Data Modeling: CSCI E-106

Harvard Extension School

Introduction and Chapter 1

Course Information

Teaching Staff:

Instructor: Dr. Hakan Gogtas

Teaching Assistants:

- Petr Chovanec
- Andrea Hatch
- Dr. Ahmet Sezer

Course Information, cont'd

- Class Dates: Mondays, 7:20 - 9:20 pm EST
  - There will be a 5 min break at 8:20 pm EST
- TA Sessions: Thursdays, 7:20 - 9:20 pm EST.
- Office hours are right after the lecture or by appointment
- Midterm Exam Date: October 18th to October 21st
- Final Exam Date: December 16th to December 17th

Course Information, cont'd

- Textbook :
  - [Required] Applied Linear Statistical Models, 5th Edition, by Kutner, Nachtsheim, Neter & Li
  - [Optional] Mastering Scientific Computing with R, by Paul Gerrard, Radia M. Johnson
  - [Optional] Linear Models with R, Julian J. Faraway

- Computing :
  - Create a free account at http://www.rstudio.cloud
  - Required to generate PDF reports using the knitr R package for homework, midterm and final exam.

Course Information, cont'd

Grading:

Assignments	30%
Midterm Exam	30%
Final Exam	40%

- The worst homework score will be dropped
- No Late Homework! No Exceptions!

Course Information, cont'd

Suggestions:

- Watch the recording for all classes and TA sections if you can't attend
- TA sessions will be mainly focused on doing exercised in R
- Read the Chapter's before each class and TA sessions
- Email your questions if you can't attend the class, we will go over your questions and you can see your questions answered on the recordings
- Time management is important

Applied Linear Statistical Models
Chapter 1

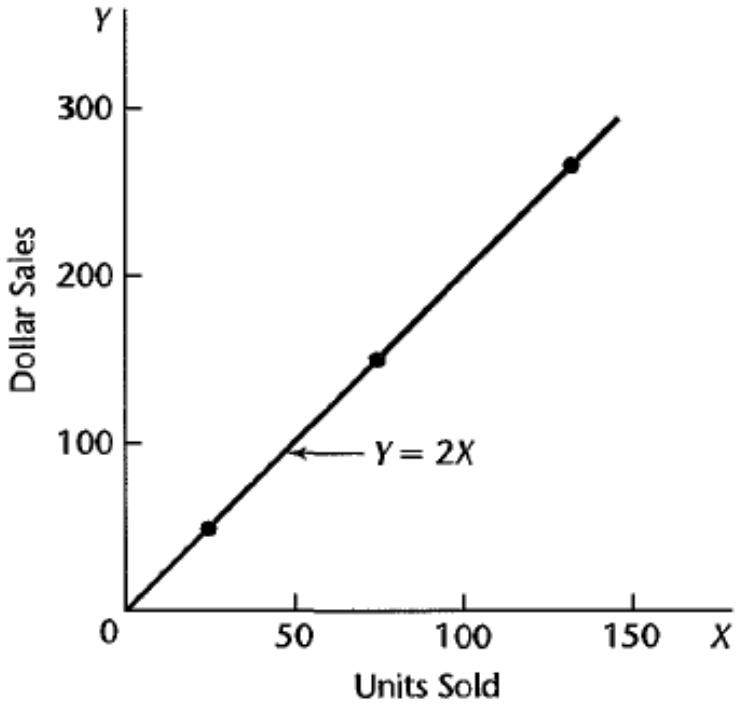
Relation Between Variables

- A functional relation between two variables is expressed by a mathematical formula. If X denotes the independent variable and Y the dependent variable, a functional relation is Y=f(X).
- Given a particular value of X, the function f indicates the corresponding value of Y.

Example 1

- Consider the relation between dolar sales (Y) of a product sold at a fixed price and number of units sold (X). If the selling price is $2 per unit, the relation is expressed by the equation: Y=2X .
- Data is below

Period	Number of Units Sold	Dollar Sales
1	75	$150
2	25	50
3	130	260



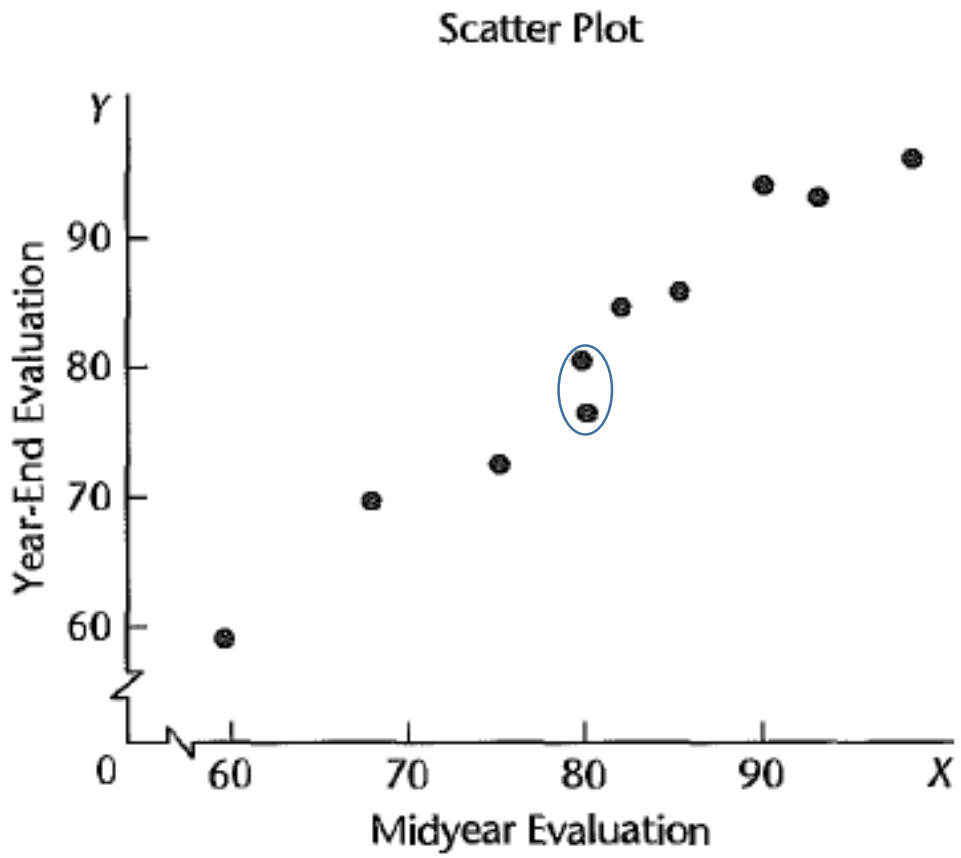
Statistical Relation between Two Variables

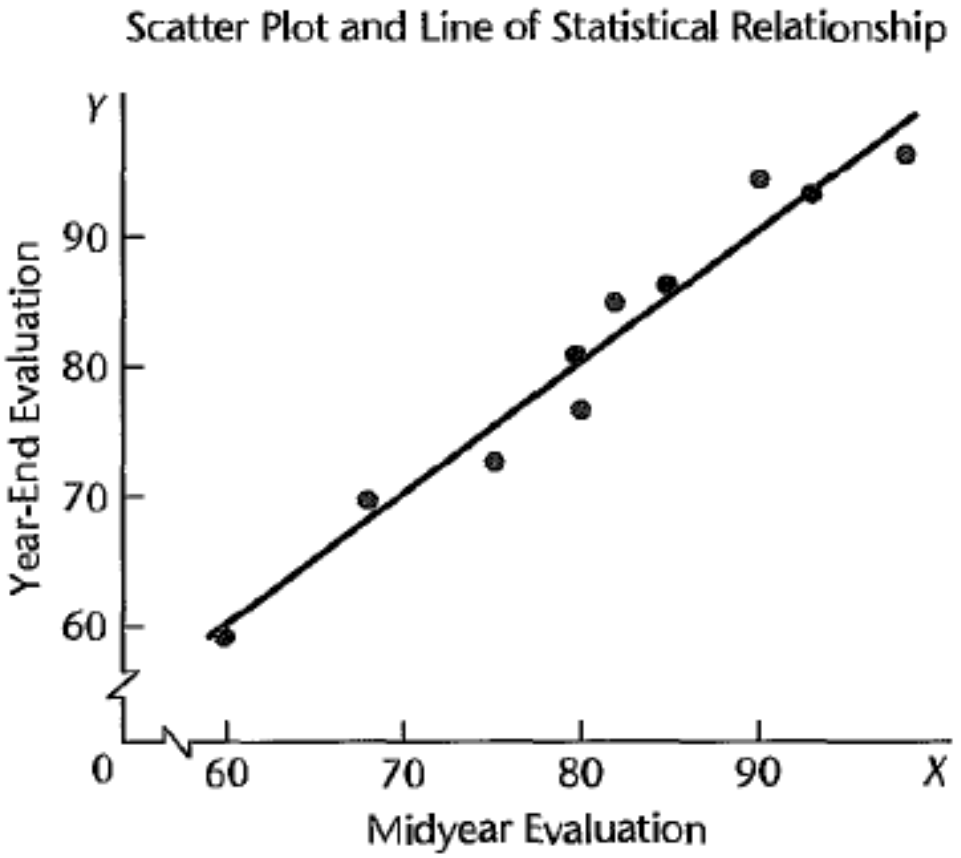
- A statistical relation, unlike a functional relation, is not a perfect one. In general, the observations for a statistical relation do not fall directly on-the curve of relationship.

- Example 2: Performance evaluations for 10 employees were obtained at midyear and at year-end. These data are plotted in Figure 1.2a. Year-end evaluations are taken as the dependent or response variable Y, and midyear evaluations as the independent, explanatory, or predictor variable X.

Example 2, cont'd

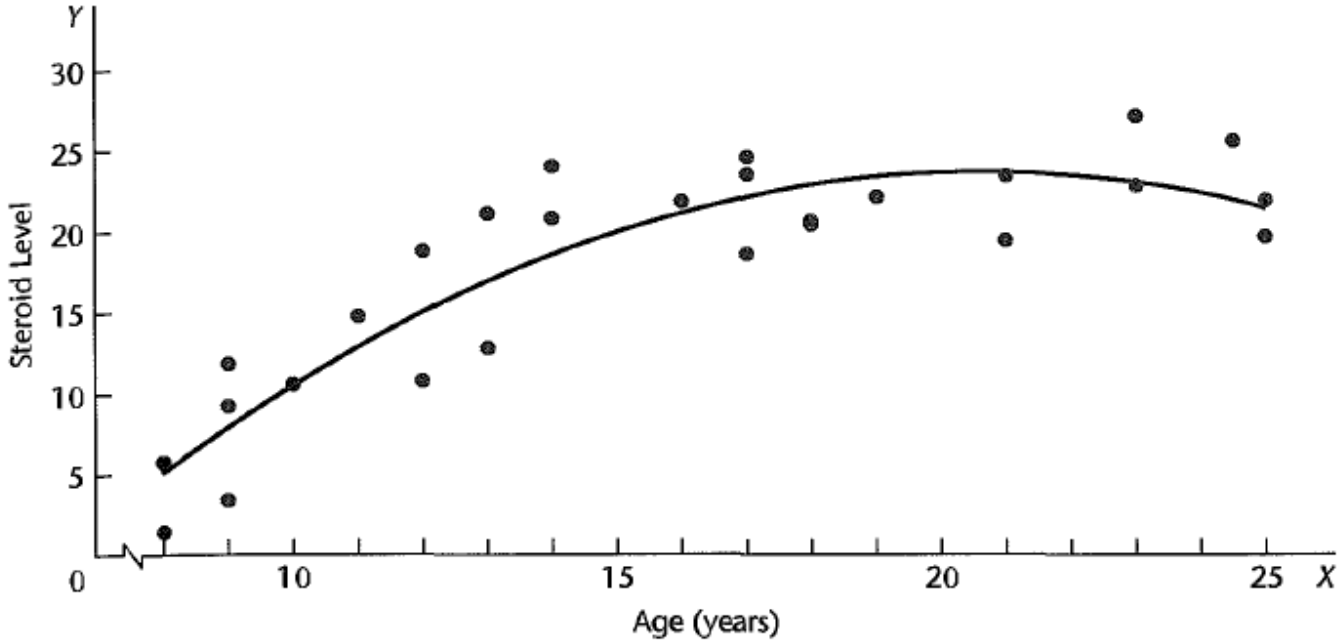
- Which kind of pattern can be observed in the figure?
- What kind of the relation between midyear and year-end evaluations?





Example 3

- Data, on age and level of a steroid in plasma for 27 healthy females between 8 and 25 years old, strongly suggest that the statistical relationship is curvilinear (not linear).



- The curve of relationship implies that, as age increases, steroid level increases up to a point and then begins to level off.
- Note again the scattering of points around the curve of statistical relationship, typical of all statistical relations.

What is Regression?

A regression model is a formal means of expressing the two essential ingredients of a statistical relation:

1. 1. A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
2. 2. A scattering of points around the curve of statistical relationship.

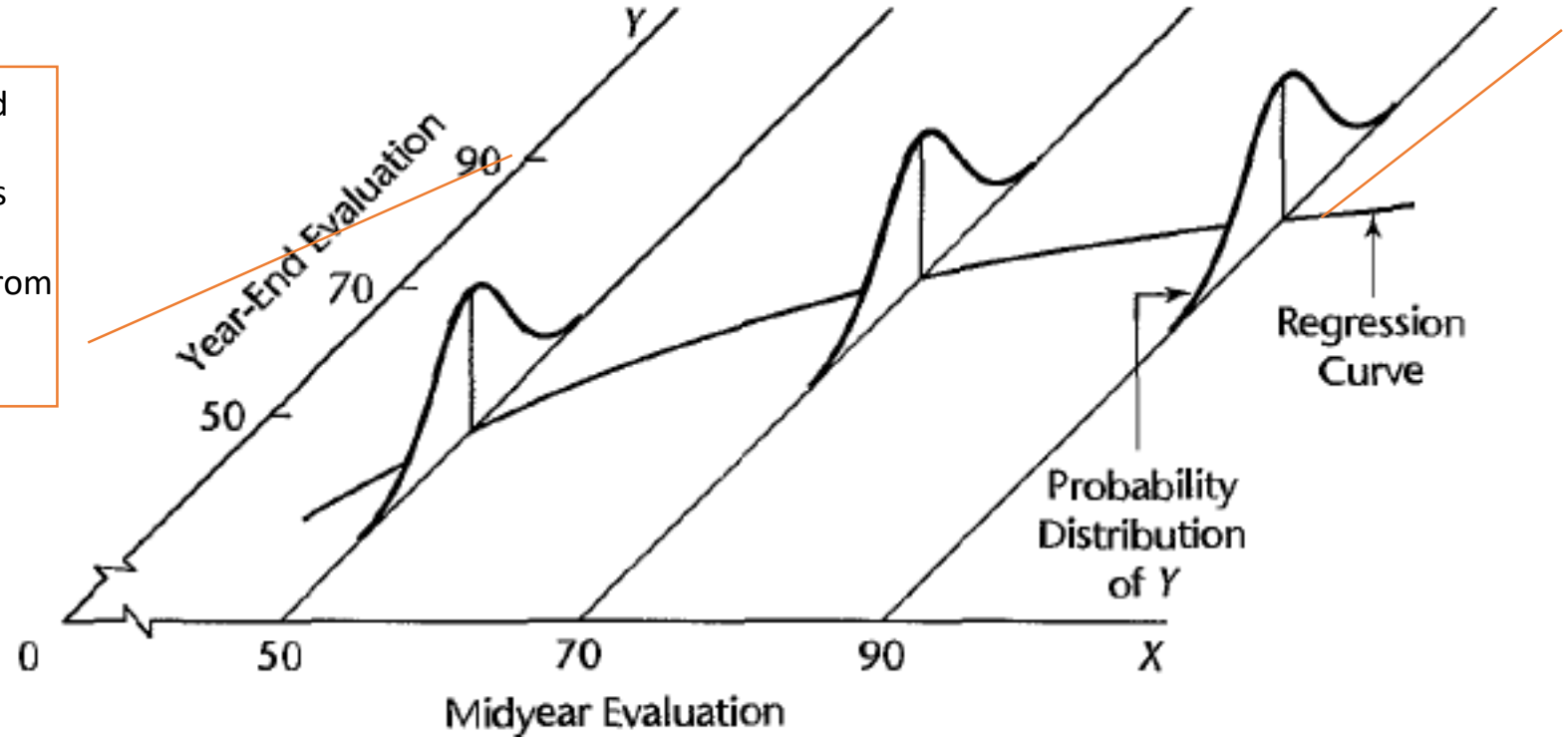
These two characteristics are embodied in a regression model by postulating that:

1. There is a probability distribution of Y for each level of X.
2. The means of these probability distributions vary in some systematic fashion with X.

Example 2, cont'd

- Consider again the performance evaluation example. The year-end evaluation Y is treated in a regression model as a random variable. For each level of midyear performance evaluation, there is postulated a probability distribution of Y..

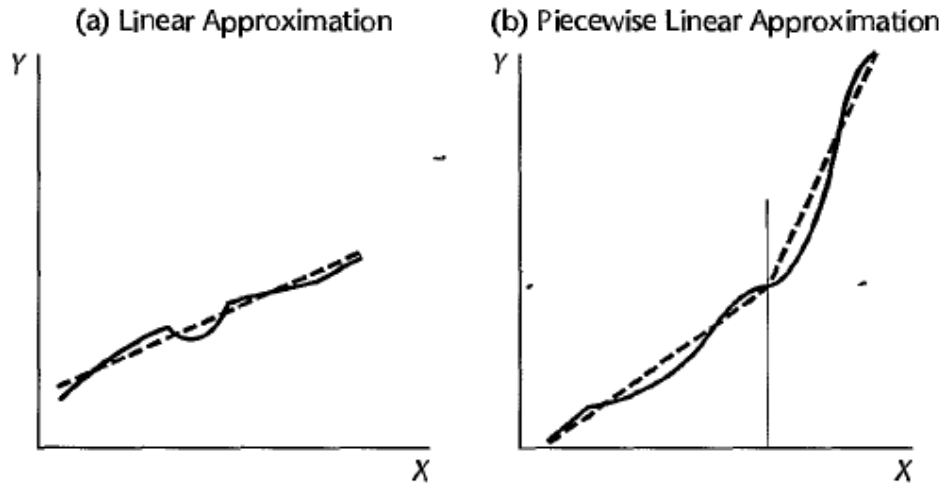
The actual year-end evaluation of this employee, Y = 94, is then viewed as a random selection from this probability distribution.



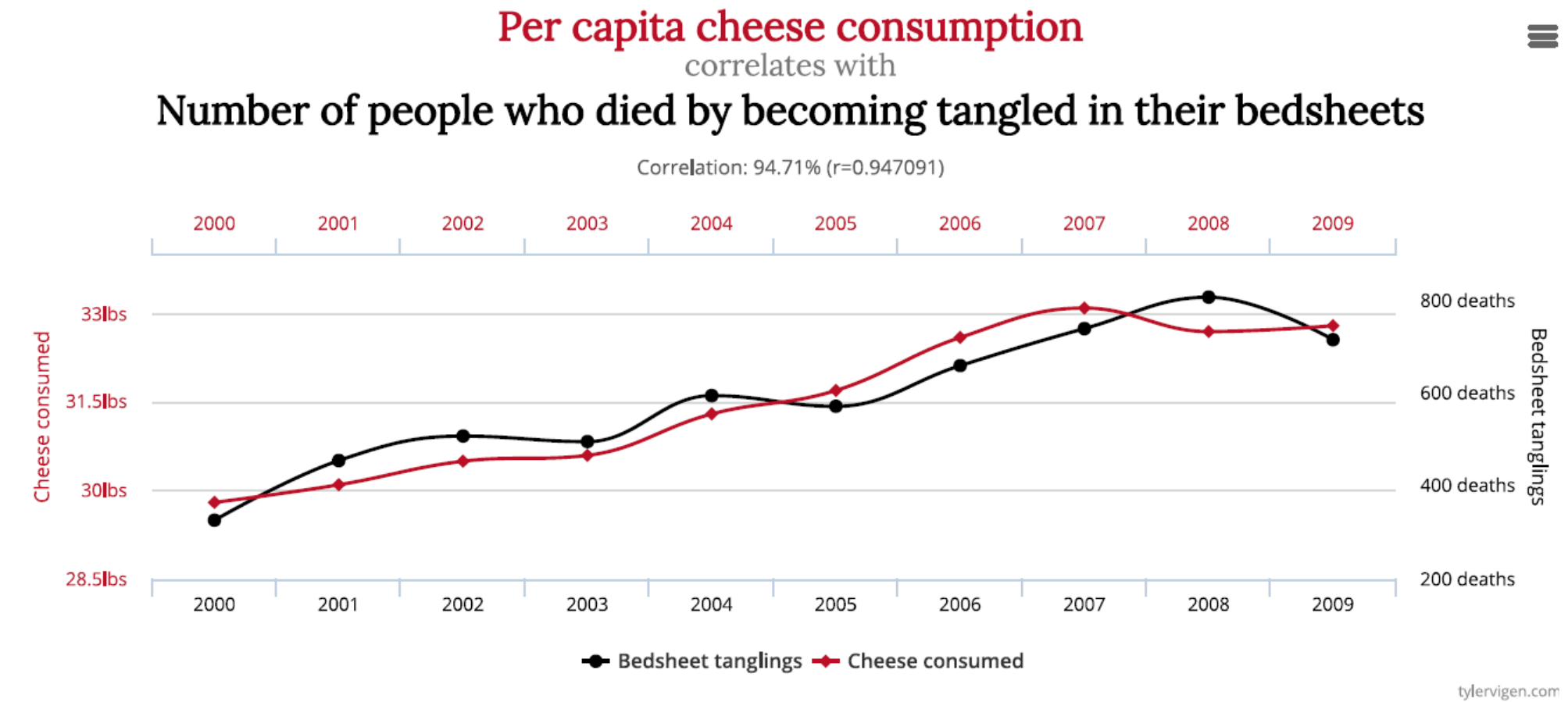
shows such a probability distribution for X = 90, which is the midyear evaluation for the first employee

Construction of Regression Models

- Definition of a Dependent Variable or Problem: Regression models could be used for different purposes
  - Analyzing data
  - Prediction
  - Comparison of Means
  - Quality Controls
  - Experimental Designs, and etc.…
- Data Availability and Data Quality: What are independent variables? And are they accurate?
- Functional relationships linear vs. non linear?



Regression and Causality



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Regression Model Definition

The linear regression function with one predictor variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

- Y_i: the value of response variable in the ith trial
- β_0, β_1: parameters
- X_i: a known constant; the value of the predictor variable in the ith trial
- ε_i: random error term; $E\{\varepsilon_i\} = 0$; $\sigma^2(\varepsilon_i) = \sigma^2$; uncorrelated ($\sigma\{\varepsilon i, \varepsilon j\} = 0$, i≠j)

Features

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ is the sum of two components:

1. $\beta_0 + \beta_1 X_i$ ----------> *Constant Term*
2. ε_i ----------> *Random Term*

$E\{\varepsilon_i\} = 0$:

- $E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \beta_0 + \beta_1 X_i$

- The regression function:

$$E\{Y\} = E\{\beta_0 + \beta_1 X + \varepsilon\} = \beta_0 + \beta_1 X$$

- The regression function relates the means of the probability distribution of Y for given X to the level of X

Features, Cont'd

- Y_i in the ith trial exceeds or falls short of the value of the regression function by the error term amount ε_i
- $\sigma^2\{\varepsilon i\} = \sigma^2$ *and* $(\sigma\{\varepsilon_i, \varepsilon_j\} = 0, i \neq j)$
- $\sigma^2\{Y_i\} = \sigma^2(\beta_0 + \beta_1 X_i + \varepsilon_i)$

  $= \sigma^2(\beta_0 + \beta_1 X_i) + \sigma^2(\varepsilon_i)$

  $= \sigma^2\{\varepsilon_i\} = \sigma^2$
- The error terms are assumed to be uncorrelated, so are the responses Y_i and Y_j.

Example

- A consultant for an electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids. Suppose that regression model is applicable and is as follows:

$$Yi = 9.5 + 2.1X_i + \varepsilon_i \quad \text{and} \quad E\{Y\} = 9.5 + 2.1\,X$$

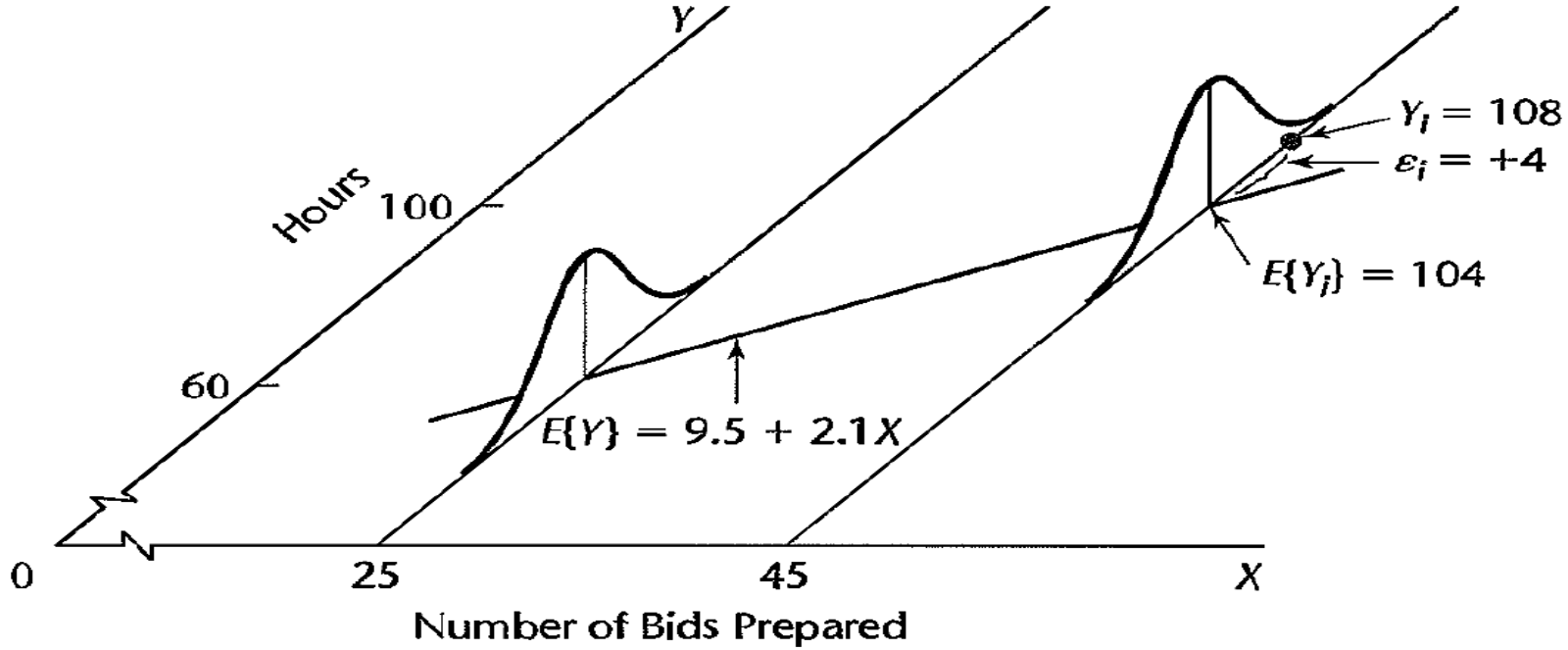
  - Y is the number of hours required to prepare the bids
  - X is the number of bids prepared in a week

- Suppose that in the ith week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. In that case, the error term value is $\varepsilon_i = 4$, for we have

$$E\{Y_i\} = 9.5 + 2.1(45) = 104 \text{ and } Y_i = 108 = 104 + 4$$

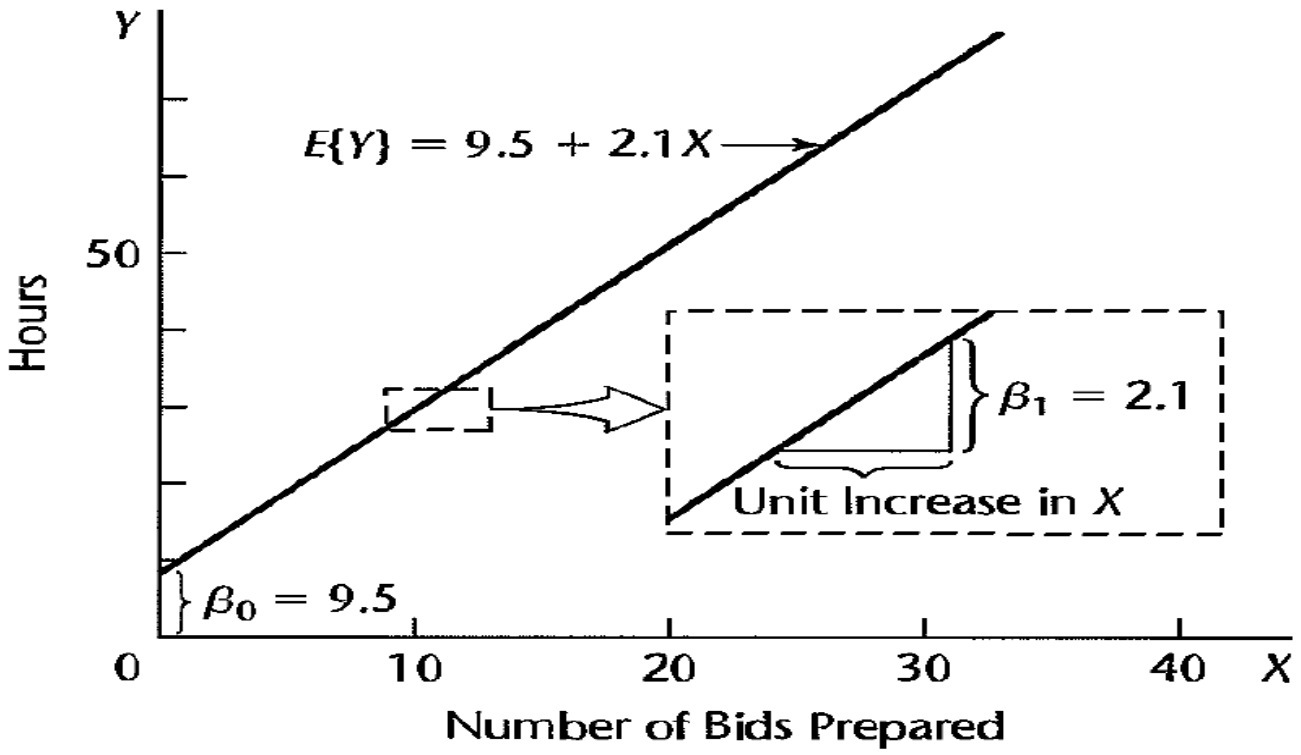
Example, cont'd

- Suppose that in the i^{th} week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. In that case, the error term value is $\varepsilon_i = 4$, for we have $E\{Y_i\} = 9.5 + 2.1(45) = 104$ and $Y_i = 108 = 104 + 4$



Meaning of Regression Parameters

Regression coefficients: β_0(*intercept*), β_1(*slope*)



- β_0 gives the mean of the probability distribution of Y at X = 0. When the scope of the model does not cover X = 0, β_0 does not have any particular meaning as a separate term in the regression model.
- β_1 indicates that the preparation of one additional bid in a week leads to an increase in the mean of the probability distribution of Y of 2. 1 hours.

Alternative Model Specification

- Original regression model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \ldots, n$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_1 \bar{X} + \varepsilon_i \\
&= (\beta_0 + \beta_1 \bar{X}) + \beta_1(X_i - \bar{X}) + \varepsilon_i \\
&= \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i
\end{aligned}
$$

Thus, this alternative model version is:

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i$$

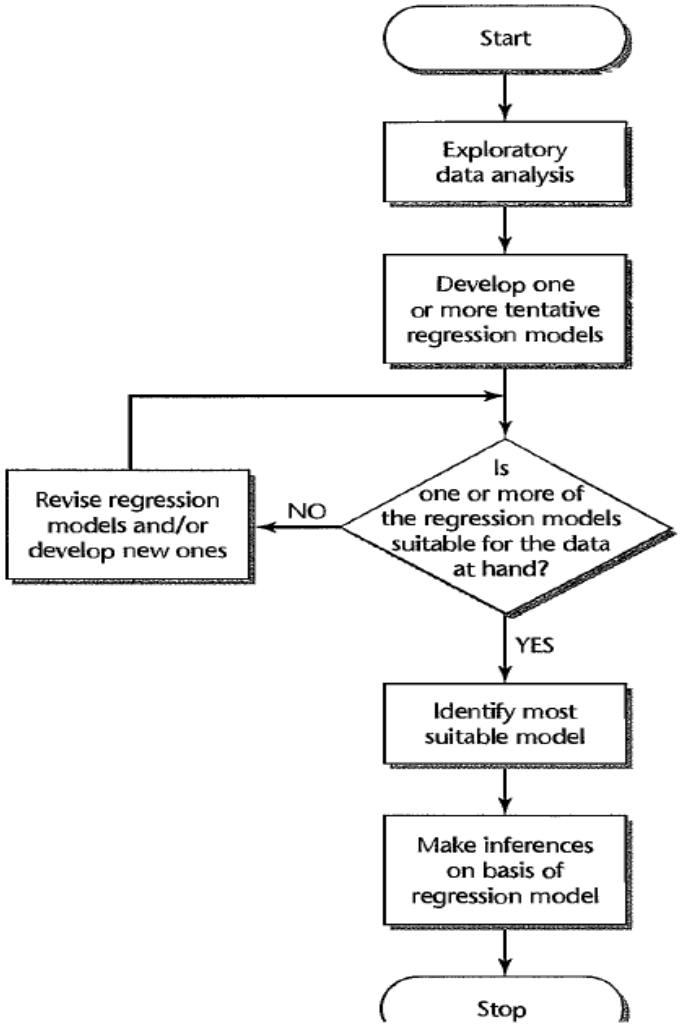
where:

$$\beta_0^* = \beta_0 + \beta_1 \bar{X}$$

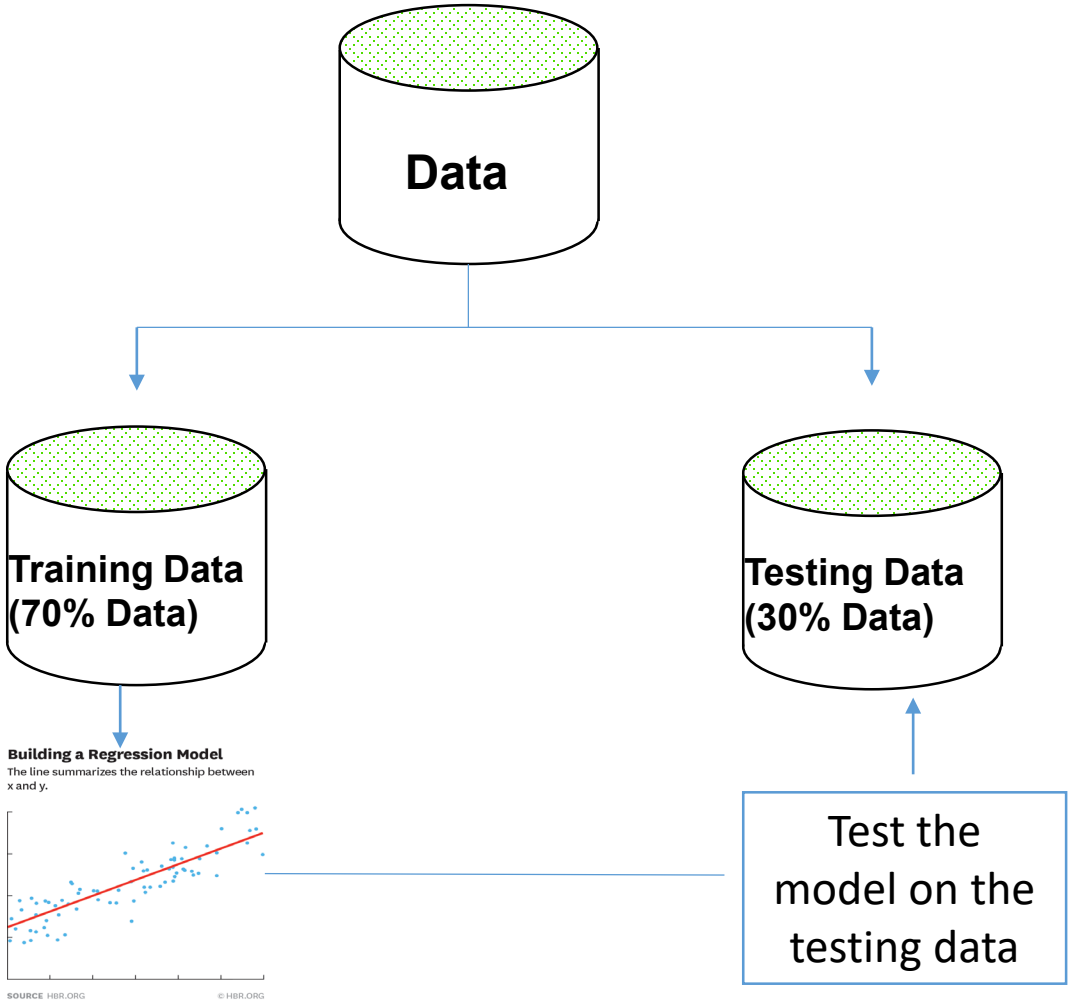
Data From Regression Analysis

- Unknown the regression parameters β_0, β_1
- Estimate parameters from relevant data
- Rely on an analysis of the data for developing a suitable regression model

Steps in Regression Analysis



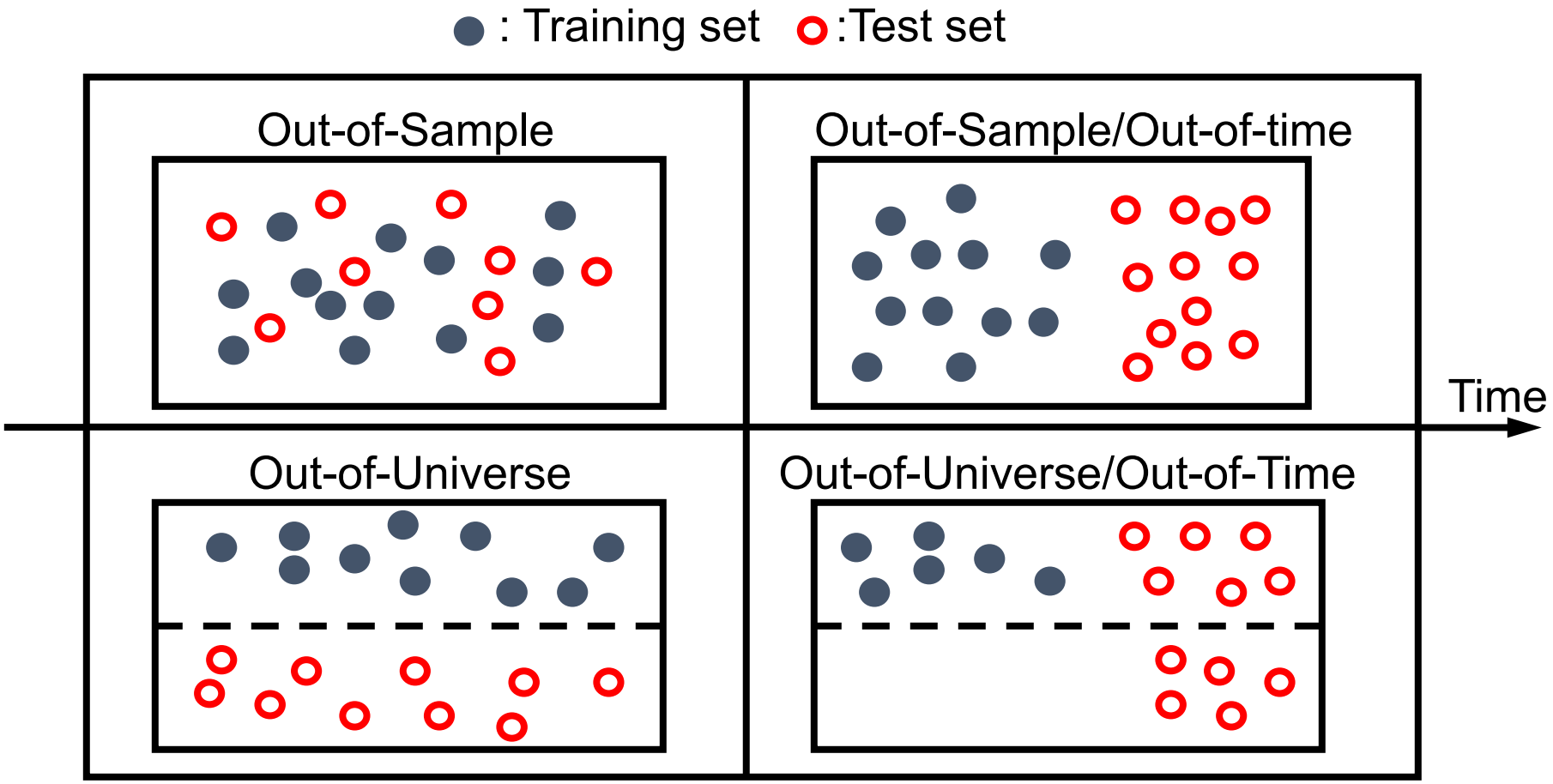
Steps in Regression Analysis, cont'd



- Training Data is also called development sample.
- Testing Data is also called hold-out sample

Steps in Regression Analysis, cont'd

- **Out-of-Sample Tests**



Estimate: Method of Least Squares

- For the observations (X_i, Y_i) for each case, the method of least squares considers the deviation of Y_i from its expected value:

$$Y_i - (\beta_0 + \beta_1 X_i)$$

- In particular, the method of least squares requires that we consider the sum of the n squared deviations. This criterion is denoted by Q:

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

- According to the method of least squares, the estimators of β_0 and β_1 are those values b_o and b_1 respectively, that minimize the criterion Q for the given sample observations (X_1 Y_1), (X_2, Y_2), ... , (X_n, Y_n).

Example

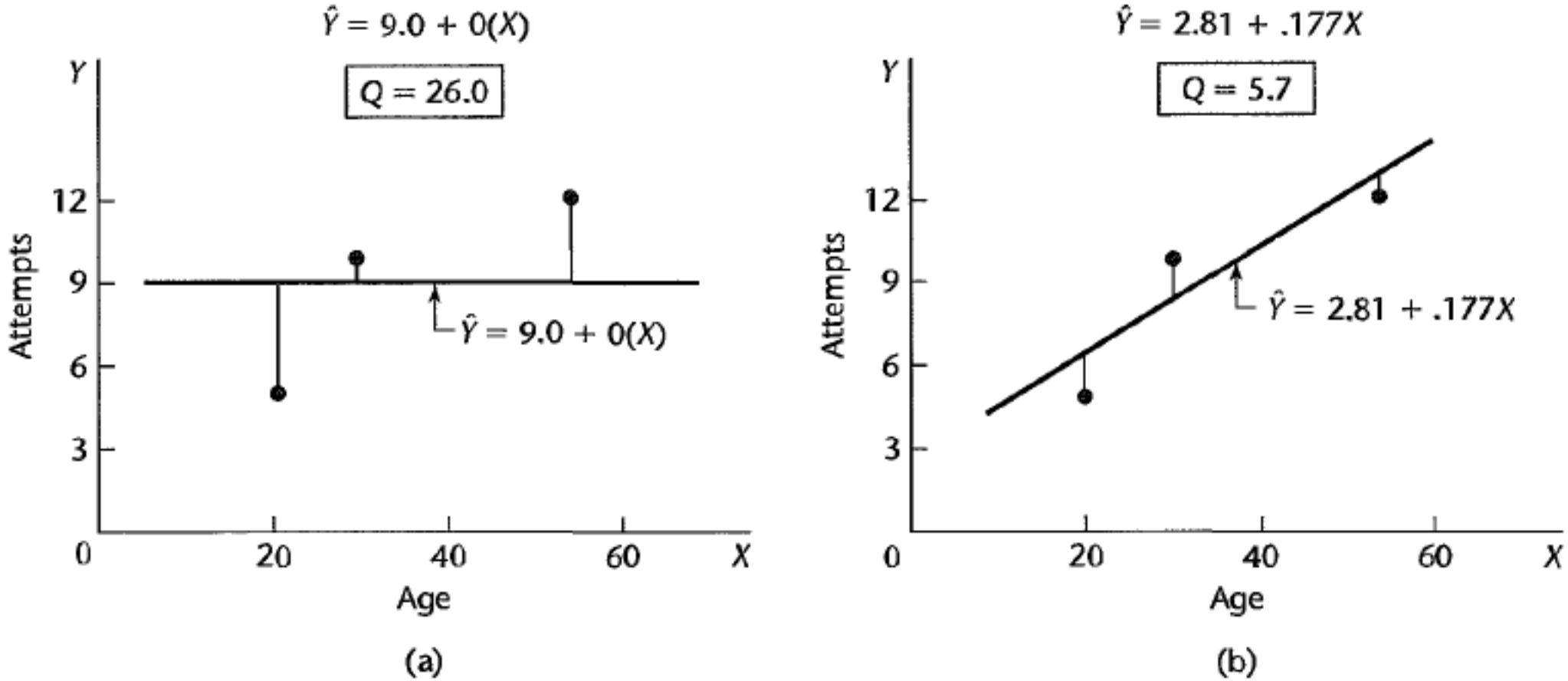
In a small-scale study of persistence, an experimenter gave three subjects a very difficult task. Data on the age of the subject (X) and on the number of attempts to accomplish the task before giving up (Y) follow:

Subject i:	1	2	3
Age X_i:	20	55	30
Number of attempts Y_i:	5	12	10

In terms of the notation to be employed, there were $n = 3$ subjects in this study, the observations for the first subject were $(X_1, Y_1) = (20, 5)$, and similarly for the other subjects.

Example, cont'd

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.



Least Squares Estimation

- The property of "Good" estimators?
- The least squares estimators b_0, b_1 minimize the criterion Q for the given sample observations.
- How to obtain the estimators b_0, b_1 ?

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

Goal: select values of β_0, β_1 that minimize Q and label them as b_0, b_1

$$(i): \frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)(-1) \overset{\text{set}}{=} 0 \Rightarrow \sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$(ii): \frac{\partial Q}{\partial \beta_1} = 2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)(-X_i) \overset{\text{set}}{=} 0 \Rightarrow \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

Least Squares Estimation , Cont'd

Solving (by multiplying (i) by $\sum_{i=1}^{n} X_i$ and (ii) by n and taking $(ii)-(i)$):

$$n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i = b_1 \left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right) \Rightarrow$$

$$\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} = b_1 \left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i \right)^2}{n} \right) \Rightarrow$$

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i \right)^2}{n}} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2} = \frac{SS_{XY}}{SS_{XX}} = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X} \right)}{SS_{XX}} Y_i = \sum_{i=1}^{n} k_i Y_i$$

From (i): $b_0 = \overline{Y} - b_1 \overline{X} = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{\overline{X}\left(X_i - \overline{X} \right)}{SS_{XX}} \right) Y_i = \sum_{i=1}^{n} l_i Y_i$

Properties of Least Squares Estimation

- An important theorem, called the GaussMarkov theorem, states: Under the conditions of regression model, the least squares estimators b_o and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

  - Unbiased: $E\{b_0\} = \beta_0; \quad E\{b_1\} = \beta_1$
  - More Precise: The estimators b_0 and b_1 are more precise than any other estimators belonging to the class of unbiased estimators that are linear functions of the observations $Y_1, \ldots, Y_n$. Linear function of Y_i:

    - $b_1 = \sum k_i Y_i$
    - $b_0 = \sum l_i Y_i$
    - $E(Y_i) = \sum (l_i + k_i) Yi$

- Since the k_i and l_i are known constants (because the Xi are known constants), b_o and b_1 are a linear combination of the Yi and hence are linear estimators.

Point Estimation of a Mean Response

- response: a value of the response variable Mean response: $E\{Y\}$
- Estimated regression function:

$$\hat{Y} = b_0 + b_1 X$$

($\hat{Y}$: the value of the estimated regression function at X of the predictor variable and an unbiased estimator of $E\{Y\}$)

Fitted value $\hat{Y}_i$:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad i = 1, \ldots, n$$

Fitted Values and Residuals

True Regression Function: $E\{Y\} = \beta_0 + \beta_1 X$ (Unknown, since $\beta_0, \beta_1 \equiv$ parameters)

Estimated Regression Function (Fitted): $\hat{Y} = b_0 + b_1 X$

For the i^{th} observation: $\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \ldots, n$

Residuals: Differences between observed and fitted (predicted) values:

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) \quad i = 1, \ldots, n$$

Properties of Residuals:

$$\sum_{i=1}^{n} e_i = 0 \quad (\text{From LS eq } (i))$$

$$\sum_{i=1}^{n} X_i e_i = 0 \quad (\text{From LS eq } (ii))$$

$$\Rightarrow \sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} (b_0 + b_1 X_i) e_i = b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i = 0$$

Estimating Error Variance σ^2

$$\sigma^2 = \sigma^2\{\varepsilon\} = E\left\{\left(\varepsilon - E\{\varepsilon\}\right)^2\right\} = E\left\{\left(\varepsilon - 0\right)^2\right\} = E\left\{\varepsilon^2\right\}$$

ε unobservable since $\varepsilon = Y - \left(\beta_0 + \beta_1 X\right)$

We use residual e to "estimate" ε

$$e = Y - \hat{Y} = Y - \left(b_0 + b_1 X\right)$$

Obtain the "average" squared residual to estimate σ^2 :

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

Normal Error Regression Model

The normal error regression model:

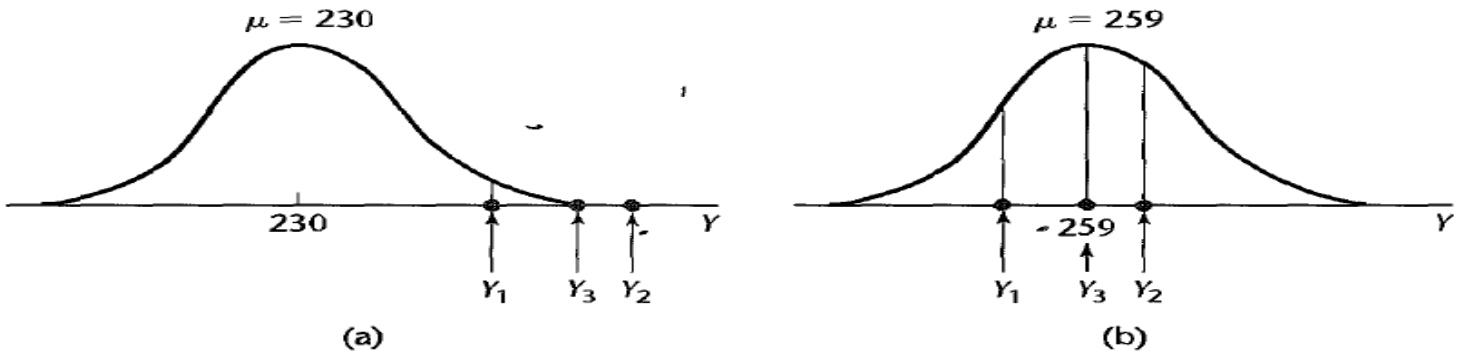
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i: the observation response
- X_i: a known constant
- β_0, β_1: parameters
- ε_i, $i = 1, \ldots, n$: independent $N(0, \sigma^2)$

The estimators of the parameters β_0, β_1 and σ^2 van be estimated be the method of *maximum likelihood*. (MLE)

Normal Error Regression Model, cont'd

FIGURE 1.13
Densities for Sample Observations for Two Possible Values of μ: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.



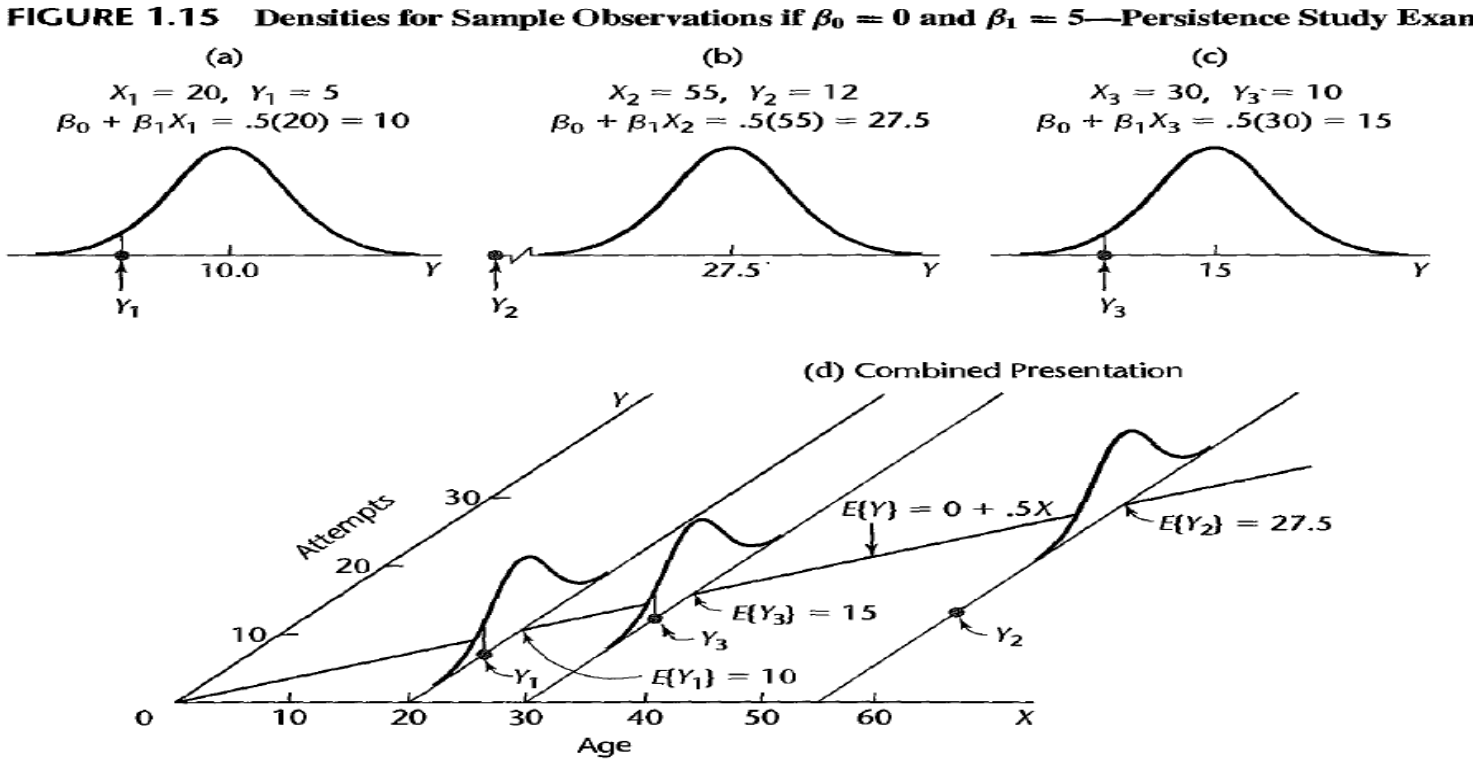
The method of maximum likelihood chooses as the maximum likelihood estimate that value for which the likelihood value is largest.

Two methods for finding MLE:

1. a systematic numerical search
2. use of an analytical solution

Estimator of μ is the sample mean $\bar{Y}$

Normal Error Regression Model, cont'd



$\sigma = 2.5; \beta_0 = 0; \beta_1 = 0.5$

Normal Error Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1,\ldots,n \quad \varepsilon_i \sim N\left(0,\sigma^2\right) \quad \text{(independent)}$$

$$f\left(y_i\right) = f_i = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{y_i - \left(\beta_0 + \beta_1 X_i\right)}{\sigma}\right)^2\right\} \quad i = 1,\ldots,n$$

$$L\left(\beta_0,\beta_1,\sigma^2\right) = \prod_{i=1}^{n} f_i = \left(2\pi\right)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \left(\beta_0 + \beta_1 X_i\right)}{\sigma}\right)^2\right\}$$

Goal: Choose values $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ that maximize L (or equivalently $l = \ln(L)$):

$$l = \left(-\frac{n}{2}\right)\ln\left(2\pi\right) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \left(\beta_0 + \beta_1 X_i\right)}{\sigma}\right)^2$$

Note: maximizing l wrt β_0, β_1 is same as minimizing $\sum_{i=1}^{n}\left(\frac{y_i - \left(\beta_0 + \beta_1 X_i\right)}{\sigma}\right)^2$

$$\Rightarrow \hat{\beta}_0 = b_0, \quad \hat{\beta}_1 = b_1$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) - (-1)\frac{1}{2}\sum_{i=1}^{n}\frac{\left(y_i - \left(\beta_0 + \beta_1 X_i\right)\right)^2}{\left(\sigma^2\right)^2} \overset{set}{=} 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n} = \frac{n-2}{n}s^2$$

Example: Toluca Company

The Toluca Company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size for producing this part. The production of this part involves setting up the production process (which must be done no matter what is the lot size) and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and labor hours required to produce the lot To determine this relationship, data on lot size and work hours for 25 recent production runs were utilized.

The production conditions were stable during the six-month period in which the 25 runs were made and were expected to continue to be the same during the next three years, the planning period for which the cost improvement program was being conducted.

R Studio Cloud - R Commands

Upload Toluca Data to R Studio Cloud


```
# attaching the data frame
attach(toluca_data)
# fitting the regression model
toluca.reg <- lm(workhrs ~ lotsize)
# getting the summary regression output:
toluca.reg <- lm(workhrs ~ lotsize)
# getting the ANOVA table:
anova(toluca.reg)
# getting the fitted values:
fitted(toluca.reg)
# getting the residuals:
names(toluca.reg)
resid(toluca.reg)
toluca.reg$residuals
plot(lotsize, workhrs)
# overlaying the regression line on this scatter plot:
abline(toluca.reg)
```

Example: Antioxidant Data – One Factor Analysis

- Dataset: lager_antioxidant_reg.csv

- Source: H. zhao, H. Li, G. Sun, B. Yang, M. Zhao (2013). "Assessment of Endogenous Antioxidative Compounds and Antioxidant Activities of Lager Beers," Journal of the Science of Food and Agriculture, Vol. 93, pp. 910-917.
- Description: Total phenolic content, melanoidin content, various measures of antioxidant activity in 40 lager beers.
- Variables/Labels
  1. Beer ID (beer)
  2. Total phenolic content (tpc)
  3. Melanoidin content (ma)
  4. DPPH radical scavenging activity (dsa)
  5. ABTS radical cation scavenging activity (asa)
  6. Oxygen radical absorbence activity (orac)
  7. Reducing Power (rp)
  8. Metal Chelaing Activity (mca)

- Perform one factor analysis by finding the best variable to explain Total Phenolic Content (TPC). Fit one variable regression model with TPC as a dependent variable against remaining variables above, as an independent variable one at time. For example, TPC = $b_o + b_1$ MA, TPC=b_0+b_1 DSA and so on.

Code – Program in R (beer example)

```
prg1<-function(x){
out1<-list({})
out2<-list({})
for (i in 1:6){
out1[[i]]<-lm(x[,2]~x[,2+i],data=x)
out2[[i]]<-cbind(dimnames(x)[[2]][2+i],sum(out[[i]]$residuals^2))
  }
list(out1,out2)}
```

```
prg1(lager_antioxidant_reg)
```