STAT 410 - Linear Regression

Lecture 1

Meng Li

Department of Statistics

Aug. 22, 2017

Statistics: What and why?

- Statistics is the *prime* information science.
- It deals with the *extraction* of useful *information* from *data*, accounting for the appropriate *uncertainty*.
- Such information can help us understand how the natural world works, and make decisions.
- It has revolutionized virtually all areas of scientific investigation in the 20th century.
  - Biology, physics, psychology, economics, you name it, ...
- It has becoming ever more important in the past 10 years, due to the computerization of data generation.

Example I: Google and Facebook

- “[W]e create as much information in two days now as we did from the dawn of man through 2003."—Eric Schmidt in 2010.
- That is about five *exabytes* of data every two days, (and that’s back in 2010).
- What kind of information can be extracted:
  - The large: Disease epidemic trends, people’s social behavior, or even the current stage of the entire human civilization!
  - The small: What might the user like or want? (What ads are relevant? Hotels vs diapers?)
- From large to small, all need statistics!

Example II: The century of biology

- We now know that humans are coded in 3 billion DNA letters (A,T,C,G), but what do they mean?
- We know many diseases are highly inheritable (heart problems, cancer, diabetes, etc.) so information regarding these diseases must be contained in these 3 billion letters.
- If we can tell which genes contribute to the disease risk, then we can make *personalized* diagnosis and treatment.
- How to map these genes to diseases?
- We need statistics!

Statistical arbitrage in high frequency trading

- The prices of tens of thousands of securities are recorded every fraction of a second.
- Can we detect patterns (deviation from equilibrium) in the prices *quickly*, so that an arbitrage opportunity can be created in the next few seconds?
- We need statistics!

Moral of the story

- We are living in an era of data explosion. The data contain all kinds of information that can make the world a better (... or worse) place.
- Technological advances have made the generating, transporting, and storing of huge quantities of data possible.
- It is the statistician's job to make sense of the data.
- There has *never* been a better time to become a statistician.
  - A world of new exciting possibilities and challenges.

Prerequisites

- **Probability** and **Statistics** at the level of STAT 310.
- Statistics is the science of extracting information from data, accounting for the *uncertainty*.
- Probability is the tool for formulating *uncertainty* in a mathematical fashion.
- Probability modeling is the art of making the appropriate *assumptions* about the underlying randomness in the data.
  - For example: What family of distributions to use? What additional assumptions—such as independence, stationarity, etc?
  - “All models are wrong, but some are useful."—George Box.
- Based on these assumptions, statistics aims at “completing the picture"— drawing inference (“highly educated guess") about various aspects of the random mechanism under these assumptions.
- In real problems, statistical inference and probability modeling typically form a *two-way* process.

Linear regression: What and Why?

- Regression Analysis is a statistical technique for investigating and modeling the relationship between variables.
  - Random variables : (y, x)
  - y denotes the response or dependent variable
  - x is called explanatory/independent variable or predictor or regressor
  - Regression function: $E(y \mid x)$
- *Simple Linear* regression assumes: $E(y \mid x) = \beta_0 + \beta x$, or equivalently,

$$y = \beta_0 + \beta x + \varepsilon,$$

  where the random component ε has mean 0.
- A *multiple linear* regression model assumes

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon.$$

Linear regression: What and Why?

- Core values of Linear Regression (SVP):
  - Simplicity
  - Variability
  - Persistency
- Course website & Syllabus

Delivery time example I

- You are an industrial engineer employed by a soft drink beverage bottler to analyze the product delivery and service operations for vending machines. You suspect that the time required by a route deliveryman to load and service machines is related to the number of cases of product delivered. You visit 25 randomly chosen retail outlets having vending machines and observe the in-out delivery time (in minutes) and the volume of product delivered (in cases) for each.
- Why this problem is of our interest - Use of Regression:
  - Data description
  - Parameter estimation
  - Prediction
  - Control
- Then how to conduct this type of analysis?

Delivery time example II

1. Data visualization: scatter plots

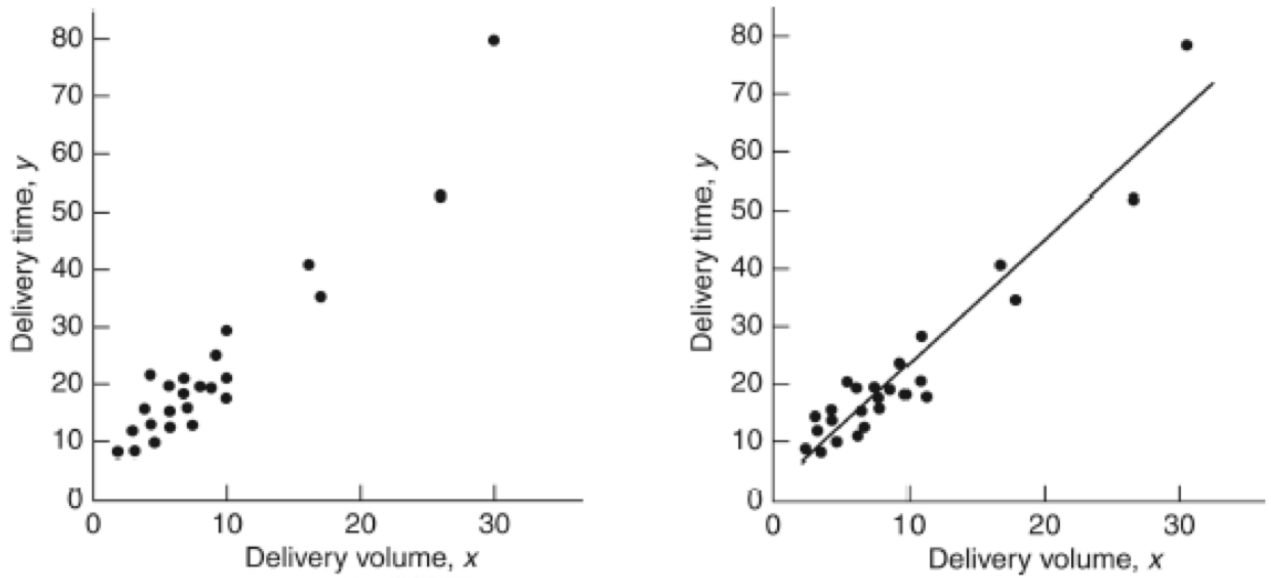

Figure: Scatter plot for the delivery data (left) and straight-line relationship between delivery time (y) and delivery volume (x).

Delivery time example III

2. Modeling:
   - $y = \beta_0 + \beta_1 x$?
   - The data points do not fall *exactly* on a straight line
   - We use a simple linear regression model (SLR):

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

   - Terminology:
     - y - dependent (response) variable
     - x - independent (regressor/predictor) variable
     - β_0 - intercept
     - β_1 - slope
     - ε - random error term with mean 0
   - SLR implies $\mathrm{E}(y|x) = \beta_0 + \beta_1 x$
   - If ε has variance σ^2, then $\mathrm{Var}(y|x) = \sigma^2$ (Why?)
3. Model fitting - estimation of parameters
4. Model checking
5. Model interpretation and decision-making (data summary, prediction, control, ect.)