Simple Linear Regression
Part 1: Introduction

STAT 705: Regression and Analysis of Variance

Example

In order to assess the impact of vehicle emissions on the environment, a researcher selected several sites along a freeway. At each site, he counted the number of vehicles that passed the site in a 24 hour period. He also measured the concentration of lead in the bark of trees near the site

Do you expect to have a relationship between these two variables?

Example, continued

- Data collected by the researcher are shown in the table
- Columns are variables
  - Traffic is measured in thousands of vehicles in a 24-hour period
  - Concentration of lead is measured in micrograms of lead per gram of tree bark
- Rows are the sites (locations) along the highway

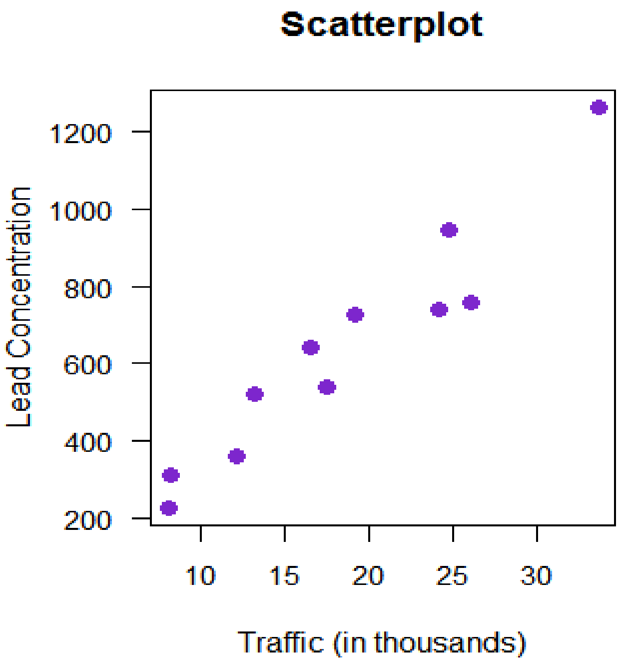
Traffic	Lead
8.1	227
8.3	312
12.1	362
13.2	521
16.5	640
17.5	539
19.2	728
24.8	945
24.1	738
26.1	759
33.6	1263

Bivariate Data

- The data in the table is an example of bivariate data
- Both Traffic and Lead are measured at each site
- Define
  - X = Traffic … explanatory (predictor) variable
    … it might explain the amount of Lead
  - Y = Lead … dependent (response) variable
    … it might depend on Traffic
- The data are (X, Y) pairs; generate a scatterplot

Information from a Scatterplot

- Stochastic or deterministic?
- Shape of the relationship?
  - Linear or curved?
- Direction (sign) of the relationship?
- Strength of the relationship?

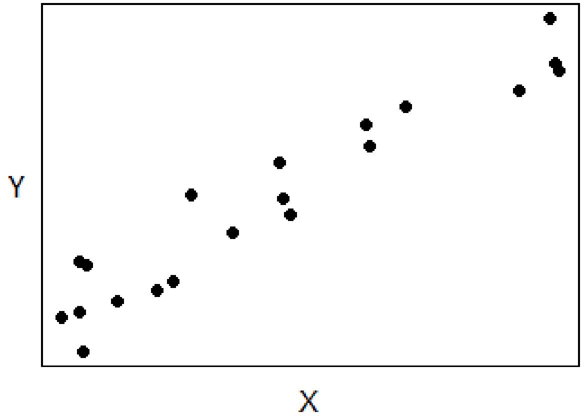


- We will return to this data after a brief introduction to regression analysis

Stochastic vs. Deterministic

Stochastic (Statistical)

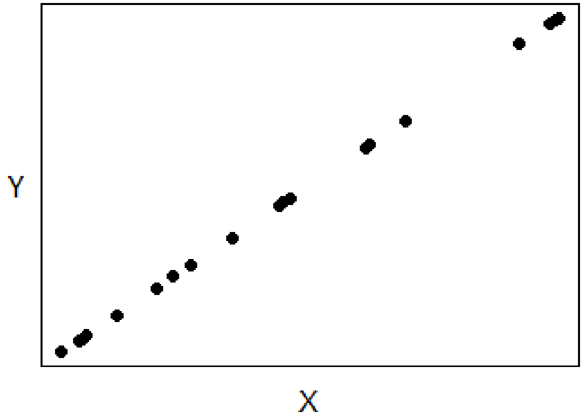
$$Y = f(X) + \varepsilon$$



Other things (besides X) can affect the value of Y.

Deterministic (Functional)

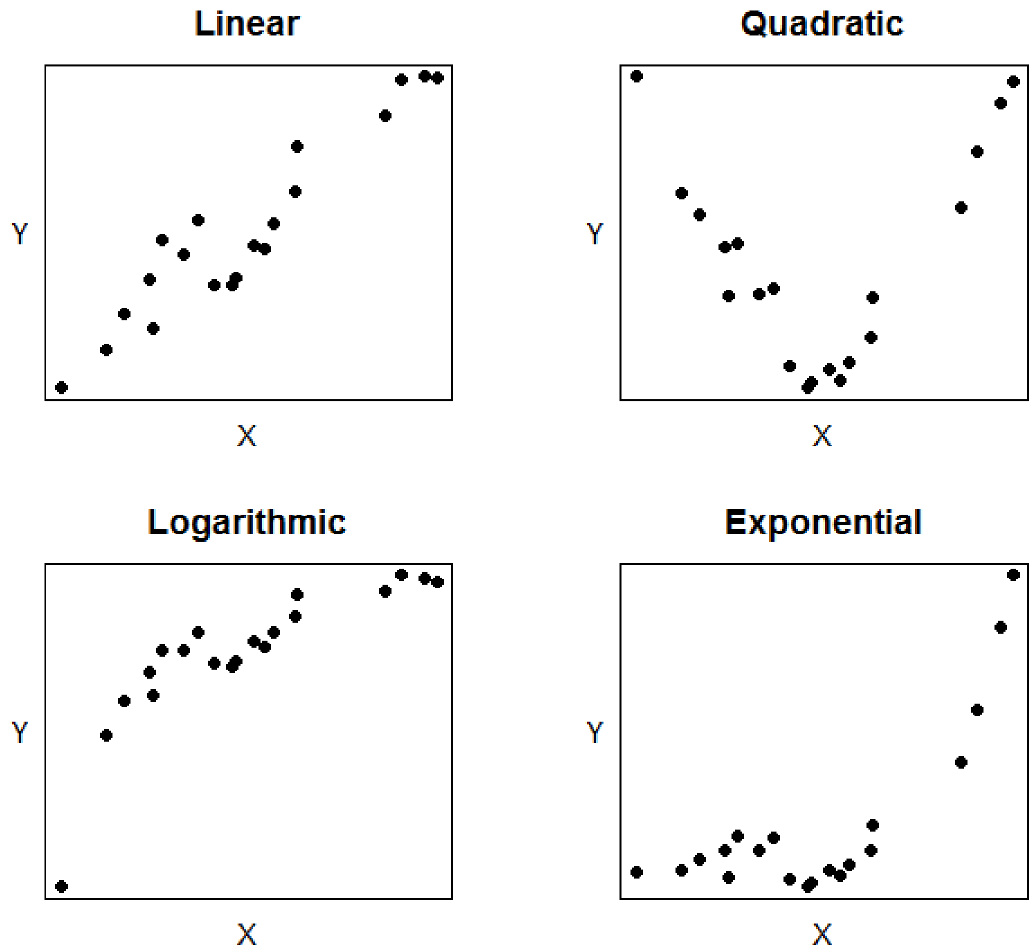
$$Y = f(X)$$



If we know the value for X, we know <u>exactly</u> the value for Y.

These scatterplots are for illustration only.
They are not related to the Lead vs. Traffic example.

Shape of the Relationship

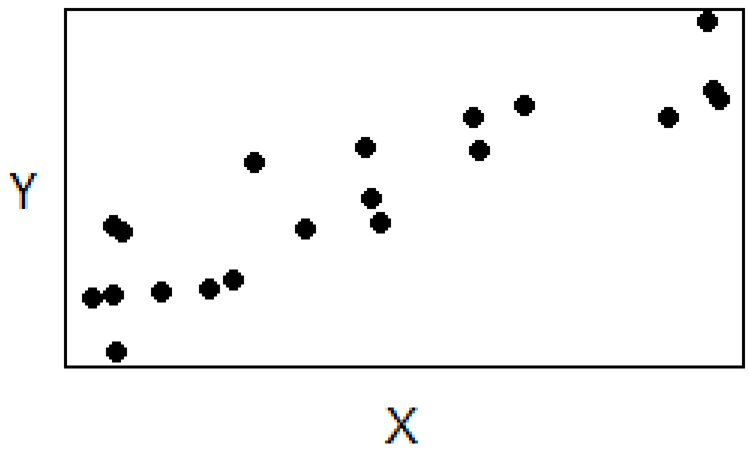


- Other shapes are possible
- We focus on linear

Direction of the Relationship

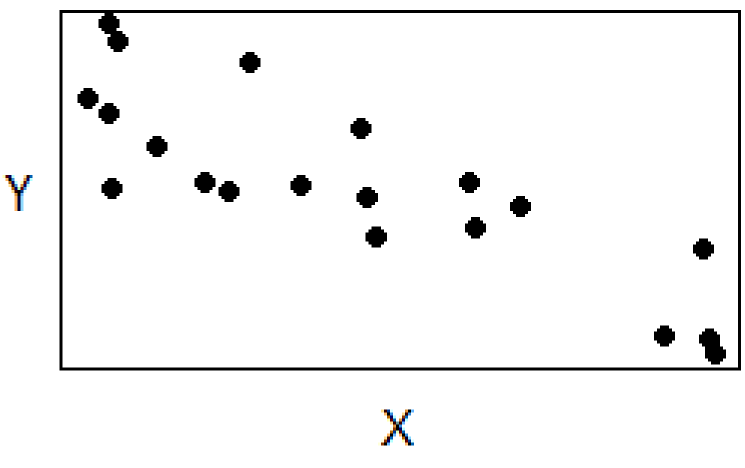
Positive

As X increases, Y increases



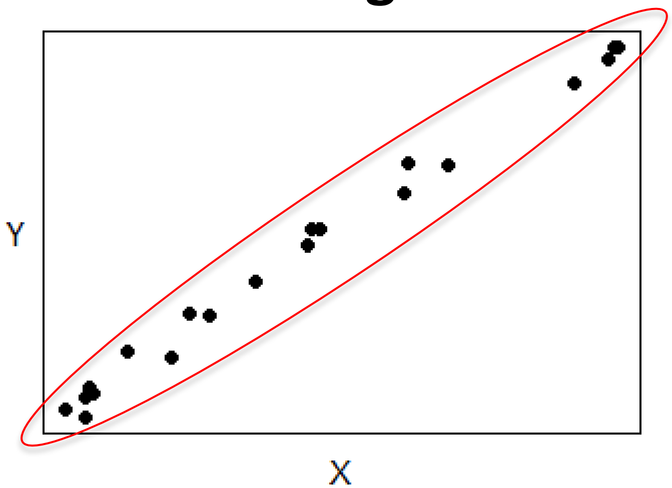
Negative

As X increases, Y decreases

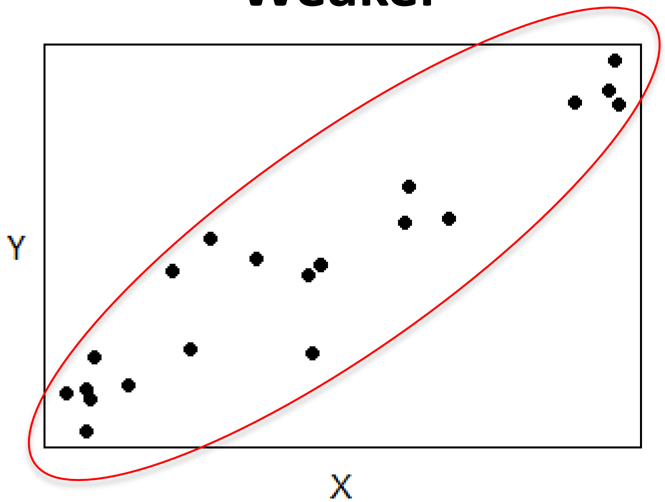


Strength of the Relationship

Stronger



Weaker



- How much variability in Y is explained by X?
- Source of unexplained variability

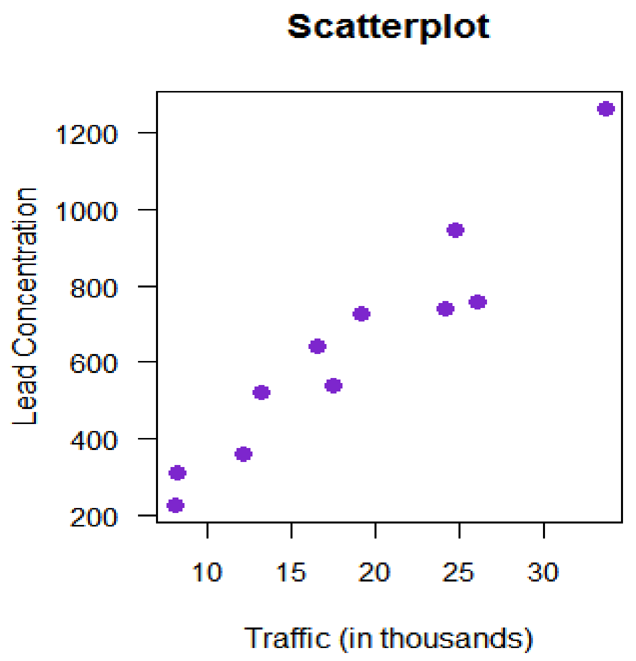
  Noise? … Measurement error? … Other variables that affect Y?

Other Examples

Y = Response Variable	X = Predictor Variable
Risk for heart disease	Cholesterol concentration
Sales of a product ($$$)	Advertising investment ($$$)
Cement compression strength	Water content of cement
Milk yield of dairy cows	Feed consumption
Person’s muscle mass	Age

Example: Lead vs. Traffic

- Relationship appears to be
  - ⇨Stochastic
  - ⇨Linear, positive
  - ⇨Fairly strong
- A simple linear regression model seems appropriate



$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0	intercept
β_1	slope
X_i	value of predictor (independent) variable for i^{th} experimental unit
ε_i	'error' for i^{th} experimental unit
Y_i	value of response (dependent) variable for i^{th} experimental unit

Note that these do <u>not</u> depend on i. There is one β_0 and one β_1 for all possible (x, y) pairs

Setting up the Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i is lead concentration in the tree bark at site i
- X_i is the traffic volume at site i
- β_0 and β_1 are parameters (intercept and slope) that define the linear relationship between X and Y
- ε_i is a leftover random error term
  - ⇨ "residual"
  - ⇨ unique to the i^{th} site
  - ⇨ measures how far "off" the model is from the observed Y

Interpreting the Model

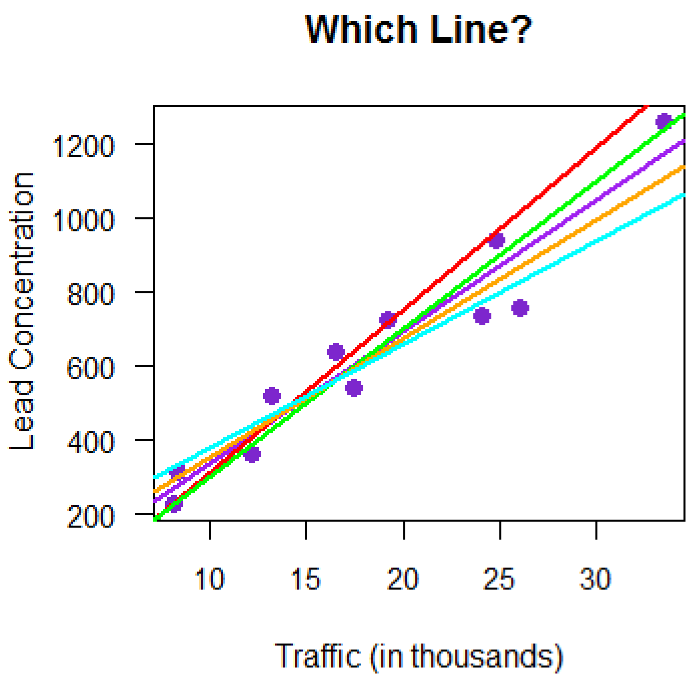
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Suppose a site has traffic flow equal to x
- The lead concentration at this site is expected to be $\beta_0 + \beta_1 x$
- But there can be other circumstances that affect the amount of lead, for example
  - tree characteristics (age, species, etc.)
  - location characteristics (nearby manufacturing plant, prevailing wind, etc.)
- These ‘other circumstances’ are captured by the error term ε, so the actual amount of lead is $\beta_0 + \beta_1 x + \varepsilon$

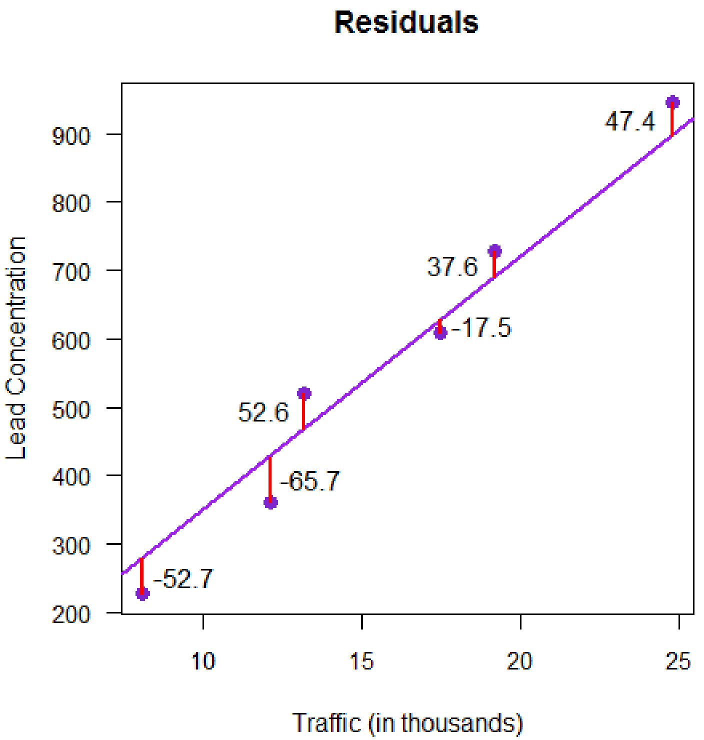
Line of “Best” Fit

- We want to find a line that is a ‘best’ fit to the data
- What is “best”?
- Use Least Squares criterion
  - minimize the sum of squared residuals

… what is a residual?



Residuals



Note: Only six of our (x, y) pairs are shown in this graph.

- Residual (i.e. 'error') for i^{th} observation is difference between the observed Y and the Y value on the line
- $r_i = Y_i - (\beta_0 + \beta_1 X_i)$
- We want to minimize

$$\sum r_i^2 = \sum \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

Calculating Least Squares Estimates

- Want to find the values of β_0 and β_1 that minimize

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2$$

- If you know calculus, take derivatives, set them to 0, and solve the simultaneous equations
- If you don't know calculus …

Least Squares Estimates

Basic Statistics	Sums of Squares
$n =$ number of (x, y) pairs	$SS_{XX} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\left(\bar{X}\right)^2$
$\bar{X} = \frac{1}{n}\sum X_i$	$SS_{YY} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\left(\bar{Y}\right)^2$
$\bar{Y} = \frac{1}{n}\sum Y_i$	$SS_{XY} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n} X_i Y_i - n\,\bar{X}\,\bar{Y}$

- slope estimate is $\hat{\beta}_1 = \dfrac{SS_{XY}}{SS_{XX}}$
- intercept estimate is $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Note: A bar on the top indicates the average. A caret ("hat") indicates an estimate.

Sums Needed for Least Squares

Site (i)	Traffic (X)	Lead (Y)	X^2	Y^2	X*Y
1	8.1	227	65.61	51,529	1,838.7
2	8.3	312	68.89	97,344	2,589.6
3	12.1	362	146.41	131,044	4,380.2
4	13.2	521	174.24	271,441	6,877.2
5	16.5	640	272.25	409,600	10,560.0
6	17.5	539	306.25	290,521	9,432.5
7	19.2	728	368.64	529,984	13,977.6
8	24.8	945	615.04	893,025	23,436.0
9	24.1	738	580.81	544,644	17,785.8
10	26.1	759	681.21	576,081	19,809.9
11	33.6	1263	1,128.96	1,595,169	42,436.8
Sums	203.5	7034	4,408.31	5,390,382	153,124.3

Estimated Slope and Intercept

Sums from the table

$\sum X_i = 203.5$ $\sum X_i^2 = 4,408.31$

$\sum Y_i = 7,034$ $\sum Y_i^2 = 5,390,382$ $\sum X_i Y_i = 153,124.3$

Sample means: $\bar{X} = \frac{1}{11}(203.5) = 18.5$ and $\bar{Y} = \frac{1}{11}(7034) = 639.45$

Sums of Squares: $SS_{XX} = 4,408.31 - (11)(18.5)^2 = 643.56$

$$SS_{YY} = 5,390,382 - (11)(639.45)^2 = 892,522.67$$

$$SS_{XY} = 153,124.3 - (11)(18.5)(639.45) = 22,996.23$$

$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{22,996.23}{643.56} = 35.7$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 639.45 - 35.7 * 18.5 = -21$

Summary

- For regression, the data consist of (X, Y) pairs
- Scatterplots can reveal the nature and strength of the relationship between X and Y
- The Least Squares criterion is used to generate the “best” line that describes the relationship between X and Y
- Know how to calculate the estimates for slope and intercept (including the table of sums)