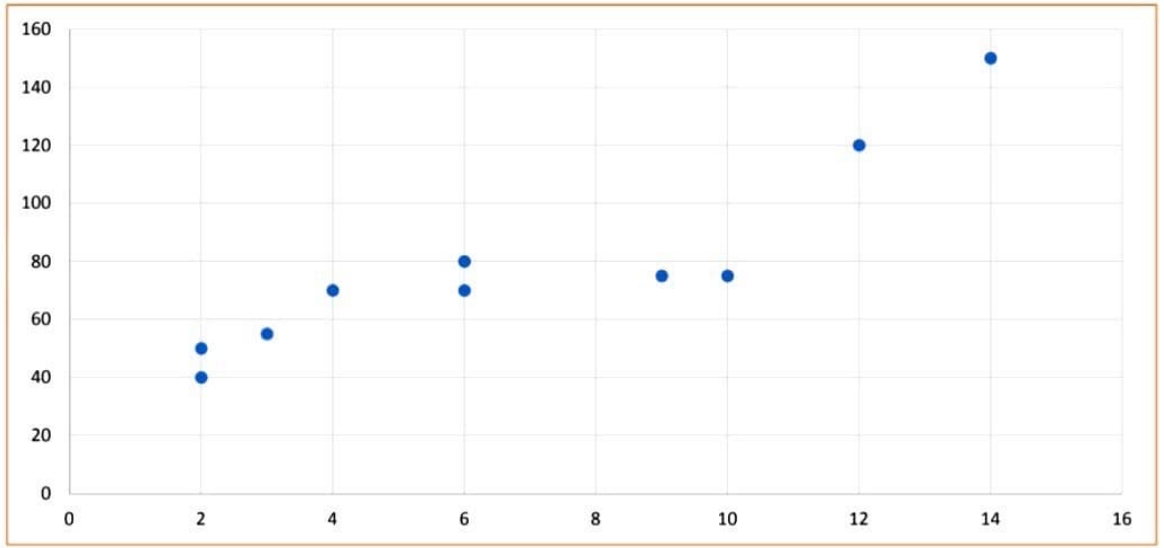
For example, relationship between height and weight.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

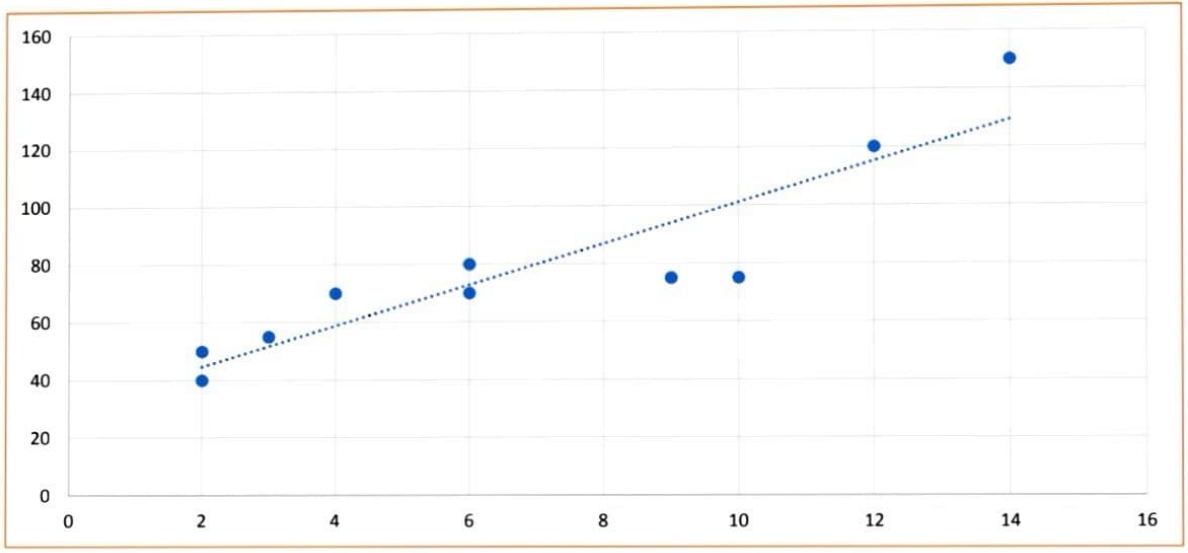
- ✓In Simple Linear Regression, a single variable "X" is used to define/predict Y
  - ✓E.g. Used car cost = B1 + (B2) x (Miles driven) + E (error)
  - ✓Simple Regression Equation: Y = B1 + (B2) *(X)+ E (error)

Regression

Exp	Salary
2	50
4	70
3	55
9	75
12	120
14	150
10	75
2	40
6	80
6	70



Exp(Yrs)	Salary(KUSD)
2	50
4	70
3	55
9	75
12	120
14	150
10	75
2	40
6	80
6	70



Business Case :The Newspaper Data

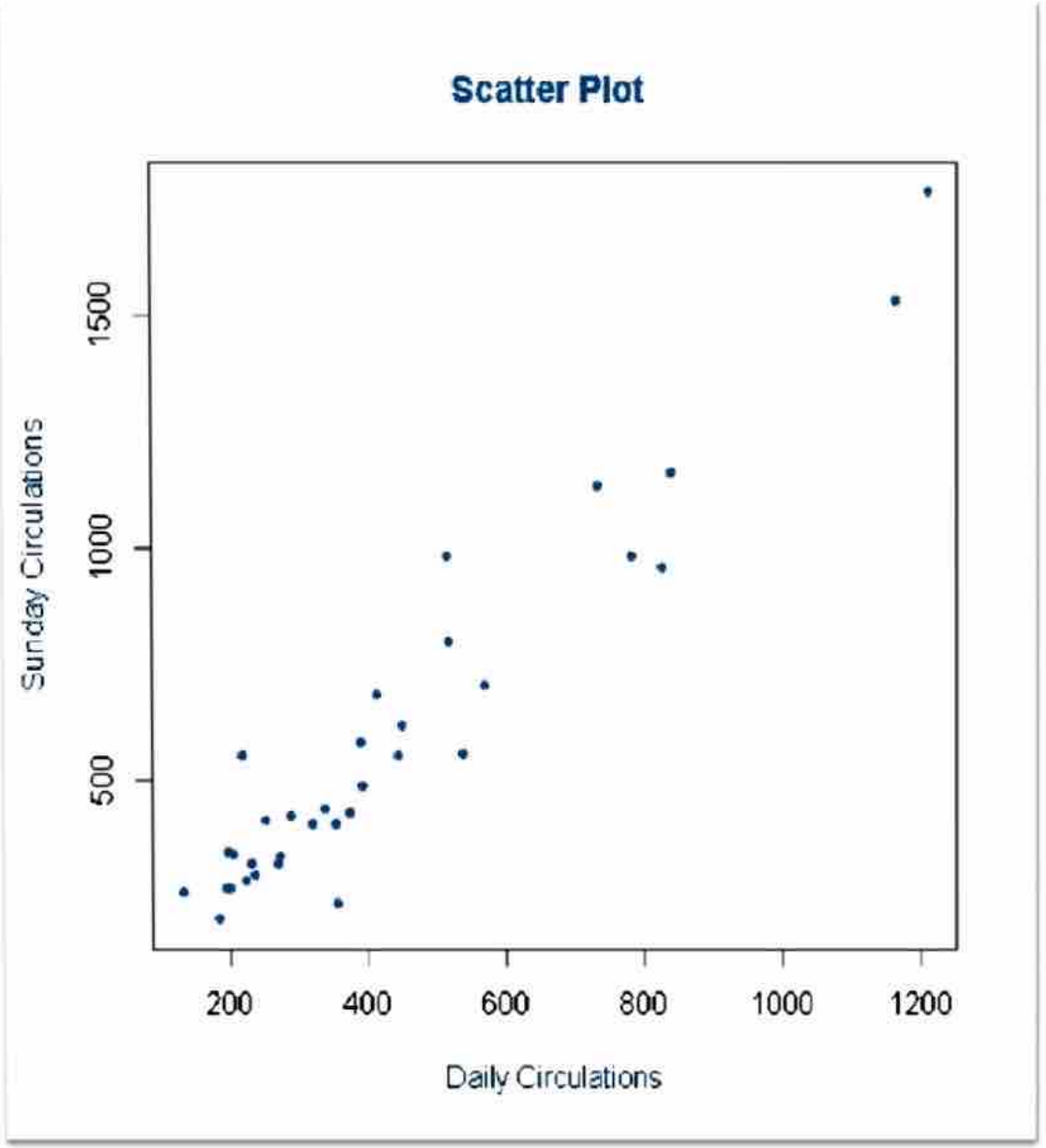
- In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands)

Newspaper	daily	sunday	Newspaper	daily	sunday
Baltimore Sun	391.952	488.506	New York Daily News	781.796	983.24
Boston Globe	516.981	798.298	New York Times	1209.225	1762.015
Boston Herald	355.628	235.084	Newsday	825.512	960.308
Charlotte Observer	238.555	299.451	Omaha World Herald	223.748	284.611
Chicago Sun Times	537.78	559.093	Orange County Register	354.843	407.76
Chicago Tribune	733.775	1133.249	Philadelphia Inquirer	515.523	982.663
Cincinnati Enquirer	198.832	348.744	Pittsburgh Press	220.465	557
Denver Post	252.624	417.779	Portland Oregonian	337.672	440.923
Des Moines Register	206.204	344.522	Providence Journal-Bulletin	197.12	268.06
Hartford Courant	231.177	323.084	Rochester Democrat & Chronicle	133.239	262.048
Houston Chronicle	449.755	620.752	Rocky Mountain News	374.009	432.502
Kansas City Star	288.571	423.305	Sacramento Bee	273.844	338.355
Los Angeles Daily News	185.736	202.614	San Francisco Chronicle	570.364	704.322
Los Angeles Times	1164.388	1531.527	St. Louis Post-Dispatch	391.286	585.681
Miami Herald	444.581	553.479	St. Paul Pioneer Press	201.86	267.781
Minneapolis Star Tribune	412.871	685.975	Tampa Tribune	321.626	408.343
New Orleans Times-Picayune	272.28	324.241	Washington Post	838.902	1165.567

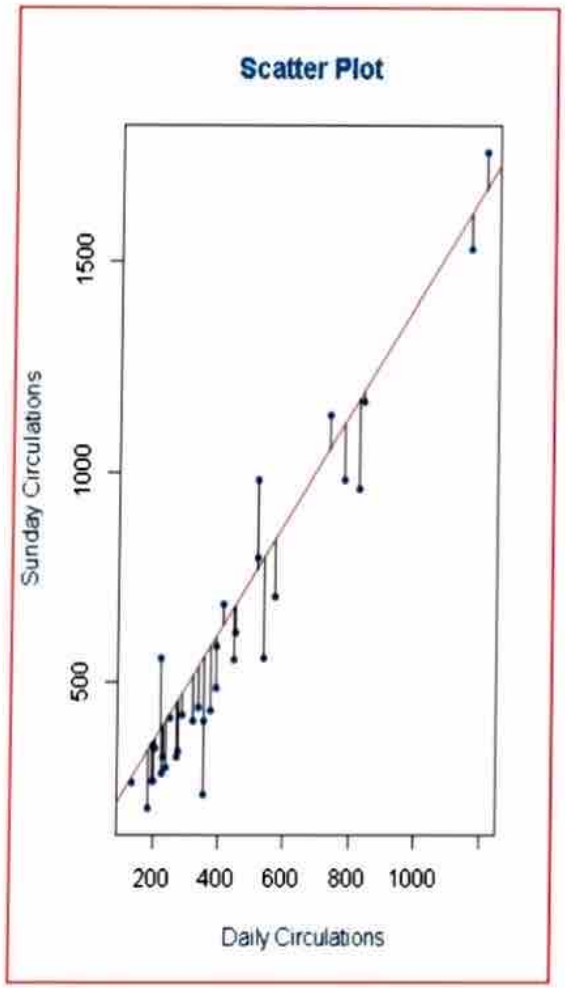
Data Set

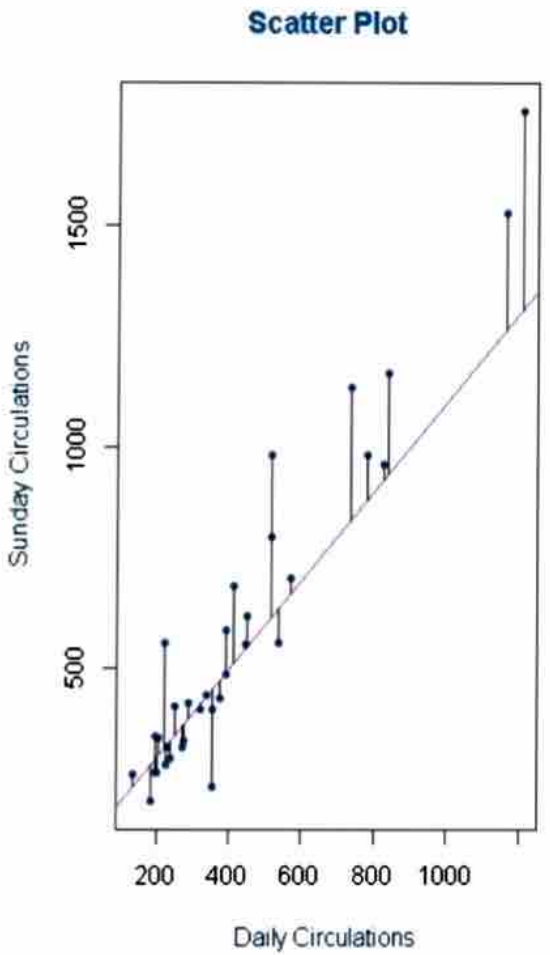
Source: Gale Directory of Publications, 1994

Scatter Plot: The Newspaper data



Which Straight Line? … The Newspaper data





The Best Line: Least Squares Method

- **The line of our interest is:**

$$Sunday = \beta_0 + \beta_1 Daily + \varepsilon$$

or

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Let us do the Regression in Python