CHAPTER

10

Duxbury Thomson Learning

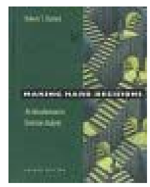
Making Hard Decision

Third Edition

Using Data

A. J. Clark School of Engineering •Department of Civil and Environmental Engineering

FALL 2003



By
Dr . Ibrahim. Assakkaf

ENCE 627 – Decision Analysis for Engineering

Department of Civil and Environmental Engineering

University of Maryland, College Park

Introduction

Using Data

Source for information about probabilities and historical data

Introduction

- So far we have seen two ways to calculate probability for decision models.
  - Subjective probabilities (chapter 8)
  - Theoretical probability models (chapter 9)
- In chapter 9, given a particular probability model, we had just assumed certain parameter values for the any particular distribution.

Introduction

- Example:
  - Poisson model for tornadoes occur in a particular area an average of two times a year. In this case, $\lambda = 2$/year.
  - The parameters u_n and α_n in the example of the Extreme Value Distribution, Type I for maximum wind velocity V_n were assumed as

$$\alpha_n = \sqrt{\frac{\pi^2}{6\sigma_{X_n}^2}} = \sqrt{\frac{\pi^2}{6(7.52)^2}} = 0.17055 \quad \text{and} \quad u_n = \mu_{X_n} - \frac{\gamma}{\alpha_n} = 61.3 - \frac{0.5772}{0.17055} = 57.9157$$

Introduction

- Now we will learn to calibrate models to data finding the best parameters

Using Data to Construct Probability Distributions

- Imagine that you are in charge of a manufacturing plant, and you are trying to develop a maintenance policy for your machines.
- You may collect the following data over 260 days:

No failures	217 days
One failure	32 days
Two failures	11 days

Using Data to Construct Probability Distributions

- These data lead to the following relative frequencies, which could be used as estimates in your analysis: Data collected: out of 260 days (= 52 weeks × 5 days/week):

Basically one year's worth of working days

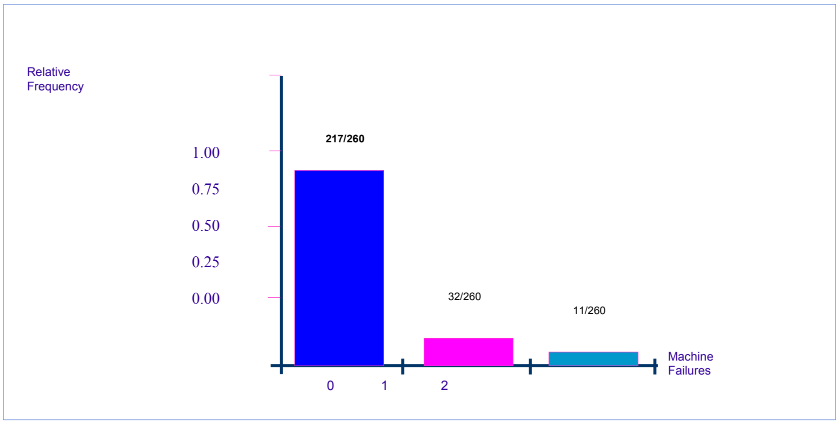
No failures	0.835 = 217/260
One failure	0.123 = 32/260
Two failures	0.042 = 11/260
	1.000

Using Data to Construct Probability Distributions

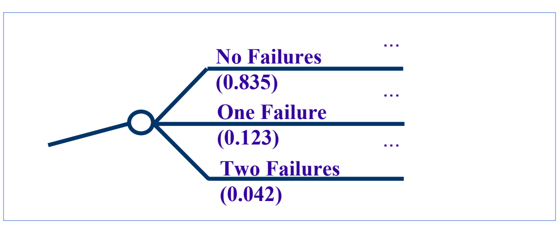
- The only serious consideration to keep in mind is that you should have enough data to make a reliable estimate of the probabilities.

Using Data to Construct Probability Distributions

- Relative frequency histogram for machine failure:



Decision-tree Representation of Uncertainty Regarding Machine Failures



- Note:
  - The data requirements depend on the particular problem, but the minimum should be approximately five observations in the least likely category. The other categories, of course, will have more observations.

Decision-tree Representation of Uncertainty Regarding Machine Failures

- Keep in mind that your probability estimates are just that - estimates.
- Ask yourself whether the probabilities estimated on the basis of the data truly reflect the uncertainty that you face.
- If you are not satisfied with the representation based on the data, you may need to model your uncertainty using subjective assessment methods. In particular, this might be the case if you think that the past may not indicate what the future holds.

Using Data to Construct Probability Distributions: Empirical CDFs

- Notes:
  - You should have enough data to make the %'s reliable.
  - Also, you should ask the question whether the data was collected over a period in which it was representative?

In some settings one year may not be enough, in other settings it may be ok.

Using Data to Construct Probability Distributions: Empirical CDFs

- Slightly more sophisticated approach
  - Empirical CDFs
    - Halfway House Example:
      - Ease transition from prison life to normal civilian life.
      - Increase chances to re-integrate into society

Using Data to Construct Probability Distributions: Empirical CDFs

- Example: Halfway House
  - Yearly per-bed rental costs (in $), let *C* be the random variable representing costs; 35 halfway houses randomly sampled.

C =										
	1	**52**	8	**205**	15	**303**	22	**400**	29	**643**
	2	**76**	9	**250**	16	**313**	23	**402**	30	**693**
	3	**100**	10	**257**	17	**317**	24	**408**	31	**732**
	4	**136**	11	**264**	18	**325**	25	**417**	32	**749**
	5	**137**	12	**280**	19	**345**	26	**422**	33	**750**
	6	**186**	13	**282**	20	**373**	27	**472**	34	**791**
	7	**196**	14	**283**	21	**384**	28	**480**	35	**891**

18 values $\leqslant$ 325
18 values $\geqslant$ 325

Relative Cumulative Frequency

$P(C \leqslant 325) = P(C \leqslant 326) = \ldots = P(C \leqslant 344.99) = 18/35 = 0.514$

Using Data to Construct Probability Distributions: Empirical CDFs

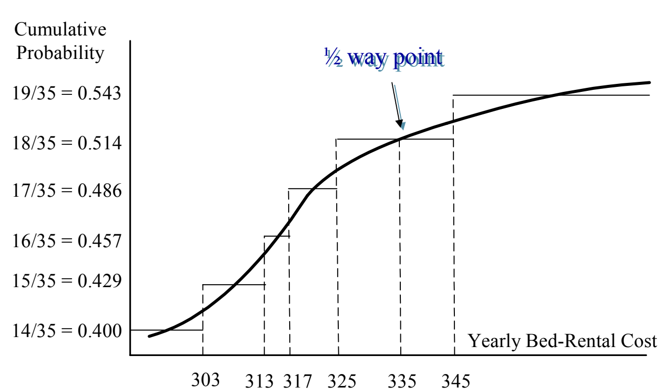
- Example (cont'd): Halfway House

  Question: Which value to be used for the 0.514 fractile?

  Answer: One way is to take a compromise: (325+345)/2 = 335, do the same things for all points.

Using Data to Construct Probability Distributions: Empirical CDFs

- Example (cont'd): Halfway House



Using Data to Construct Probability Distributions: Empirical CDFs

- Example (cont'd): Halfway House

Estimated cumulative probabilities for the halfway-house data

Obs. No.	Cost	x_m	Cumulative Probability	Obs. No.	Cost	x_m	Cumulative Probability
1	52	64.0	0.029	19	345	369.0	0.543
2	76	88.0	0.057	20	373	378.5	0.571
3	100	118.0	0.086	21	384	392.0	0.600
4	136	136.5	0.114	22	400	401.0	0.629
5	137	161.5	0.143	23	402	405.0	0.657
6	186	191.0	0.171	24	408	412.5	0.686
7	196	200.5	0.200	25	417	419.5	0.714
8	205	227.5	0.229	26	422	447.0	0.743
9	250	253.5	0.257	27	472	476.0	0.771
10	257	260.5	0.286	28	480	561.5	0.800
11	264	272.0	0.314	29	643	668.0	0.829
12	280	281.0	0.343	30	693	712.5	0.857
13	282	282.5	0.371	31	732	740.5	0.886
14	283	293.0	0.400	32	749	749.5	0.914
15	303	308.0	0.429	33	750	770.5	0.943
16	313	315.0	0.437	34	791	841.0	0.971
17	317	321.0	0.486	35	891		
18	325	335.0	0.514				

Using Data to Construct Probability Distributions: Empirical CDFs

- Example (cont'd): Halfway House

½ way point
(52+76)/2=64
etc.

n: total points
(n=35 here)
m: typical point

(303+313)/2=308

Obs. No.	Cost	x_m	Cumulative Probability	Obs. No.	Cost	x_m	Cumulative Probability
1	52	64.0	0.029	19	345	369.0	0.543
2	76	88.0	0.057	20	373	378.5	0.571
3	100	118.0	0.086	21	384	392.0	0.600
4	136	136.5	0.114	22	400	401.0	0.629
5	137	161.5	0.143	23	402	405.0	0.657
6	186	191.0	0.171	24	408	412.5	0.686
7	196	200.5	0.200	25	417	419.5	0.714
8	205	227.5	0.229	26	422	447.0	0.743
9	250	253.5	0.257	27	472	476.0	0.771
10	257	260.5	0.286	28	480	561.5	0.800
11	264	272.0	0.314	29	643	668.0	0.829
12	280	281.0	0.343	30	693	712.5	0.857
13	282	282.5	0.371	31	732	740.5	0.886
14	283	293.0	0.400	32	749	749.5	0.914
15	303	308.0	0.429	33	750	770.5	0.943
16	313	315.0	0.437	34	791	841.0	0.971
17	317	321.0	0.486	35	891		
18	325	335.0	0.514				

Using Data to Construct Probability Distributions: Empirical CDFs

- Example (cont'd): Halfway House

The value for x_m is the m^{th} one.

∴ it should have cumulative probability of m/n.

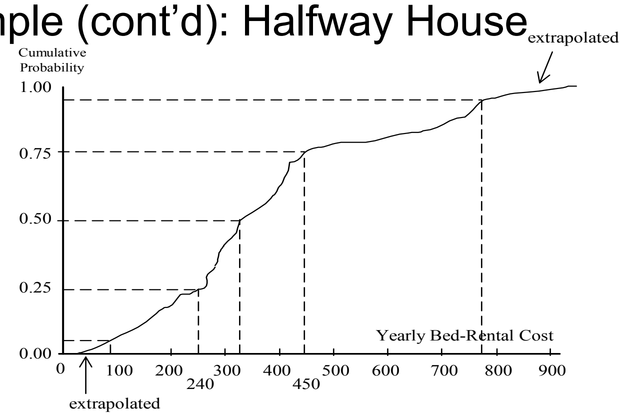
Example: P(C ⩽ 64) =1/35 ≅ 0.029
P(C ⩽ 308) =15/35 ≅ 0.429
P(C ⩽ 335) =18/35 ≅ 0.514
Do for all 35 points and then smoothly extrapolate for the tails

Using Data to Construct Probability Distributions: Empirical CDFs

- Example (cont'd): Halfway House

can automate easily

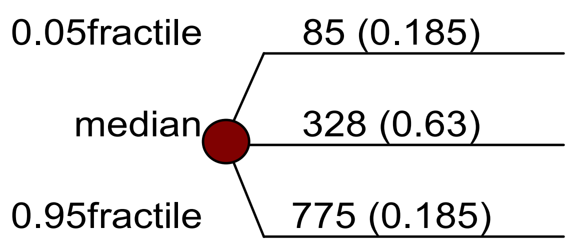
(e.g. RiskView)



We can say that there is:
- **50% chance that the yearly bed-rental cost will fall between $240 and $450.**
- **25% chance that the cost would fall below $240.**
- **25% chance that it would fall above $450.**

Using Data to Construct Probability Distributions: Empirical CDFs

- Example (cont'd): Halfway House
  - Alternatively, could use a discrete approximation e.g. three-point Pearson-Tukey method



Using Data to Fit Theoretical Probability Models

Method A: One way to deal with data is simply to fit a theoretical distribution to it.

Step 1: Decide what kind of distribution is appropriate (binomial, Poisson, normal, and so on)

- What distribution is best? Need to understand the setting
  - Defects maybe Poisson
  - Value in [0,1] maybe beta
  - Normal? Need symmetry as well as other things

Using Data to Fit Theoretical Probability Models

Step 2: Choose the values of the distribution parameters

- Having chosen the distribution, need to calibrate, i.e., choose the values for the parameters. Bernoulli (*P*), Binomial (*n*, *p*), Poisson (λ), etc.
- Easy way (probably adequate in a number of settings.
- Take sample mean and sample variance:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Statistical reasons why not *n*

Using Data to Fit Theoretical Probability Models

- Example: Calculate the sample mean (*x*) and sample variance (S_2) for the 35 halfway house observations

$$n = 35$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 380.4$$

$$S^2 = \frac{1}{n-1}\sum_{i=}^{n}\left(x_i - \overline{X}\right)^2 = 47{,}344.3$$

$$S = \sqrt{47{,}344.3} = 217.6$$

We might choose a normal distribution

with mean μ = 380.4 and standard deviation

σ = 217.6 to represent the distribution of the yearly bed-rental costs.

Using Data to Fit Theoretical Probability Models

Method B: Fit a theoretical distribution using fractiles. That is, find a theoretical distribution whose fractiles match as well as possible with the fractiles of the empirical data. In this case we would be fitting a theoretical distribution to a data-base distribution.

Using Data to Fit Theoretical Probability Models

Method C: For most initial attempts to model uncertainty in a decision analysis, it may be adequate to use the sample mean and variance as estimates of the mean and variance of the theoretical distribution and to establish parameter values in this way. Refinement of the probability model may require more careful judgment about the kind of distribution as well as more care in fitting the parameters.

Testing the Validity of Assumed Distribution

- When a theoretical distribution has been assumed, the validity of the assumed distribution may be verified or disproved statistically by goodness-of-fit tests.
- Two tests are commonly used:
  - The Chi-square
  - The Kolmogorov-Smirnov test

Testing the Validity of Assumed Distribution

- Chi-square Test for Goodness of Fit
  - Consider a sample of *O* observed values of a random variable.
  - The chi-square goodness-of-fit test compares the observed frequencies O_1, *O*2,…, O_k of *k* values (*k* intervals) of the variate with the corresponding frequencies E_1, E_2,…,E_k from an assumed or theoretical distribution.

Testing the Validity of Assumed Distribution

- Chi-square Test for Goodness of Fit
  – The basis for the appraising the goodness of the comparison is given by the following test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where χ^2 is the computed value of a random variable having a chi-square distribution with $k - 1$ degrees of freedom; O_i and E_i are the observed and expected frequencies in cell (or interval) i, and k is the number of discrete cells (intervals) into which data were separated.

Testing the Validity of Assumed Distribution

- Chi-square Test for Goodness of Fit
  – Degrees of Freedom
    - If the mean and standard deviation of the sample are needed to compute the expected frequencies, then two additional degrees of freedom are subtracted (i.e., $k - 3$).
    - If the mean and standard deviation are obtained from past experience or other sources, then the number of degrees of freedom for the test statistic remains $k - 1$.

Testing the Validity of Assumed Distribution

- Chi-square Test for Goodness of Fit
  - If the assumed distribution yields

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} < \chi^2_{\alpha,\nu}$$

1. The assumed theoretical distribution is an acceptable model if $\chi^2 < \chi^2_{\alpha,\nu}$
2. Otherwise, it is not acceptable at the α significance level.

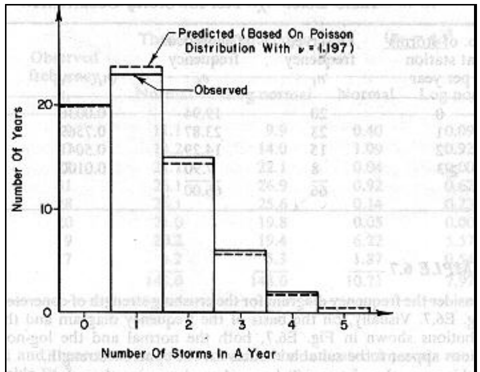
Testing the Validity of Assumed Distribution

- Example: Rainstorms

Severe rainstorms have been recorded at a given station over a period of 66 years. During this period, there were 20 years without severe rainstorms; and 23, 15, 6, and 2 years, respectively, with 1, 2, 3, and 4 rainstorms annually. Judging from the shape of the histogram, a Poisson distribution seems an appropriate model for the annual number of rainstorms. Is this claim valid? Use a significance level of 5%.

Testing the Validity of Assumed Distribution

- Example (cont'd): Rainstorms



Histogram and Poisson model for storm occurrences

Testing the Validity of Assumed Distribution

- Example (cont'd): Rainstorms

No. of storms at station per year	Observed frequency, O_i	Theoretical frequency, E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	20	19.94	0.0036	0.0002
1	23	23.87	0.7569	0.0317
2	15	14.29	0.5041	0.0353
>3	8	7.90	0.0100	0.0013
∑	99	66.00		0.0685

Testing the Validity of Assumed Distribution

- Example (cont'd): Rainstorms

$\alpha = 0.05 \quad \Rightarrow 1 - \alpha = 1 - 0.05 = 0.95$

$$\lambda = \frac{\overline{X}}{t} = \frac{23 + 2\times 15 + 3\times 6 + 4\times 2}{66} = \frac{79}{66} = 1.197 \text{ rainstorms/year}$$

From Chi-squares Table, for $\alpha = 0.05$, and $\nu = k - 2 = 4 - 2 = 2$,

$$\chi^2_{0.05,2} = 5.995$$

Since,

$$\left(\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.068\right) < \left(\chi^2_{\alpha,\nu} = 5.995\right)$$

Hence, the Poisson distribution is a valid model at the 5% significance level.

Software for Fitting Distributions: BestFit and @RISK

- Excellent software exists to help a decision analysis fit theoretical distributions to data.
- BestFit, a program published by Palisade Software. Providing a way for an analyst to fit theoretical distributions to subjective assessments elicited from experts.

Software for Fitting Distributions: BestFit and @RISK

- Fitting distributions to the half-way house data with BestFit.
- Read procedure of BestFit with @RISK packages in book pages 405 – 411.

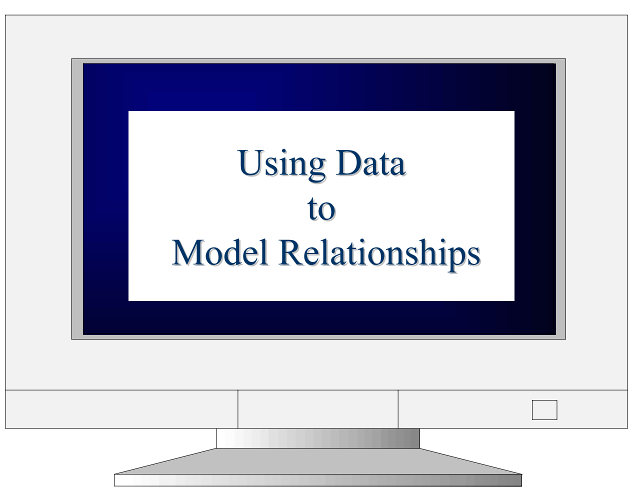
Summary

- We have seen some ways in which data can be used in the development of probabilities and probability distributions for decision analysis.
- The basis of constructing histograms and empirically based CDFs.
- Use of data to estimate parameters for theoretical distributions.

References

1. Berry, D. (1995). Statistics: A Bayesian Perspective, Belmont, CA: Duxbury.
2. Chatterjee, S., and B. Price (1977) Regression Analysis by Example. New York: Wiley.
3. DeGroot, M. (1970) Optimal Statistical Decisions. New York: McGraw-Hill.
4. Draper, N., and H. Smith (1981) Applied Regression Analysis, 2nd Ed. New York: Wiley.
5. Neter, T., W. Wasserman, and M. Kutner (1989). Applied Linear Regression Models, 2nd Ed. Homewood., IL: Irwin.
6. Olkin,I., L.J. Gleser, and C. Derman (1980) Probability Models and Applications. New York: Macmillan.
7. Raiffa, H., and R. Schlaifer (1961) Applied Statistical Decision Theory. Cambridge, MA: Harvard University

Using Data to Model Relationships



Using Data to Model Relationships

- We use data to try to understand the relationships that exist among different phenomena in our world.
- Examples:
  - Causes of Cancer
  - Sales Revenue
  - Economic Conditions
  - Natural Processes

Using Data to Model Relationships

Note:

1. The motivation for studying relationships among phenomena that we observe is to gain some degree of control over our world. In many cases we hope to make changes in those areas where we have direct control in order to accomplish a change in another area.
2. We will focus on the problem of using data on a number of auxiliary variables (which we will denote as $X_1, \ldots, X_k$) to determine the distribution of some other variable of interest (Y) that is related to the x's.

Using Data to Model Relationships

3. *Y* is sometimes called a ***response*** variable or **dependent** variable, because its probability distribution changes in response to changes in the *X*'s.
4. The *X*'s sometimes are called ***explanatory*** variables or ***independent*** variables, because they can be used to help explain the changes in *Y*.

Using Data to Model Relationships

– The use of data to understand relationships is not trivial. Consider the influence diagrams.

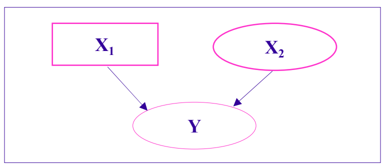
- The brute force approach would require obtaining enough data to estimate the conditional distribution for the particular variable of interest (Y) *for every possible combination of values for its conditioning or predecessor variables* (X_1 and X_2).
- We would need to know what are feasible values for the decision variable (X_1), and we would have to assess a distribution for the possible values for the uncertain variable (X_2).

Note: We would require a lot of data, and even in simple problems this could be a tedious or infeasible task.

Using Data to Model Relationships

Example:
X_2 → Pearson-Tukey three-point approximation
X_1 → low, medium, high

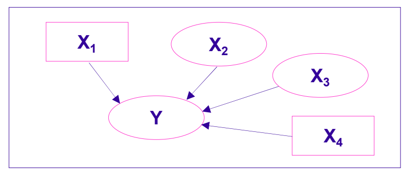
Nine different conditional probability distributions for Y based on the possible scenarios.



An influence diagram for modeling relationship among uncertain quantities X_1, X_2, and Y.

Example:
X_3 and X_4 → Three possible values
X_1 and X_2 → Three point approximation

We would need to come up with 81 conditional distributions $3^4 = 81$



An influence diagram relating two uncertain quantities and two decision variables to Y.

The Regression Approach

- One way to model the relationships between variables
  - Determine the conditional expected value of Y given the X's, $E(Y \mid X_1, \ldots, X_k)$.
  - Consider the conditional probability distribution around that expected value.

The Regression Approach

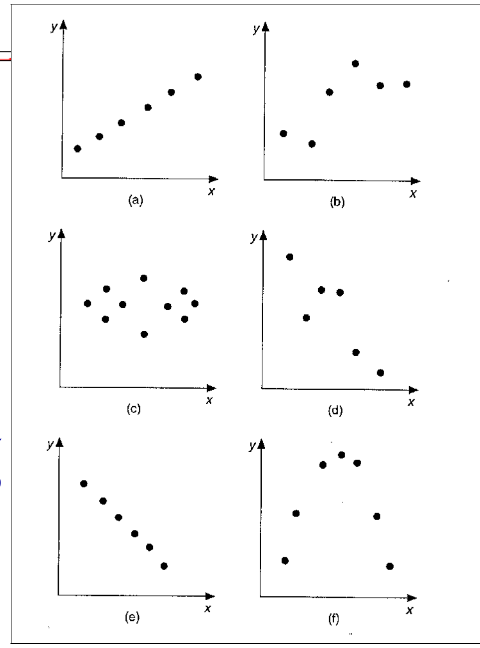
- Correlation
  - The study of the degree of linear interrelation between random variables is called correlation analysis.
  - Correlation analysis provides a means of drawing inferences about the strength of the relationship between two or more variables.

The Regression Approach

- Correlation
  - Correlation is a measure of the degree to which the values of these variables vary in a systematic manner.
  - It provides a quantitative index of the degree to which one or more variables can be used to predict the values of another variable

Correlation

- Different degrees of correlation between variables *X* and *Y* .
  - High degree of correlation in Fig. a and e.
  - No correlation in Fig. c.
  - The degree of correlation is moderate in Fig's b and d.
  - In Fig. b, exact change in *Y* for change in *X* is difficult to predict.
  - In Fig. f, very predictable trend, but poor correlation.



The Regression Approach

- Limitations of Correlation Analysis
  - Correlation analysis does not provide an equation for predicting the value of a variable..
  - Also, it does not indicate whether a relationship is causal, that is whether there is a cause-and-effect relationship between the variables.

The Regression Approach

- Correlation
  - Example random variables having causal relationship and strong correlation:
    - The cost of living and wages
    - The volumes of rainfall and flood runoff
  - Example random variables not having causal relationship and strong correlation:
    - The crime rate and the sale of chewing gum last decade
    - Annual population growth in 19th century France and annual cancer deaths rate in the U.S. in the 20th century

The Regression Approach

- Correlation

  Separation of Variation

$$\text{TV} = \text{EV} + \text{UV}$$

TV = total variation

EV = explained variation

UV = unexplained variation

The Regression Approach

- Correlation
  - Separation of Variation: A Set of Observations on a Random Variable *Y*

$$\mathrm{TV} = \mathrm{EV} + \mathrm{UV} \qquad \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

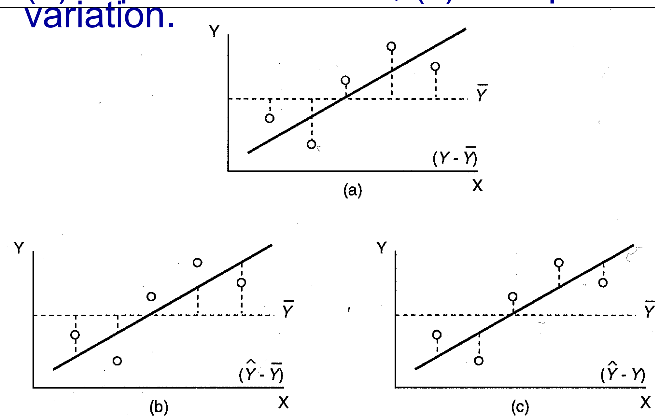
$$1 = \frac{\mathrm{EV}}{\mathrm{TV}} + \frac{\mathrm{UV}}{\mathrm{TV}}$$

$$R = \frac{\mathrm{EV}}{\mathrm{TV}} \qquad R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}}$$

The Regression Approach

- Correlation

  Separation of variation: (a) total variation; (b) explained variation; (c) unexplained variation.



The Regression Approach

- Need for Regression
  - When dealing with two or more variables, the functional relationship between the variables is often of interest.
  - However, if one or two variables (in two-variable case) are random, there is no unique relationship between the values of the two variables.

The Regression Approach

- Need for Regression
  - Given a value of one variable (the controlled or independent variable), there is a range of possible values of the other.
  - Thus, a probabilistic description is required.

The Regression Approach

- Regression Analysis

  Regression analysis is the probabilistic relationship between random variables when this relationship is described in terms of the mean and variance of one random variable as a function of the value of the other random variables.

The Regression Approach

- Optimization
  - The process of deriving a relationship between a random variable and measured values of other variables is called "*optimization*" or model "*calibration*"
  - The objective of optimization is to find the values of vectors of unknowns that provides the minimum or maximum value of some function.

The Regression Approach

- Correlation Versus Regression
  - *Correlation analysis* provides a measure of goodness of fit.
  - *Regression analysis* is a means of calibrating the unknown coefficients of a prediction equation.
  - *Correlation* has its usefulness in model formulation and verification.

The Regression Approach

- Elements of Statistical Optimization
  1. An objective function, which defines what is meant by the best fit.
  2. A mathematical model, which is a n explicit function relating a criterion variable (i.e., *Y*) to vectors of unknowns and predictor (i.e., *X*) variable(s)
  3. A matrix of measured data

The Regression Approach

- Example: Evaporation
  - In irrigation projects, it is necessary to provide estimates of evaporation.
  - Evaporation can be a function of other variables such as air temperature, humidity, and air mass.
  - If measurement of air temperature are available, a relationship or a *model* can be developed.

The Regression Approach

- Example (cont'd): Evaporation

$$\hat{E} = \beta_0 + \beta_1 T$$

b_0 and b_1 = the unknown coefficients

$\hat{E}$ = the predicted value of E

T = air temperature

The Regression Approach

- Example (cont'd): Evaporation
  - If we are interested in daily evaporation rates, we may measure both the total evaporation for each day in a year and the corresponding mean daily temperature.
  - An objective function should be established to evaluate the unknowns.
  - Regression minimizes the sum of the squares of the differences between the predicted and measured values.

The Regression Approach

- Regression Definitions
  - The objective of regression is to evaluate the coefficients of an equation relating the criterion variable to one or more variables, which are called the *predictor variables*.
  - The predictor variables are variables in which their variation is deemed to cause or agree with variation in criterion variable

The Regression Approach

- Linear Regression
  - The conditional expected value of Y is linear in the X's
  - In symbols:

$$E(Y|X_1,\ldots,X_k) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

The β's are coefficients, and they serve the purpose of combining the X values to obtain a conditional expected value for Y.

The Regression Approach

- Note:
  1. It is important to remember that the equation defines a relationship between the explanatory variables and the expected Y. The actual Y value will be above or below this expected value to some extend; this is where the uncertainty and the conditional probability distribution of Y come into play.

The Regression Approach

2. The distribution around the conditional expected value has the same shape regardless of the particular X values

$$\hat{Y} = \mathrm{E}(Y \mid X_1, \ldots, X_k) + \varepsilon$$

The conditional distribution (and the corresponding density) of Y, given the X's, ahs the same shape as the distribution (or density) of the errors, but it is just shifted so that the distribution is centered on the expected value E(Y|X_1, . . . , X_k)

The Regression Approach

- Bivariate Model

$$\hat{Y} = \beta_0 + \beta_1 X$$

- Multivariate model

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_k X_k$$

The Regression Approach

- Principle of Least Squares
  – The principle of least squares is the process of obtaining the *best* estimates of the coefficients (β_0, β_1,.., β_k).
  – This principle is referred to as the regression method.
  – To express the principle of least squares, it is important to define the error e

The Regression Approach

- Principle of Least Squares
  – Objective Function

$$F = \min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)$$

$$\hat{y}_i = \text{the ith predicted value of } \hat{Y}$$

$$y_i = \text{the ith measured value of } Y$$

$$e_i = \text{the ith error}$$

The objective function for the principle of least squares is to minimize the sum of the squares of the errors

The Regression Approach

- Procedure Solution for the Bivariate model

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

The objective function in this case is

$$F = \min \sum_{i=1}^{n} \left(\hat{Y}_i - y_i\right)^2 = \min \sum_{i=1}^{n} \left(b_0 + b_1 x_i - y_i\right)^2$$

The derivatives of the sum of the squares of the errors with respect to the unknowns b_0 and b_1 are as follows:

The Regression Approach

$$\frac{\partial F}{\partial b_0} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_i - y_i\right) = 0$$

$$\frac{\partial F}{\partial b_1} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_i - y_i\right) x_i = 0$$

Dividing each equation by 2, separating the terms in the summation, and rearranging yields the set of normal equations:

$$\sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1 x_i - \sum_{i=1}^{n} y_i = 0 = n b_0 + b_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} b_0 x_i + \sum_{i=1}^{n} b_1 x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i$$

The Regression Approach

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

From which,

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{2}\left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{and} \quad b_0 = \bar{Y} - b_1 \bar{X} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b_1}{n}\sum_{i=1}^{n} x_i$$

The Regression Approach

- Bivariate Model

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

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b_1}{n}\sum_{i=1}^{n} x_i$$

The Regression Approach

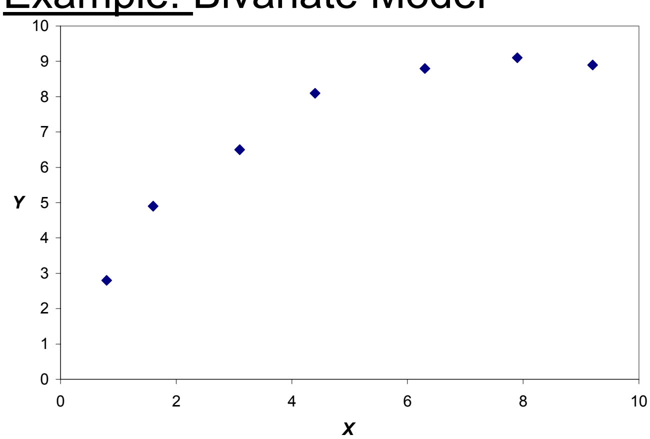
- Example: Bivariate Model

Given the following pairs of observations, compute the regression coefficients for the bivariate (linear) model using the principle of least squares.

X	0.8	1.6	3.1	4.4	6.3	7.9	9.2
Y	2.8	4.9	6.5	8.1	8.8	9.1	8.9

The Regression Approach

- Example: Bivariate Model



The Regression Approach

- Example: Bivariate Model

	x_i	y_i	x_i^2	$x_i y_i$
	0.8	2.8	0.64	2.24
	1.6	4.9	2.56	7.84
	3.1	6.5	9.61	20.15
	4.4	8.1	19.36	35.64
	6.3	8.8	39.69	55.44
	7.9	9.1	62.41	71.89
	9.2	8.9	84.64	81.88
Σ	33.3	49.1	218.91	275.08

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b_1}{n}\sum_{i=1}^{n} x_i$$
$$= \frac{1}{7}(49.1) - \frac{0.68605}{7}(33.3)$$
$$= 3.75063$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{275.08 - \frac{1}{7}(33.3)(49.1)}{(218.91) - \frac{(33.3)^2}{7}} = 0.68605$$

The Regression Approach

- Example: Bivariate Model

$$\hat{Y} = 3.751 + 0.686X$$