STEVENS INSTITUTE of TECHNOLOGY
School of Business

Amir H Gandomi; PhD
Assistant Professor
Stevens Institute of Technology
a.h.gandomi@stevens.edu

Multivariate Data Analysis – BIA 652

Class 3 – Introduction to Multivariate Analysis & Simple Regression

Homework – Class 3-Part 2

- Homework for next meeting:
  - Finish reading Chapters 1 – 5
  - Start reading Chapter 6
  - Assign the first HW Due Next Class
  - Submitting your dataset (Group assignment)
- Grading Assistant:
  - Mr. Haochen Liu (hliu56@stevens.edu)
- Office Hours:
  - Tuesdays from 2-6 PM
  - Or By Appointment

Course Topics

Topics	Text Reference
Multivariate Overview/Univariate Review	Slides
Univariate Review/Introduction	Slides/Chapters 1 – 5, +
Regression & Correlation	Chapter 6
Multiple Regression & Correlation	Chapter 7, +
Classification: Discriminant Analysis	Chapter 11
Classification: Logistic Regression, Naïve Bayes.	Chapter 12
Dimension Reduction: PCA, SVD, & Factor Analysis;	Chapter 14, 15
Cluster Analysis	Chapter 16
Additional Topics: E. g., SVM, ANOVA, MANOVA, Multi Dimensional Scaling (as time permits)	Outside References

Multivariate analysis

- Concerned with datasets that have more than one response variables for each observational unit
- N rows (cases) and P columns (variables)
  - Relationships among cases
  - Relationships among variables
- First, visualize
  - Pairs plot – plot scatter plot matrix
  - pairs plots can easily miss interesting structure
  - multivariate methods explore the data in a less coordinate-dependent way

Multivariate Dataset

Independent Variable

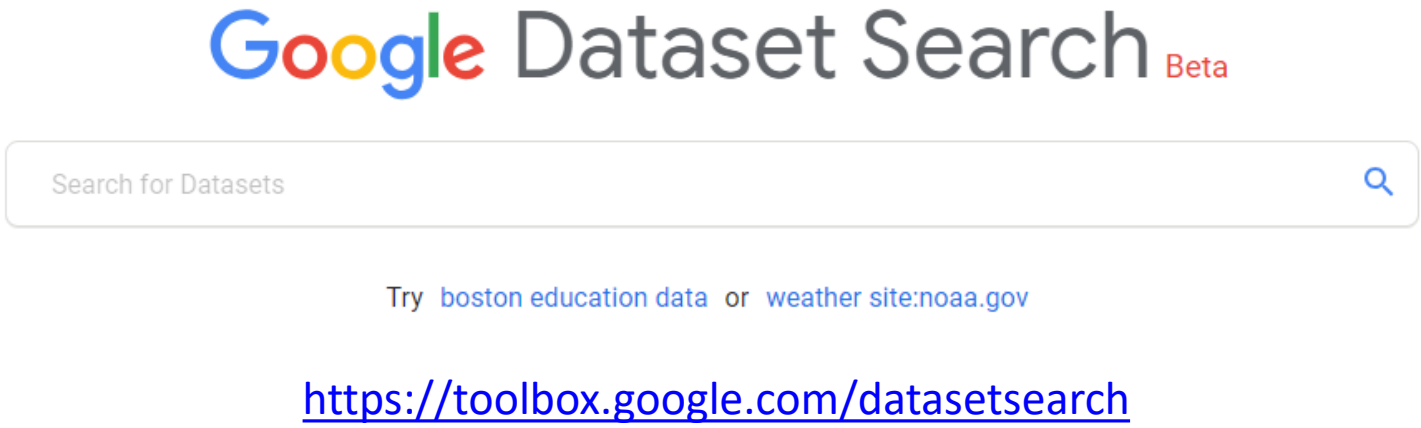
Dependent Variable

Individual	Age	Gender	Height	Weight	Health Code

Season	Time of Day	Dry Roadway Surface	Dark Not Lit	Clear Weather	DUI Driver	Seat Belt Used	Injury Severity
Winter	Night	No	No	No	N/A	Yes	Severe Injury
Summer	Evening	Yes	Yes	Yes	Yes	Yes	Minor Injury
Winter	Afternoon	Yes	No	Yes	No	No	No Injury

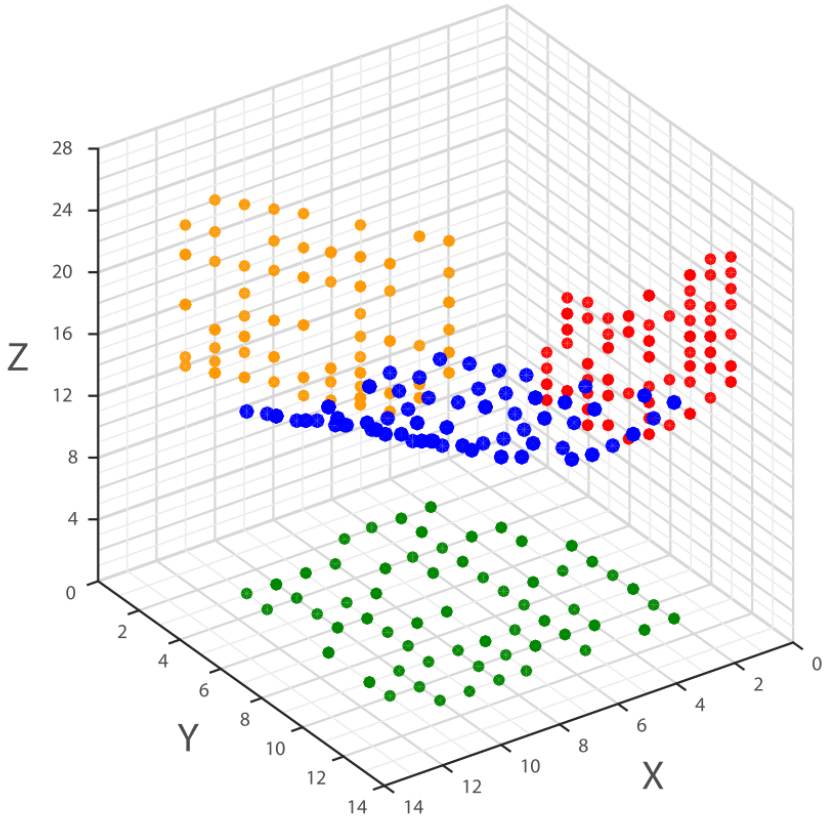
Finding a Dataset for your Project

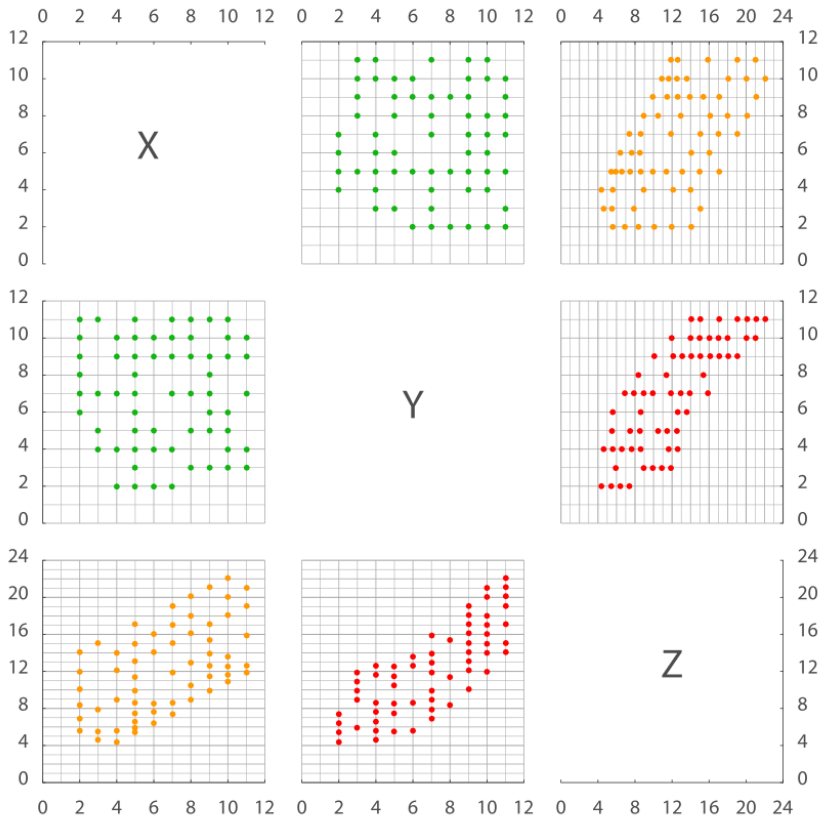
- You need a Multivariate dataset for Classification (two classes) or regression
  - Your own dataset (best)
  - Find it through the Internet (dataset journals, google, etc.)



https://toolbox.google.com/datasetsearch

Scatter plot



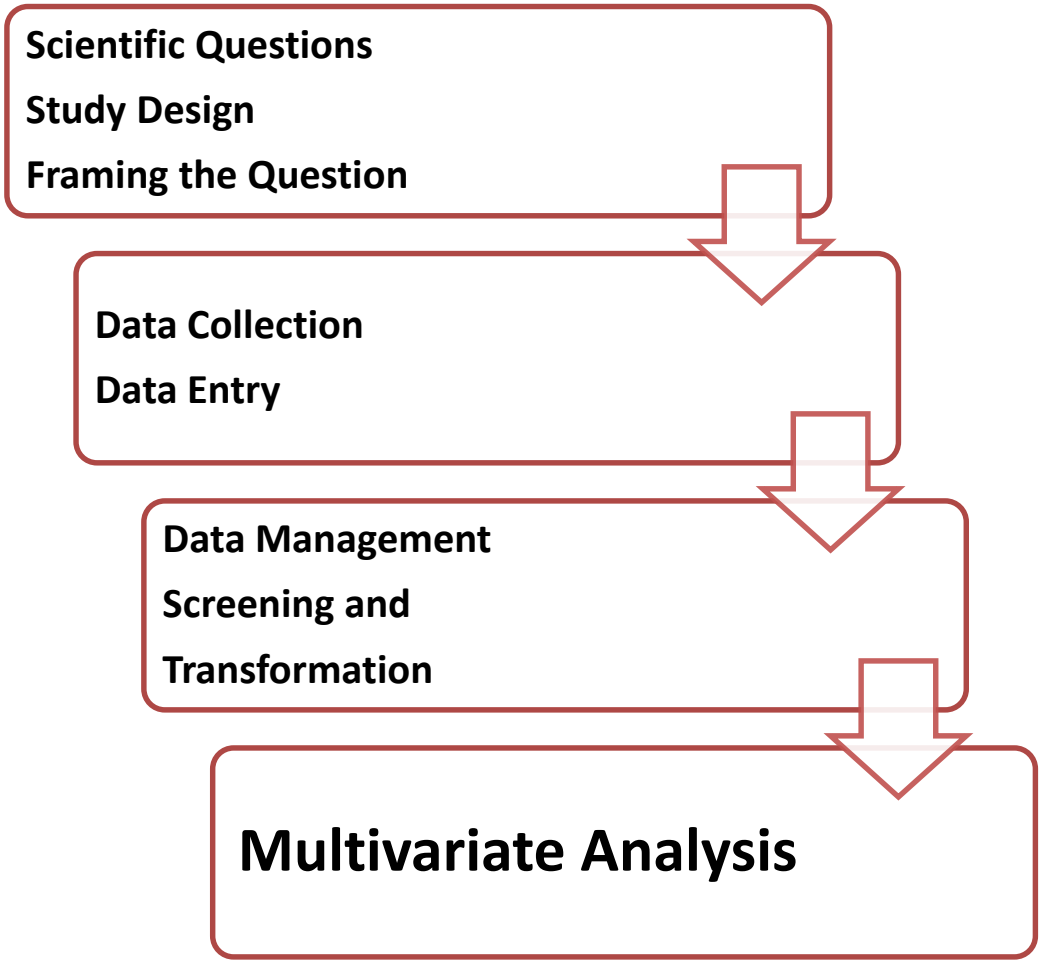


https://en.wikipedia.org

Example of Multivariate Data Depression Data from Afifi (See Page 42, Table 3.3)

Preparation for Analysis

Science & Statistics



- Structured Data
- Semi Structured Data
- Event Processing
- Unstructured Data
- Sensor Networks
- Networks & Protocols
- Real time
- Data Streams
- Multimodality

- Missing Values
- Missing Files
- Distribution
- Outliers
- Data Transformation
- Data Types
- Data Integration

- Data validation
- Statistics
- Data Mining
- Machine Learning
- Graph Analysis
- Text Mining
- Visualization

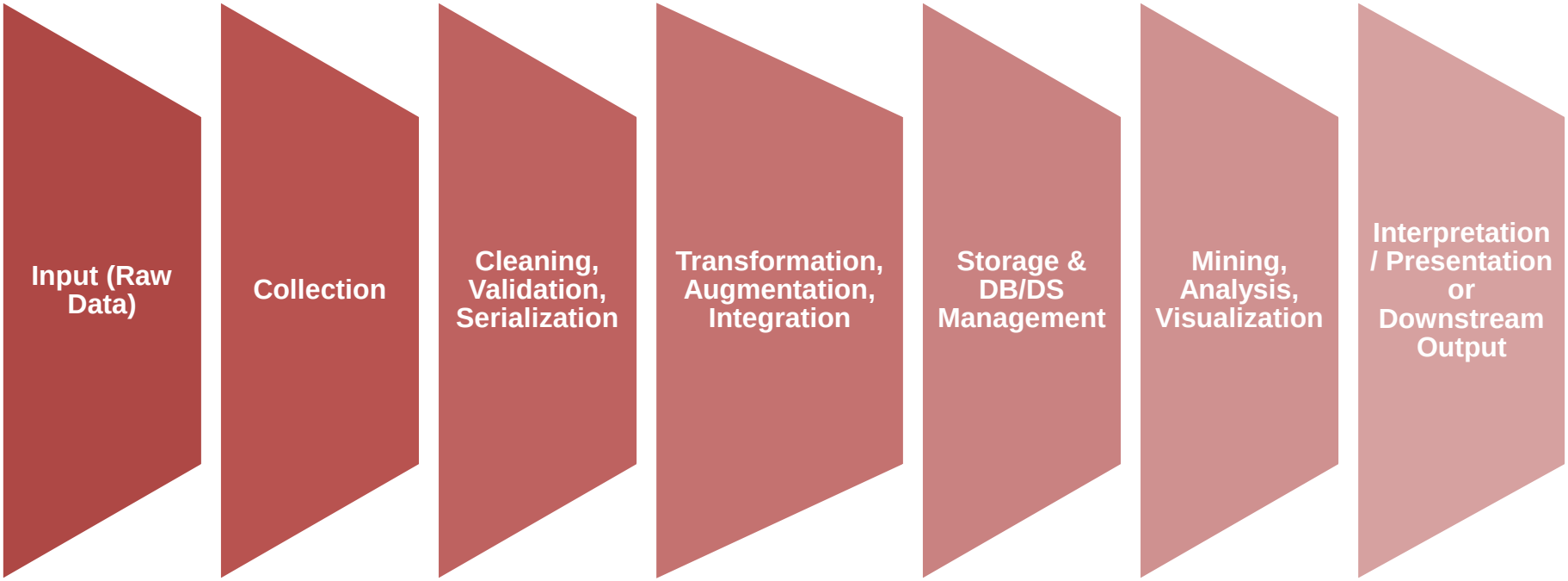
- Trust / Provenance
- Data Quality
- Annotation
- Data Validation
- Automation

- RDBMS
- CSV, JSON,
- Cloud
- Query Languages
- Models & Schema

- Decision Support
- Prediction
- Classification
- Exploration / Confirmation
- Visualization
- Simulation
- Clustering
- Domain Specific

Data Lifecycle I (Basic)

Can we completely automate the staging of data for analysis (From origination to analysis ready)?



IEEE BigData

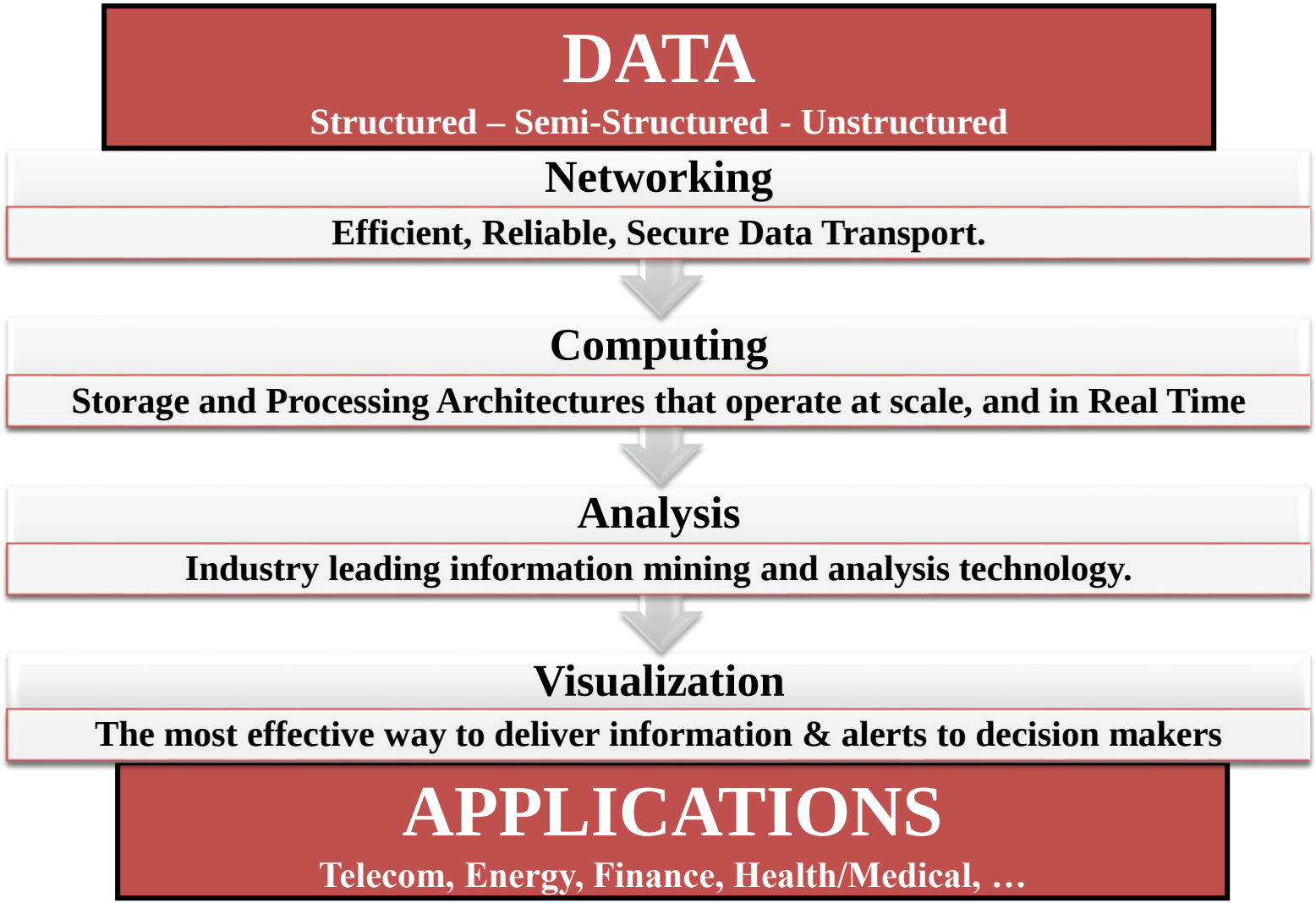
STEVENS INSTITUTE of TECHNOLOGY THE INNOVATION UNIVERSITY

Data System Lifecycle II (Extended)

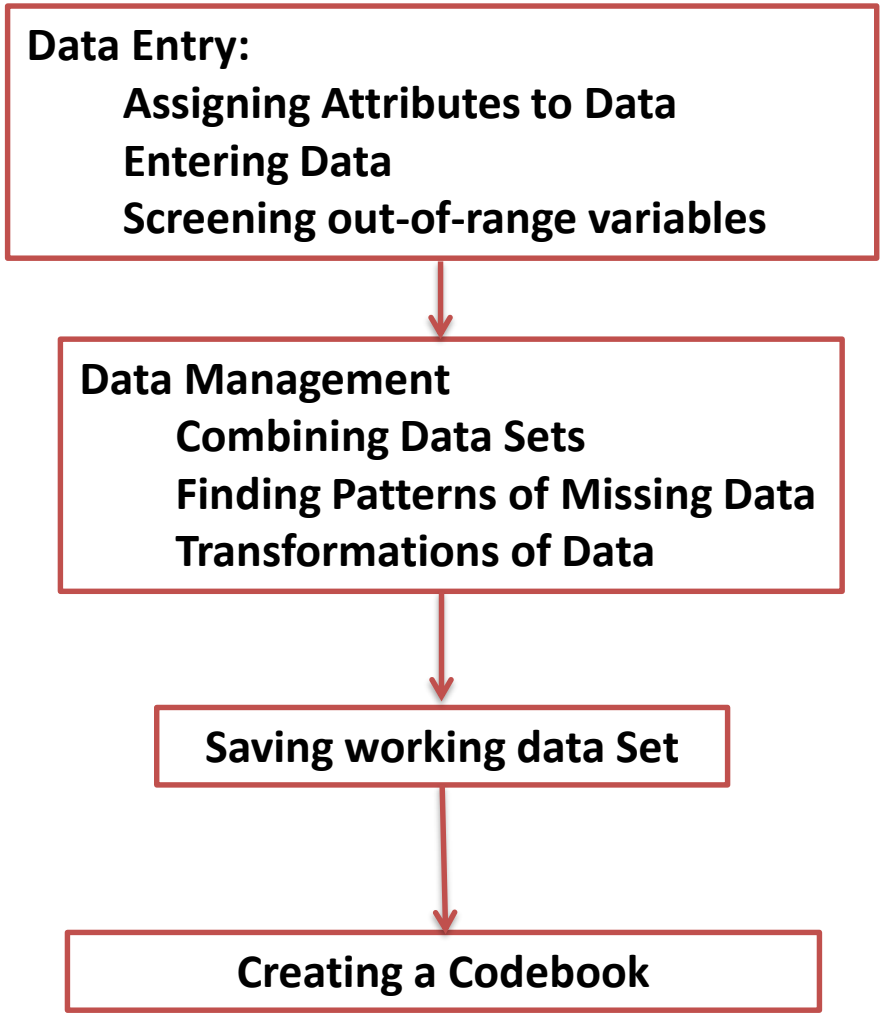
The results of analytic applications must be integrated into a production infrastructure taking into account differences between machine learning systems and classical systems

Data Lifecycle I (Basic)	Non-Functional Requirements	Testing and QA	Deployment	Operations	Maintenance
• Input Data • Collection • Cleaning, Validation, Serialization • Transformation, Augmentation, Integration • Storage & DB/DS Management • Mining, Analysis, Visualization • Interpretation/ Presentation/ Downstream Output	• Performance • APIs • Reliability: MTTF, MTTR • Security, Privacy	• Standard Testing Technology – e.g. Code Review, etc. • Test Data Structure, Version, Drift Control • Metadata and Semantics Documentation • Testing Environments: Unit, Integration, System, Load. • Concurrency, etc.	• Automated Change Control • Automated Data Feed Monitor • Resource and Capacity Management • Incubation: Sandbox to Deployment to Operations	• Upgrade Strategy • Dashboards and Logs • Configuration Management • Feature Set Management	• Version, Configuration, Build • Platform Integration

Industrial Data Analysis at Scale



Preparing Data (p. 40)



Data Screening

- Aims:
  - Identify outliers and inconsistent values
  - Assess normality of the distribution
  - Assess independence of observations
  - Explore data transformation to aid description, inference

Assess Normality

- Approaches:
  - Visual: histogram, cumulative distribution function
  - normal probability plot
  - formal statistical tests
- Caveats:
  - Violations can affect inference
  - Consider transforming data to change distribution

Data Transformation

- Change distribution to better meet assumptions for inference
- Create new variables to simplify later analyses:
  - ➢ to clarify role of extraneous factors
  - ➢ to induce linearity of response
- (Advanced) Consider non-parametric techniques
  - – data is not required to fit a normal distribution

Histograms and their Normal Probability Plots (p. 53 & 54)

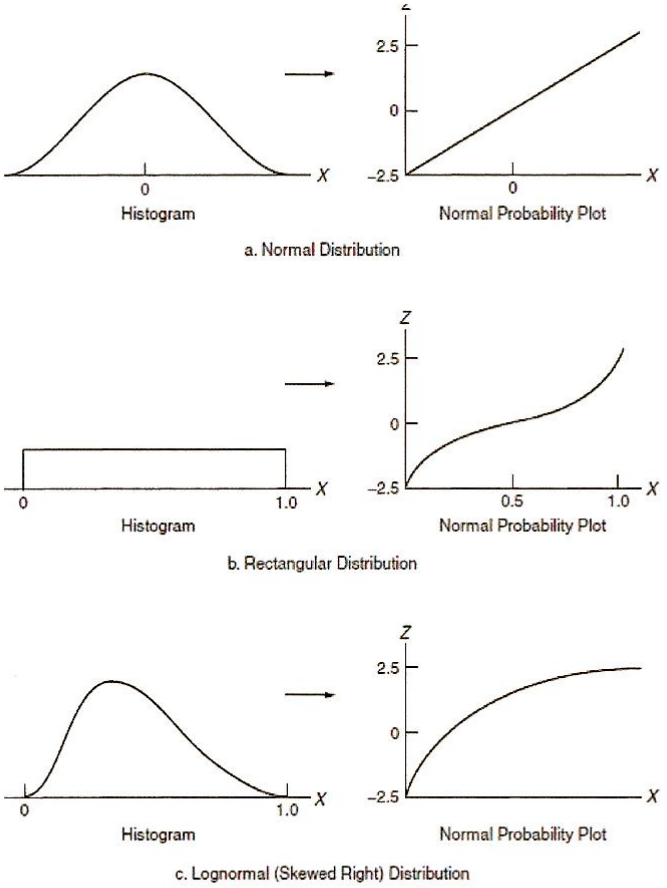


Figure 4.4: *Plots (a,b,c) of Common Histograms and the Resulting Normal Probability Plots from those Distributions*

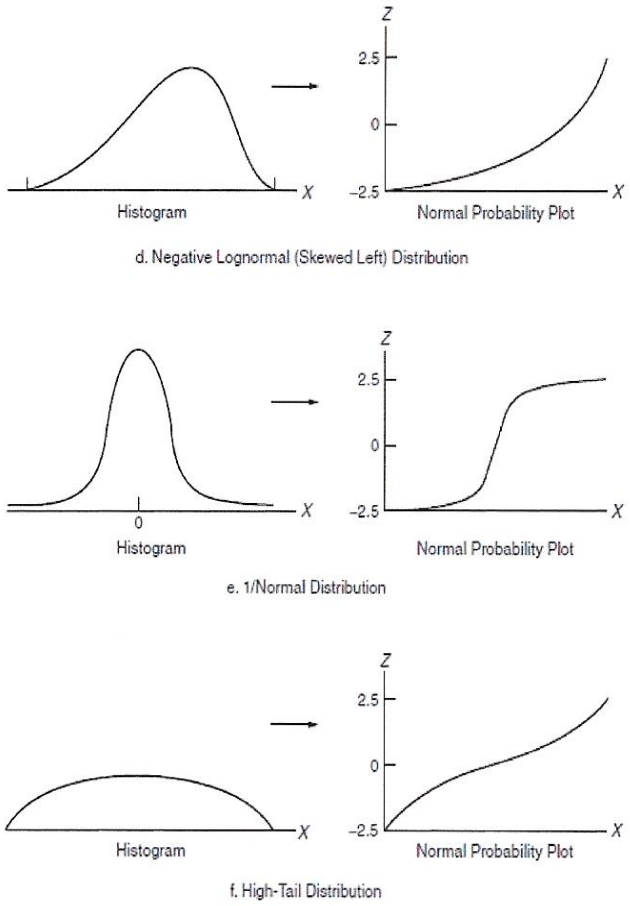


Figure 4.5: *Plots (d,e,f) of Other Histograms and the Resulting Normal Probability Plots from those Distributions*

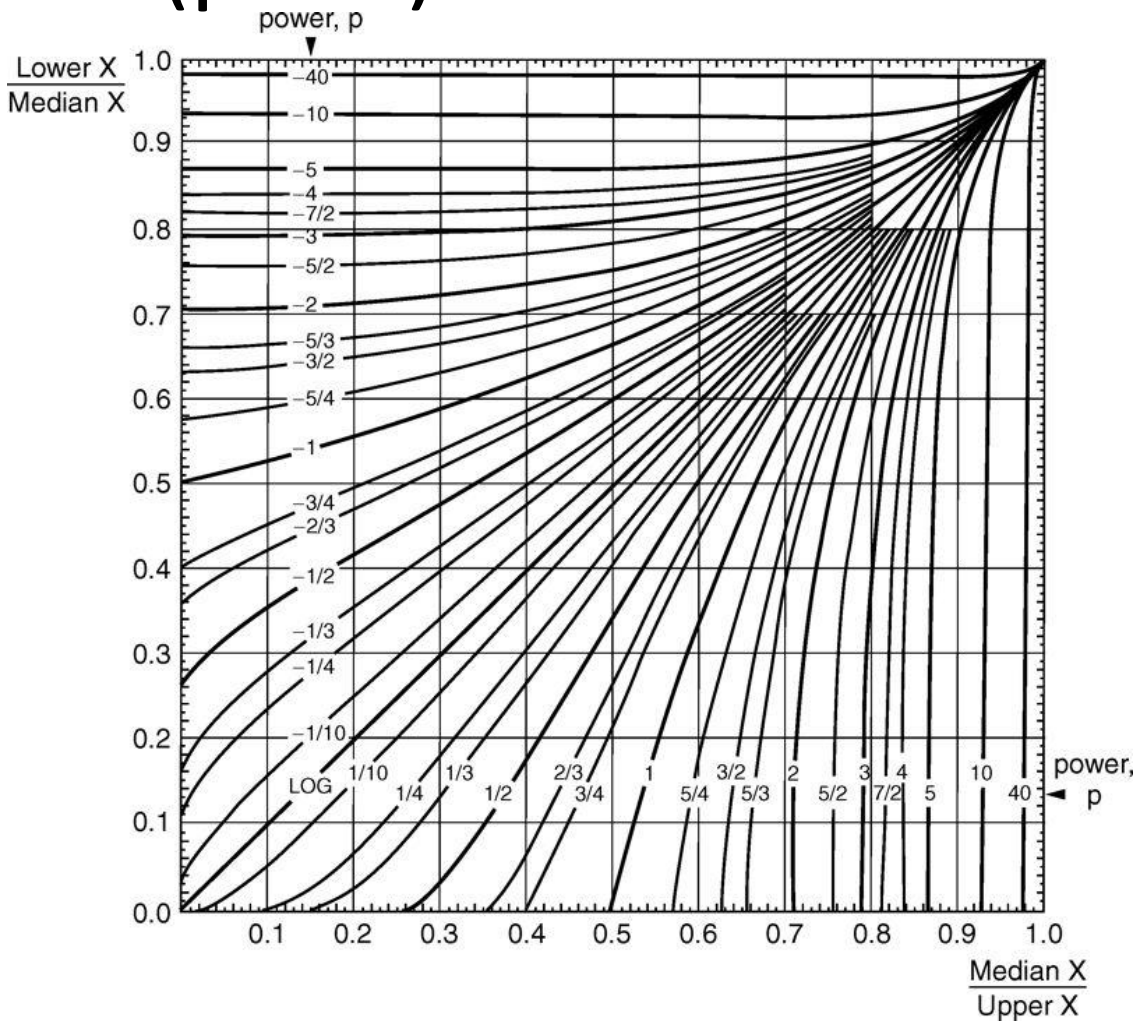
Power Transformation for Normality (p. 58)

Figure 4.6 Determination of the Value of *p* in the Power Transformation to Produce Approximate Normality

$$(X + C)^p$$

Lower Quantile: Q(0.25)

Upper Quantile: Q(0.75)



Example: Depression data set (Table 3.3, p. 42)

X = Income ($K)

- M = median = 15, Q(0.25) = 9, and Q(0.75) = 28
- Median is not halfway between the two quartiles
- M/Q(0.75) = 0.54, Q(0.25)/M = 0.60
- From Figure 4.6, power = -1/3
- Since this is between zero and -1/2,
  - Try log transformation or 1/square root
  - Log will work fairly well

Power Transformation for Normality (p. 58)

Q(0.25)/M = 0.60

M/Q(0.75) = 0.54

$(X)^p$

$$p = -\frac{1}{3}$$

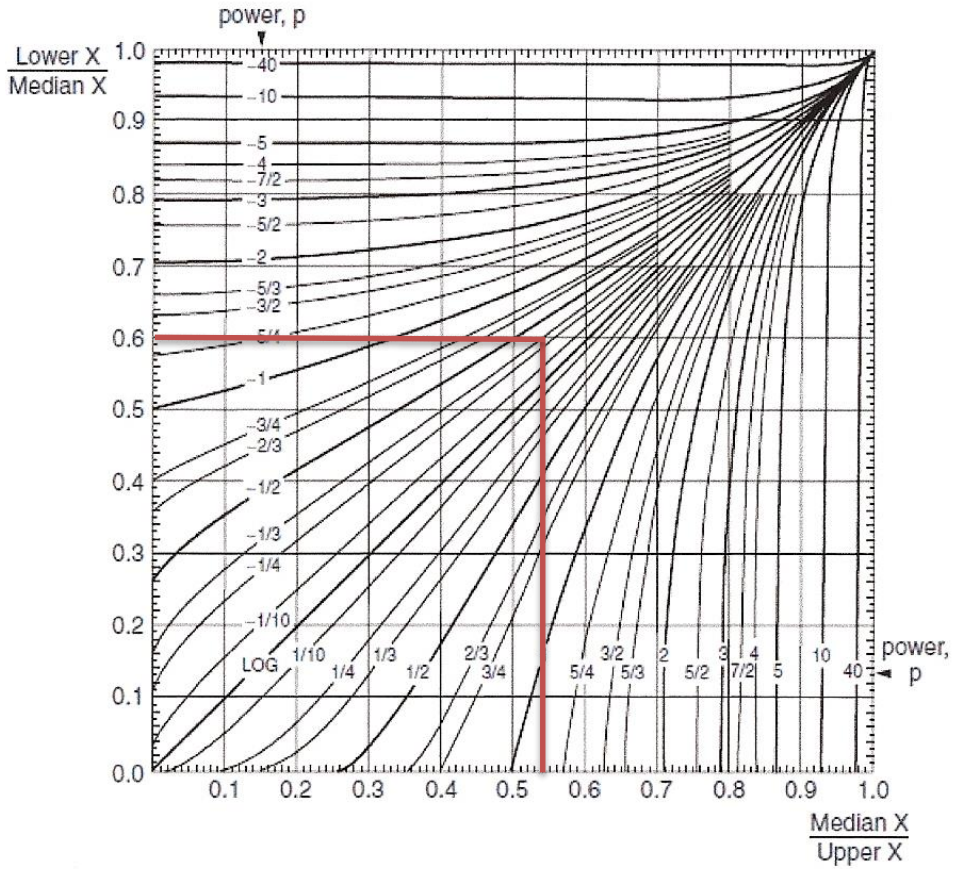
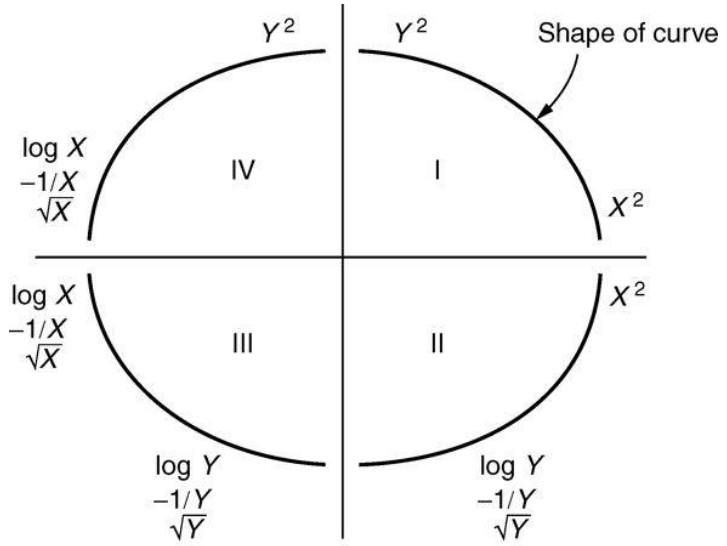


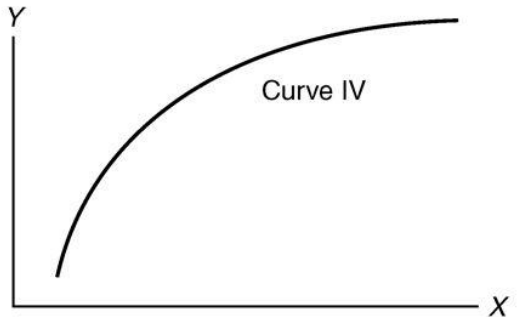
Figure 4.6: *Determination of the Value of p in the Power Transformation to Produce Approximate Normality*

Example Continued (p. 105)

Figure 6.9
Choice of Transformation: Typical Curves and Appropriate Transformation



a. Curves Not Linear in X



b. Detail of Fourth Quadrant

Skewed Data

In a variety of applications, histograms with long tails are common (e.g. figure 4.4c)

- Log and square root transformations are common approaches to this situation
- Use normal probability plots to select best fit.
- Dealing with Zeros before taking Log:
  - Add a constant to each observation before taking log
  - Constant = smallest non-zero observed value in the data set

Assessing Independence

- Between Individuals - i.e. rows
- Over Time
- Caveats:
  - Violations can affect inference

Lack of Independence (p 63)

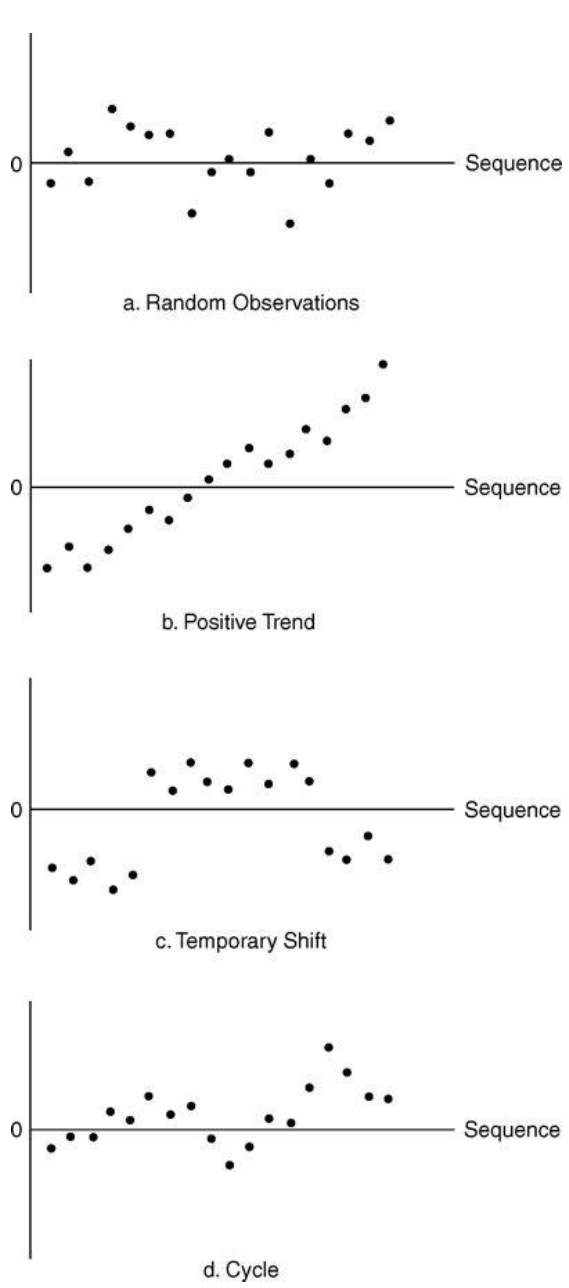


Figure 4.7
Graphs of Hypothetical Data Sequences Showing Lack of Independence

Overview of Multivariate Analysis

Many Observations from each Individual

Example: Depression Study (Chapter 3)

- 1000 adults interviewed 4 times over a 14-month period.
- Information obtained from each person included:
  - mental and physical health status
  - demographic variables
  - Life-style factors
  - Health care utilization

Many Observations from each Entity

Examples: From UCI, http://archive.ics.uci.edu/ml/datasets.html

- Credit Approval
- Internet Advertisement
- Statlog (Australian Customer Credit)
- Wines

Multivariate Analysis Toolbox is Large and Varied

- Linear Regression – relationship between a set of variables and a continuous outcome
- Discriminant analysis – classifying entities into populations
- Logistic regression – relationship between a set of variables and a categorical outcome
- Canonical correlation analysis – relationship between two sets of several variables

Multivariate Analysis Toolbox is Large and Varied (2)

- Principal component analysis – restructuring data
- Factor analysis – further restructuring data
- Cluster analysis – grouping entities
- Log-linear models – high dimensional contingency tables
- Multi Dimensional Scaling – graphical representation of data for dimension reduction
- Analysis of Variance
- Many more.

Selecting Appropriate Analyses

Some Considerations:

- Purpose of the Analysis
- Types of Variables in the Data Set
- Assumptions needed, Assumptions satisfied
- Choice is often arbitrary, consider several.

Types of Variables - Scales

- **Nominal:** several distinct categories – no natural ordering.
- **Ordinal:** Categories that can be Ordered
- **Interval:** differences between successive values are always the same
- **Ratio:** interval variable with values having a natural 0
- ***Also*:** categorical, discrete, continuous

Stevens's Measurement System (p 18)

Type of measurement	Basic empirical operation	Examples
Nominal	Determine equality of categories	Company names Race Religion Soccer players' numbers
Ordinal	Determine greater than or less than (ranking)	Hardness of minerals Socioeconomic status Rankings of wines
Interval	Determine equality of differences between levels	Temperature in degrees Fahrenheit Calendar dates
Ratio	Determine equality of ratios of levels	Height Weight Density Difference in time

Comparison of measurement scales

Incremental Progress	Measure Property	Mathematical Operators	Advanced Operations	Central Tendency
Nominal	Classification, Membership	=, !=	Grouping	Mode
Ordinal	Comparison, Level	>, <	Sorting	Median
Interval	Difference, Affinity	+, -	Yardstick	Mean, Deviation
Ratio	Magnitude, Amount	*, /	Ratio	Geometric Mean, Coeff. of Variation

Geometric mean (or mean proportional):

$$\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \cdots a_n}$$

https://en.wikipedia.org

Selecting Analysis (p 75)

	Independent variables			
	Nominal or ordinal		Interval or ratio	
Dependent variable(s)	1 variable	>1 variable	1 variable	>1 variable
No dependent variables	χ^2 goodness of fit	Measures of association Log-linear model (17) χ^2 test of independence	Univariate statistics (e.g., one-sample t tests) Descriptive measures (5) Tests for normality (4)	Correlation matrix (7) Principal components (14) Factor analysis (15) Cluster analysis (16)
Nominal or ordinal				
1 variable	χ^2 test Fisher's exact test	Log-linear model (17) Logistic regression (12) Poisson regression (12)	Discriminant function (11) Logistic regression (12) Univariate statistics (e.g., two-sample t tests)	Discriminant function (11) Logistic regression (12) Poisson regression (12)
>1 variable	Log-linear model (17)	Log-linear model (17)	Discriminant function (11)	Discriminant function (11)
Interval or ratio				
1 variable	t- test Analysis of variance Survival analysis (13)	Analysis of variance Multiple-classification analysis Survival analysis (13)	Linear regression (6, 18) Correlation (6) Survival analysis (13)	Multiple regression (7–9, 18) Survival analysis (13)
>1 variable	Multivariate analysis of variance Analysis of variance on principal components Hotelling's T^2 Profile analysis (16)	Multivariate analysis of variance Analysis of variance on principal components	Canonical correlation (10)	Canonical correlation (10) Path analysis Structural models (LISREL, Mplus)

Example

- Data:
  - 5 independent variables: 3 interval, 1 ordinal, 1 nominal
  - 1 dependent variable: interval
- Possible Analyses:
  - Multiple Regression
    - Pretend independent ordinal variable is an interval variable
    - Use dummy (0/1) variables for nominal variables
  - Analysis of Variance
    - Categorize all independent variables

Example

- More Possible Analyses:
  - Logistic Regression
    - Categorize dependent variable, high/low
  - Analysis of covariance:
    - Leave variables as they are
    - Check assumptions
  - Survival analysis
    - If dependent variable is time to an event

Compare Results

Simple Regression

Visit:

http://www.ats.ucla.edu/stat/sas/output/reg.htm

Aims

- Describe the relationship between an independent variable X, and a continuous dependent variable Y as a straight line in R^2
  - Two Cases:
    - Fixed X: values of X are preselected by investigator
    - Variable X: a random sample of (X,Y) pairs
- Draw inferences regarding the relationship
- Predict the value of Y for a given X

Mathematical Model

- The mean of Y values at a given X is:

$$\mu_{(Y|X)} = \beta_0 + \beta_1 X$$

- Variance of Y values at any X is σ^2

  (For all X)

- Y values are normally distributed at each X (needed for inference)

11.1

Overview of Linear Models

- An equation can be fit to show the best linear relationship between two variables:

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

Where Y is the dependent variable and

X is the independent variable

β_0 is the Y-intercept

β_1 is the slope

Steps in Simple Regression

1. **State the research hypothesis.**
2. **State the null hypothesis**
3. **Gather the data**
4. **Assess each variable separately first (obtain measures of central tendency and dispersion (descriptive statistics); frequency distributions; graphs); is the variable normally distributed?**
5. **Calculate the regression equation from the data**
6. **Calculate and examine appropriate measures of association and tests of statistical significance for each coefficient and for the equation as a whole**
7. **Accept or reject the null hypothesis**
8. **Reject or accept the research hypothesis**
9. **Explain the practical implications of the findings**

α and β (p 86)

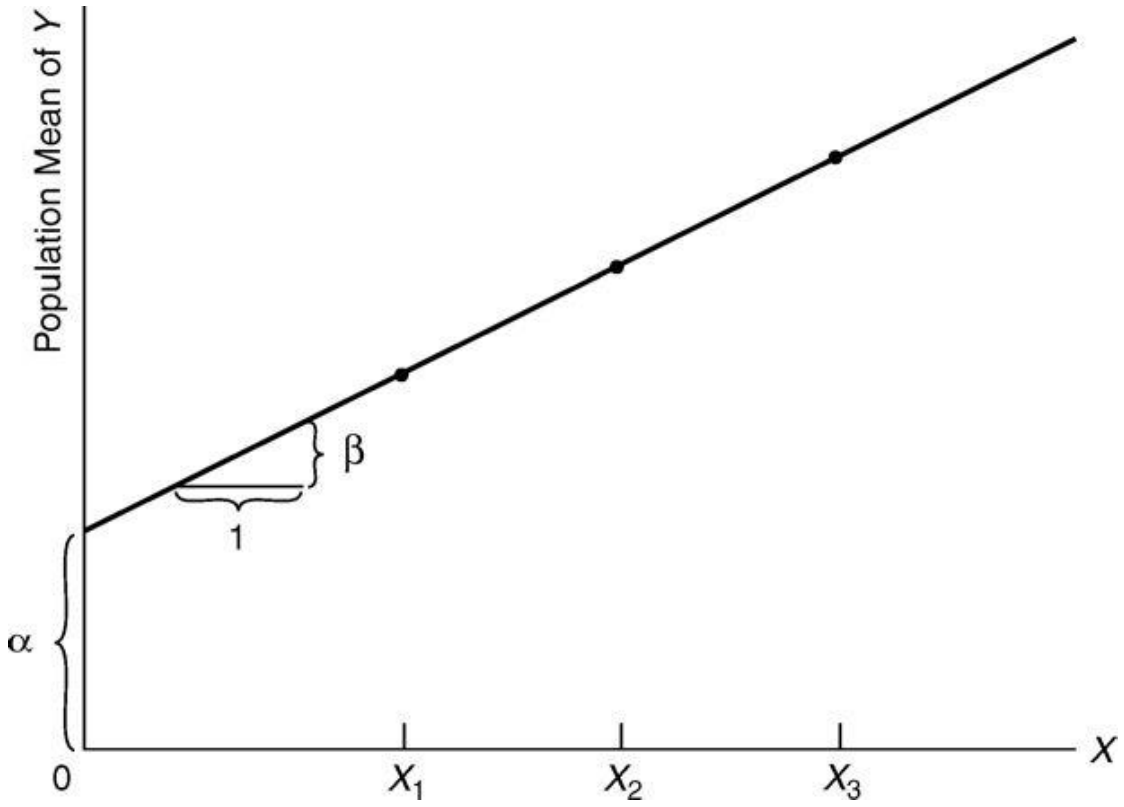


Figure 6.3 Theoretical Regression Line Illustrating α and β

Graphically (p 85)

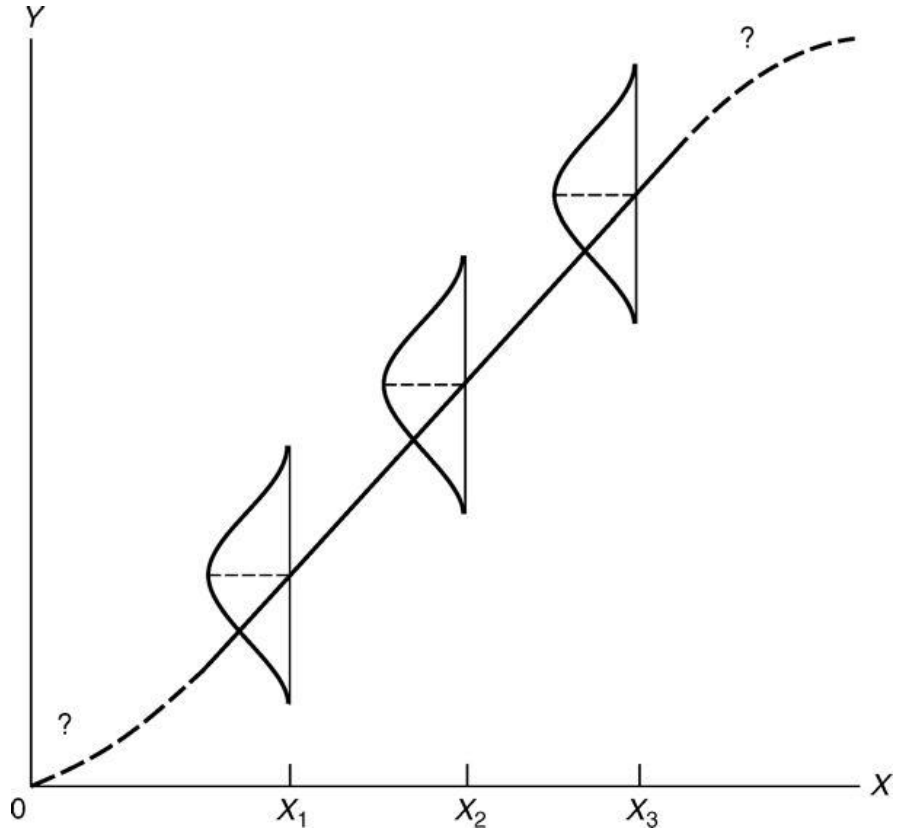
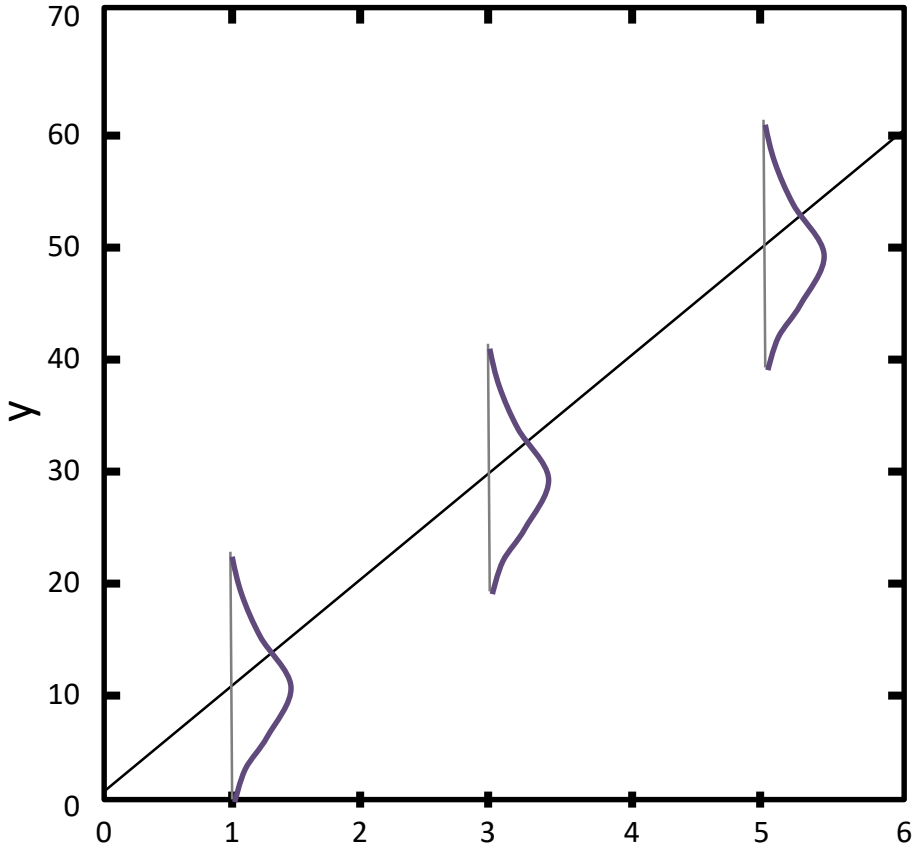
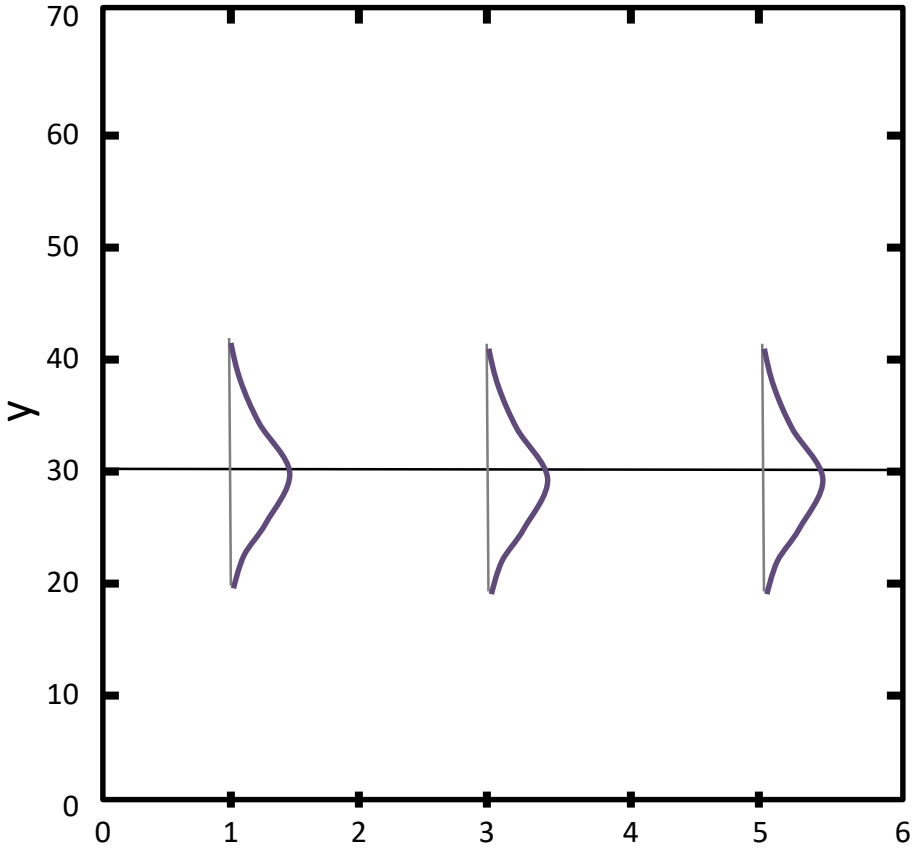


Figure 6.2 Simple Linear Regression Model for Fixed *X*'s

Hypothesis Tests



H_0: $\beta_0 = 0$
H_a: $\beta_0 \neq 0$



H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Hypothesis Tests

- Hypothesis Test for β_0
  - H_0: $\beta_0 = 0$
  - H_a: $\beta_0 \neq 0$

  $P_{\text{intercept}} \geq 0.05 \rightarrow$ Accept the Null Hypothesis

  $P_{\text{intercept}} < 0.05 \rightarrow$ Reject the Null Hypothesis

- Hypothesis Test for β_1
  - H_0: $\beta_1 = 0$
  - H_a: $\beta_1 \neq 0$

  $P_{\text{slope}} \geq 0.05 \rightarrow$ Accept the Null Hypothesis

  $P_{\text{slope}} < 0.05 \rightarrow$ Reject the Null Hypothesis

Least Squares Regression

- Estimates for coefficients b_0 and b_1 are found using a Least Squares Regression technique
- The least-squares regression line, based on sample data, is

$$\hat{y} = b_0 + b_1 x$$

- Where b_1 is the slope of the line and b_0 is the y-intercept:

$$b_1 = \frac{Cov(x,y)}{s_x^2} = r\left(\frac{s_y}{s_x}\right)$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Introduction to Regression Analysis

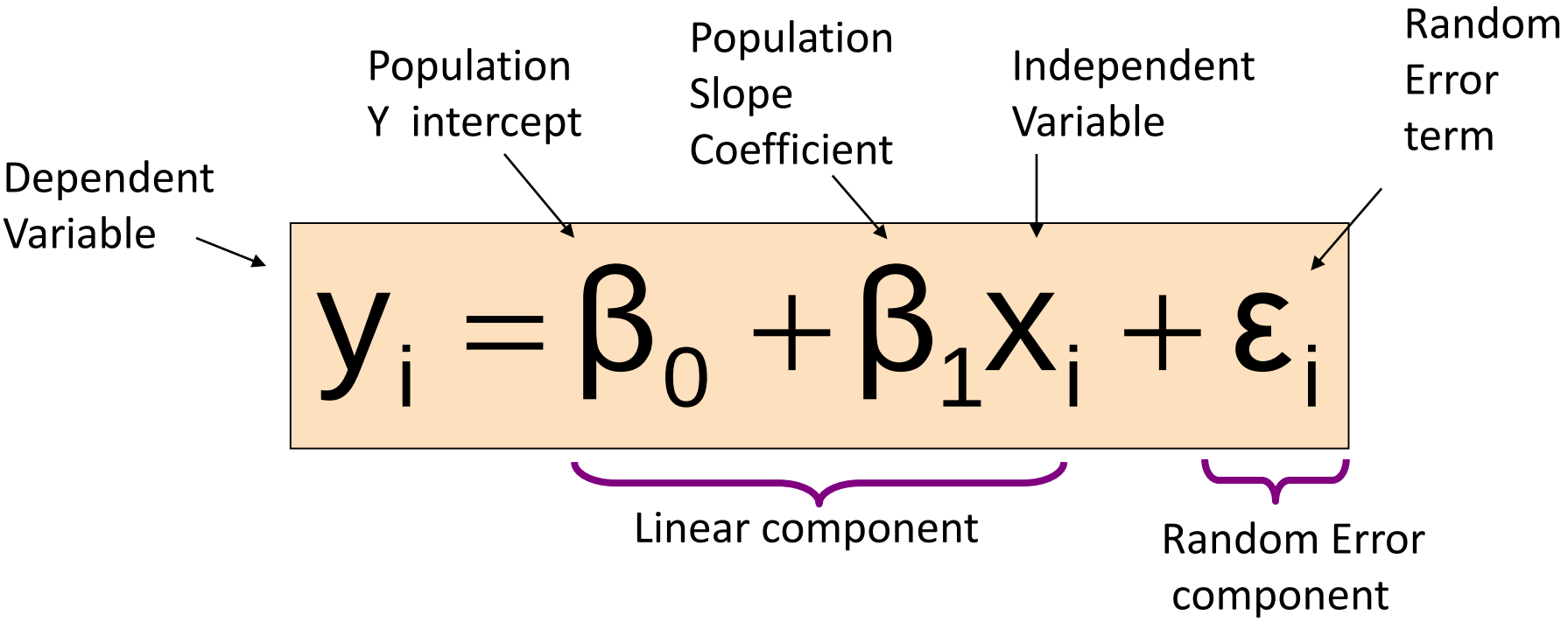
- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

Simple Linear Regression Model

The population regression model:



Linear Regression Assumptions

- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the X values
- The error terms are random variables with mean 0 and constant variance, σ^2

  (the uniform variance property is called homoscedasticity)

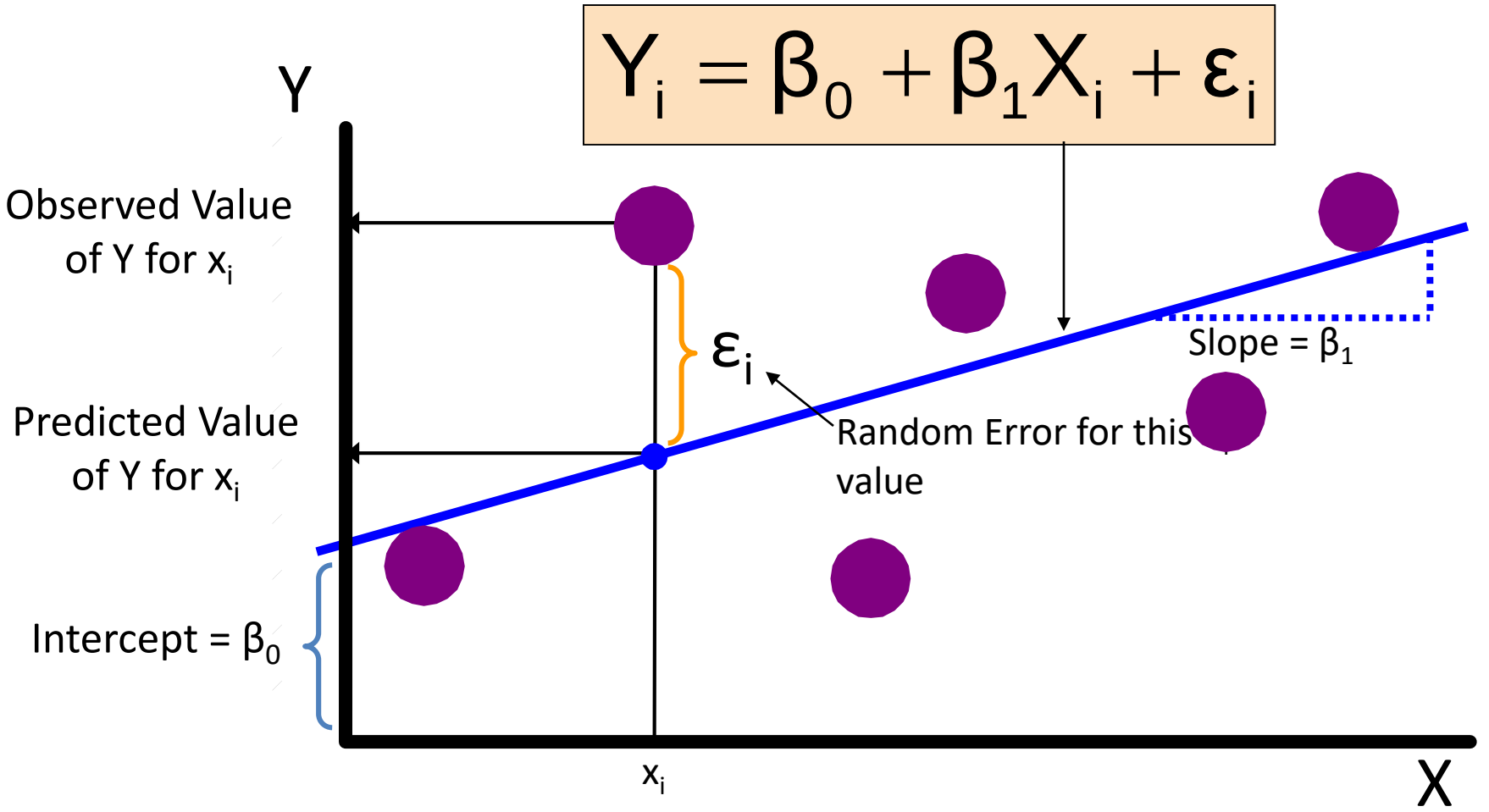
$$E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{for } (i = 1, \ldots, n)$$

- The random error terms, ε_i, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{for all } i \neq j$$

Simple Linear Regression Model

(continued)



Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

$$\hat{y}_i = b_0 + b_1 x_i$$

The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Least Squares Coefficient Estimators

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals (errors), SSE:

$$\min \text{ SSE} = \min \sum_{i=1}^{n} e_i^2$$

$$= \min \sum (y_i - \hat{y}_i)^2$$

$$= \min \sum [y_i - (b_0 + b_1 x_i)]^2$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Least Squares Coefficient Estimators

(continued)

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2} = r\frac{s_y}{s_x}$$

- And the constant or y-intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

- The regression line always goes through the mean $\bar{x}$, $\bar{y}$

Example:
Linear Regression using Least Square

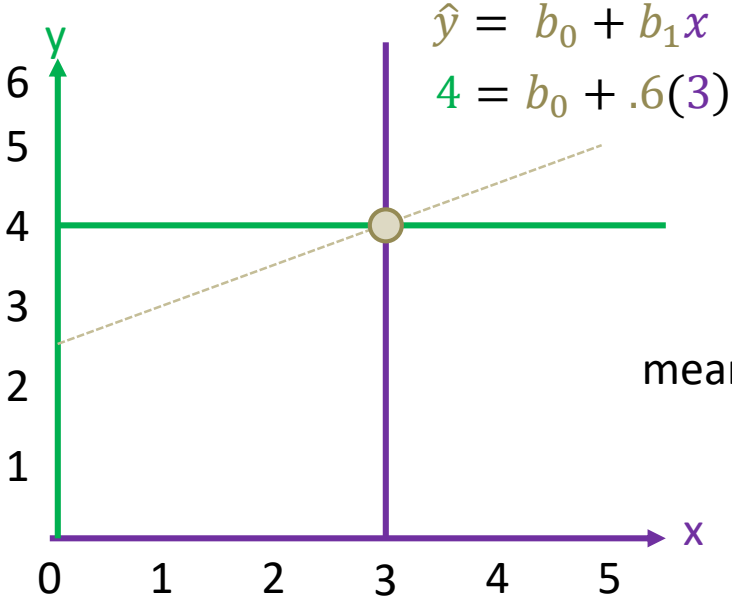
$\beta_0 = 2.2$

$\beta_1 = 0.6$

$\hat{y} = 2.2 + 0.6x$

x	y	x - $\overline{x}$	y - $\overline{y}$	$(x-\overline{x})^2$	$(x-\overline{x})(y-\overline{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
mean 3	4			10	6

$$b_1 = \frac{6}{10} = .6 = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}$$



$4 = b_0 + .6(3)$

$4 = b_0 + (1.8)$ ⇒ $2.2 = b_0$

Computer Analysis

Results:

- estimates of slope (β_1) and intercept (β_0), using least squares
- residual mean square = estimate of variance (S^2)
- test if $\beta = \beta_0$
  - Usually, test $\beta = 0$, i.e. X has no effect on Y

Model Printout

UCLA Document Reg

Annotated SAS Output: Regression

Prediction

- The regression equation can be used to predict a value for y, given a particular x

- For a specified value, x_{n+1} , the predicted value is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Confidence & Prediction Intervals

- 100(1-α) % of confidence interval
  - β_0: $\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2} \cdot se(\hat{\beta}_0)$
  - β_1: $\hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot se(\hat{\beta}_1)$

Two Types for Y (p 88):

- Confidence interval (CI) for mean of Y
- Prediction interval (PI) for individual Y

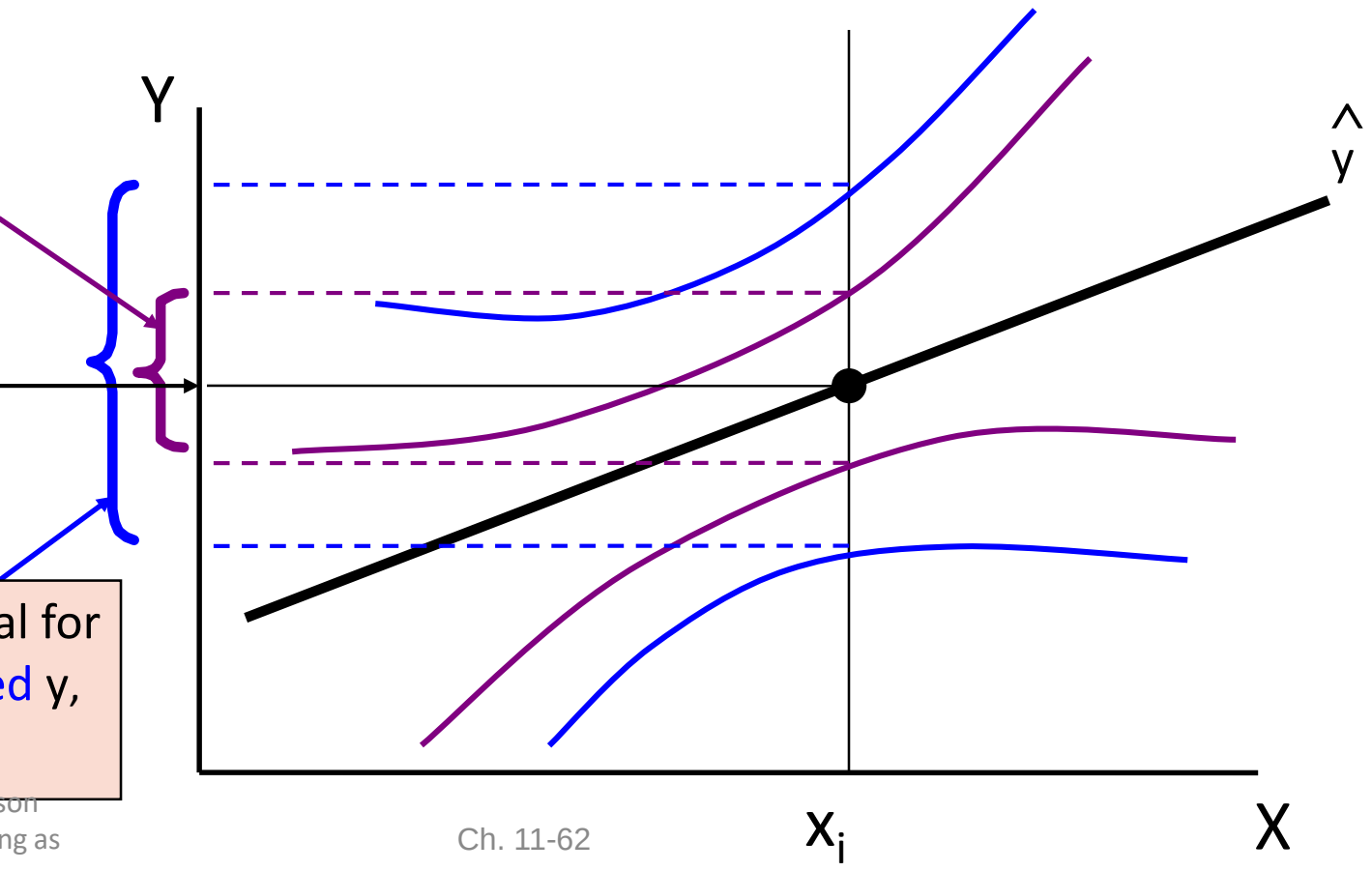
Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around y to express uncertainty about the value of y for a given x_i

Confidence Interval for the expected value of y, given x_i

$\hat{y} = b_0 + b_1 x_i$

Prediction Interval for an single observed y, given x_i



Confidence Interval for the Average Y, Given X

Confidence interval estimate for the **expected value of y** given a particular x_i

Confidence interval for $E(Y_{n+1} \mid X_{n+1})$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

Notice that the formula involves the term $(x_{n+1} - \bar{x})^2$ so the size of interval varies according to the distance x_{n+1} is from the mean, $\bar{x}$

Prediction Interval for an Individual Y, Given X

Prediction interval estimate for an **actual observed value of y** given a particular x_i

Prediction interval for $\hat{y}_{n+1}$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Correlation Coefficient - ρ

- Correlation coefficient measures the strength of linear association between X and Y in the population (ρ).
- it is estimated by sample (r)

11.7

Correlation Analysis

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation
  - Correlation was first presented in Chapter 4

Correlation Analysis

- The population correlation coefficient is denoted ρ (the Greek letter rho)
- The sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation Coefficient - ρ

- σ^2 = Var (Y|X) = Variance of Y for a specific X
- σ_y^2 = Var (Y) = Variance of Y for all X's
- $100\ (1 - \rho^2)^{1/2}$ = % of Standard Deviation NOT "explained" by X

$$\sigma^2 = \sigma_y^2 (1 - \rho^2)$$

$$\Rightarrow \sigma = \sigma_y \sqrt{1 - \rho^2}$$

$$\Rightarrow \rho^2 = \frac{\sigma_y^2 - \sigma^2}{\sigma_y^2}$$

Interpretation of ρ

- ρ^2 = reduction in variance of Y associated with knowledge of X/original variance of Y
- $100\rho^2$ = % of variance of Y "explained by X"

Caveat: correlation vs causation

- Outliers
- Linear
- There is no prove

Estimating the value of ρ (Pearson's Correlation Coefficient)

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$r = \frac{S_{XY}}{S_X S_Y}$$

$$S_{XY} = \sum (X - m(X))(Y - m(Y))/(N-1)$$

S_{XY} = covariance of X and Y
(joint variability of X and Y)

Graphically (p 92)

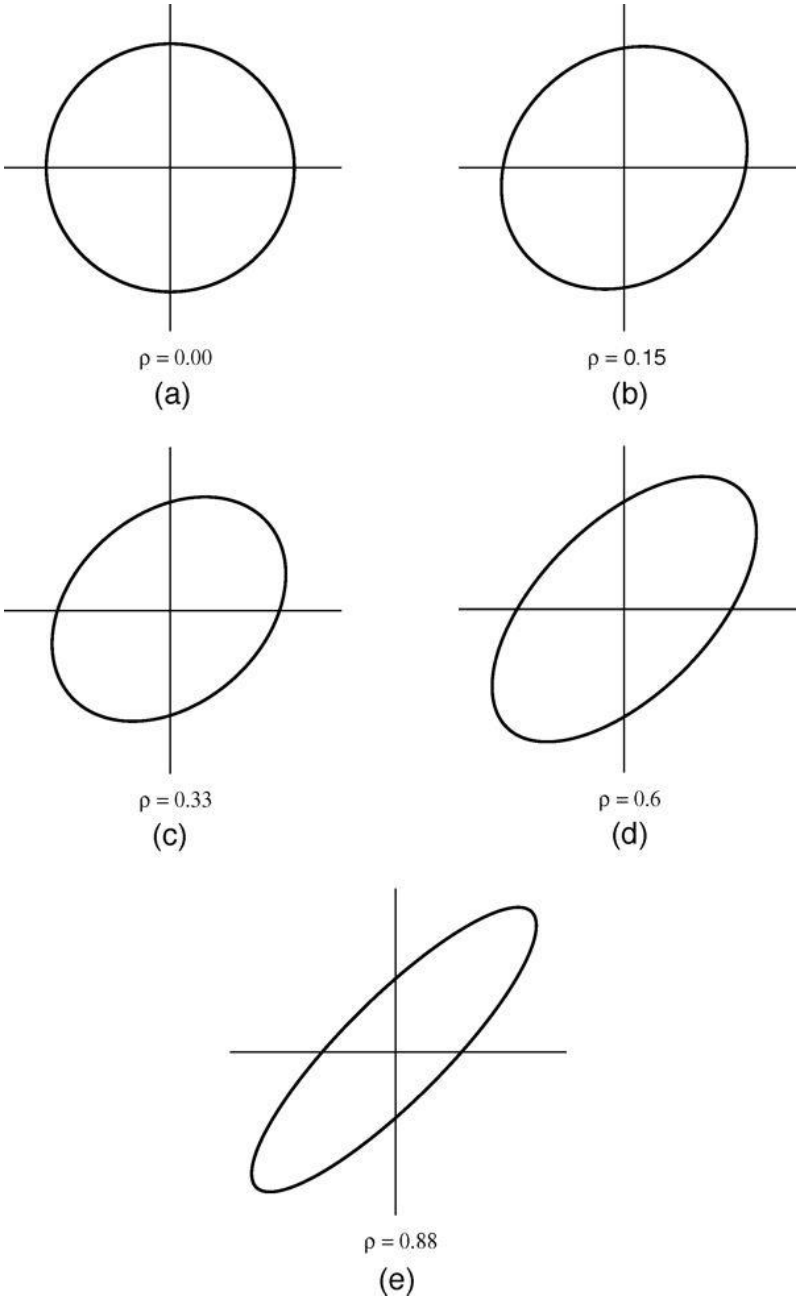
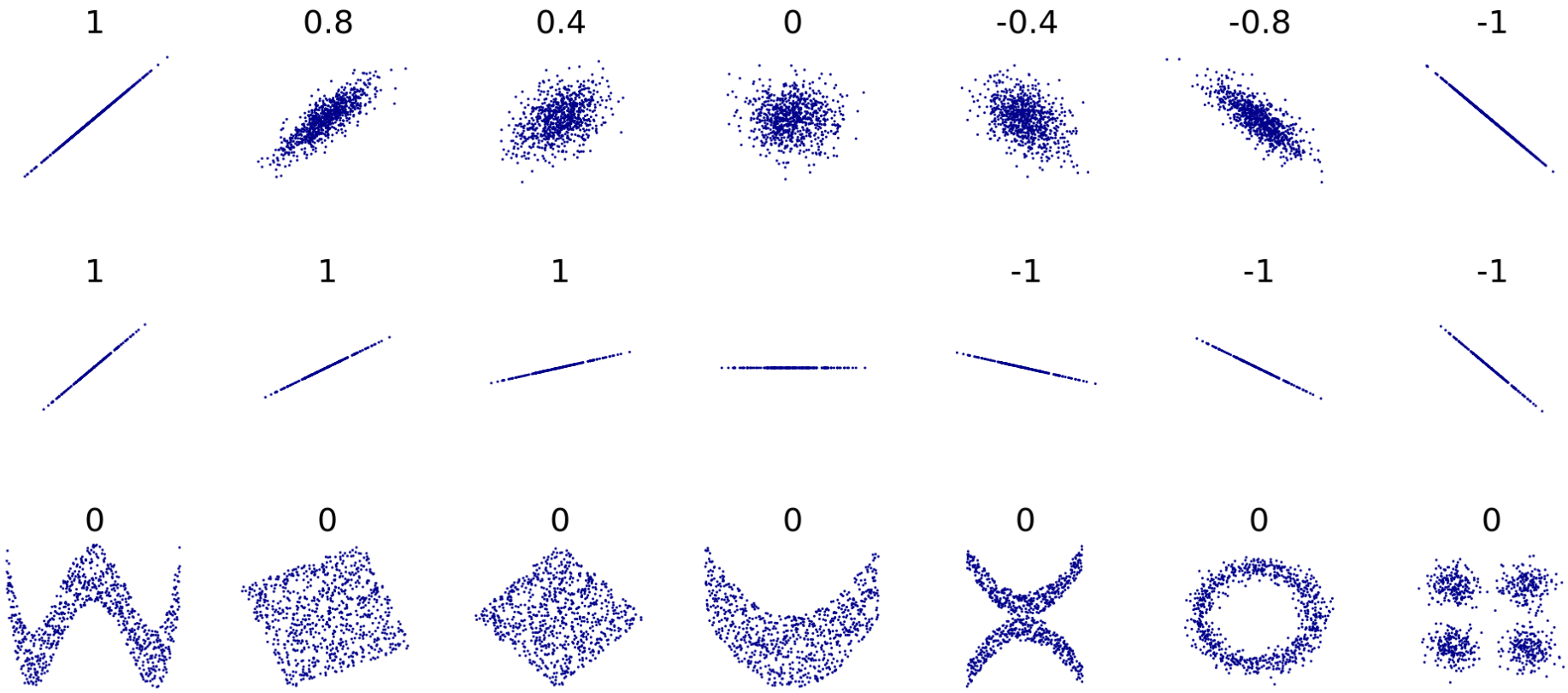


Figure 6.5
Ellipses of Concentration for Various ρ Values

Graphically



https://en.wikipedia.org/

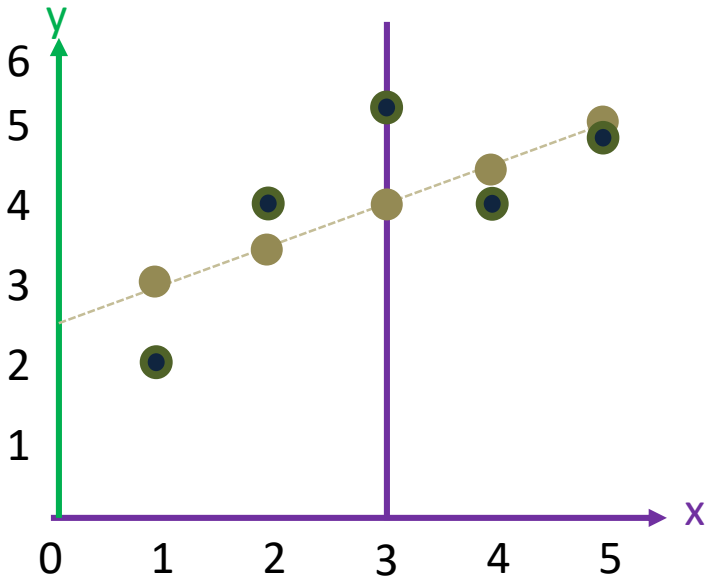
Interpretation of ρ

ρ	% of variance "explained"	% of variance not "explained"	% of SD "explained"	% of SD not "explained"
±0.3	9%	91%	5%	95%
±0.5	25%	75%	13%	87%
±0.71	50%	50%	29%	71%
±0.95	90%	10%	69%	31%

r Using Regression Analysis

For a sample:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	3	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
			6			3.6
mean	4					

$$R^2 = \frac{3.6}{6} = .6 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

Test for Zero Population Correlation

- To test the null hypothesis of no linear association,

$$H_0 : \rho = 0$$

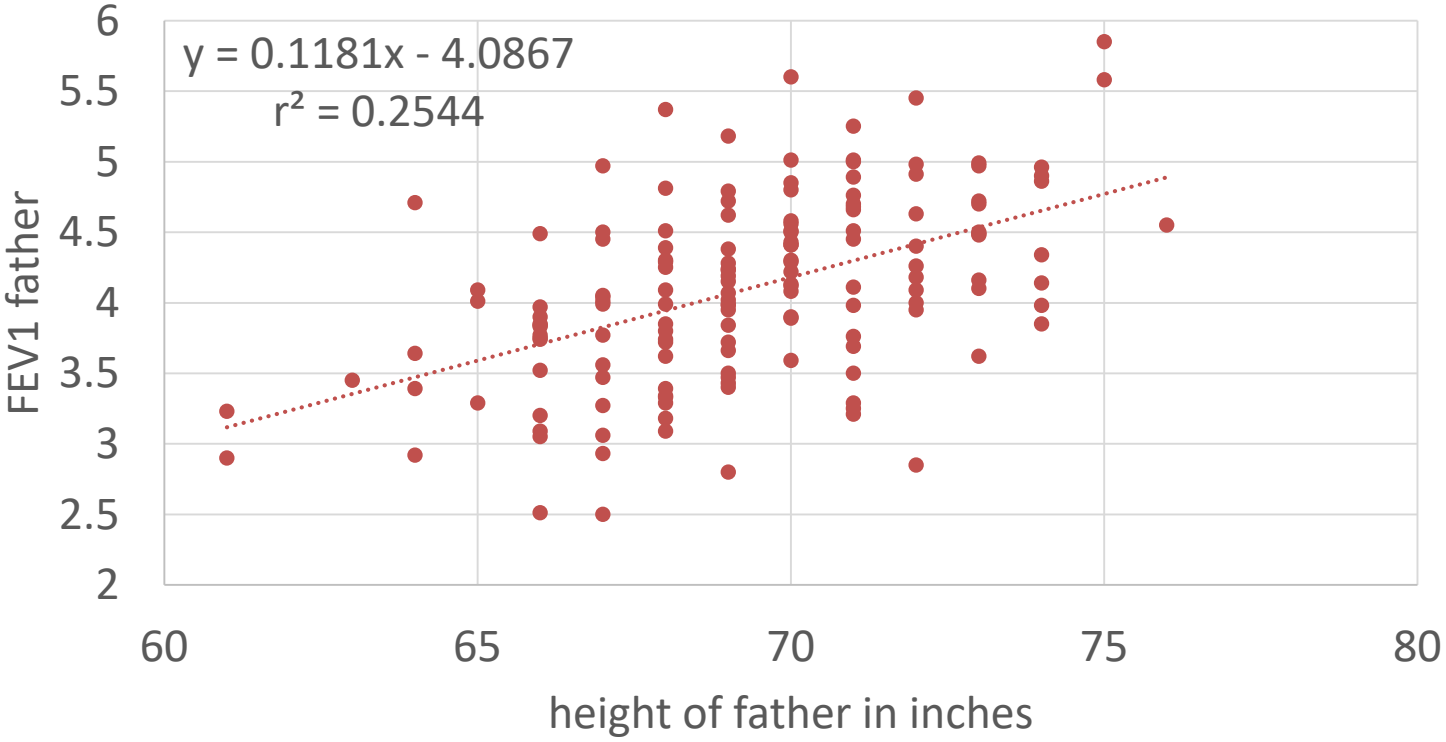
the test statistic follows the Student's t distribution with (n – 2) degrees of freedom:

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

Example from Text: Lung Function

- Data from an epidemiological study of households
  - living in four areas with different amounts and types of air pollution (Appendix A)
- Data only on non-smoking fathers
  - X = height in inches
  - Y = forced expiratory volume in 1 second (FEV1)

Scatter Plot



Example Results

- Least Squares Equation: $Y = -4.087 + 0.118\ X$
- Correlation $r = 0.504$
- Test $p = 0$,
  - $t = 7.1$ (p 94), $\rho < 0.0001$
  - t test can be one or two sided

Analysis of Variance

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between x and y

11.4

Explanatory Power of a Linear Regression Equation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error (residual) Sum of Squares
$SST = \sum(y_i - \bar{y})^2$	$SSR = \sum(\hat{y}_i - \bar{y})^2$	$SSE = \sum(y_i - \hat{y}_i)^2$

where:

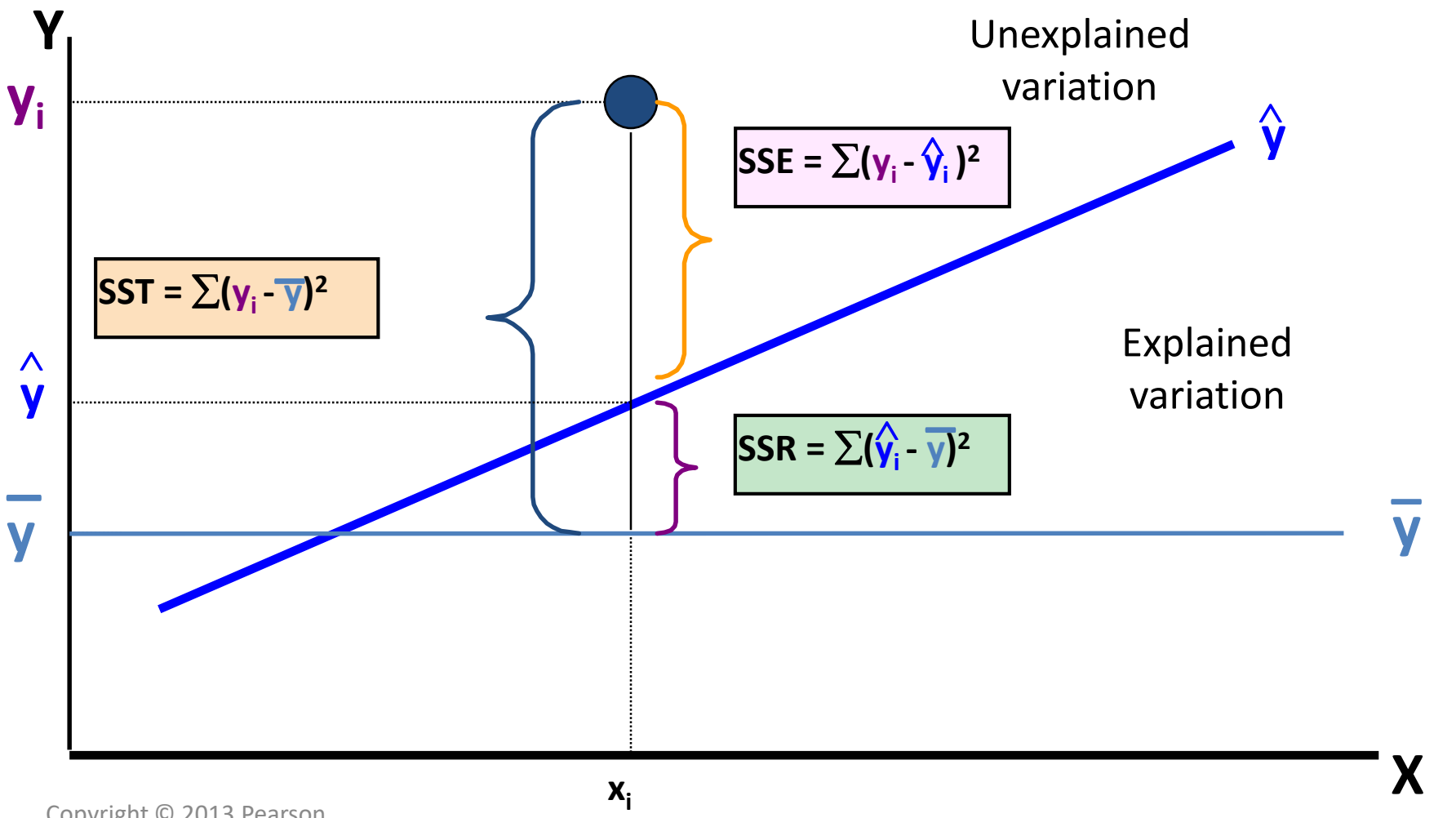
$\bar{y}$ = Average value of the dependent variable

y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of y for the given x_i value

Analysis of Variance

(continued)



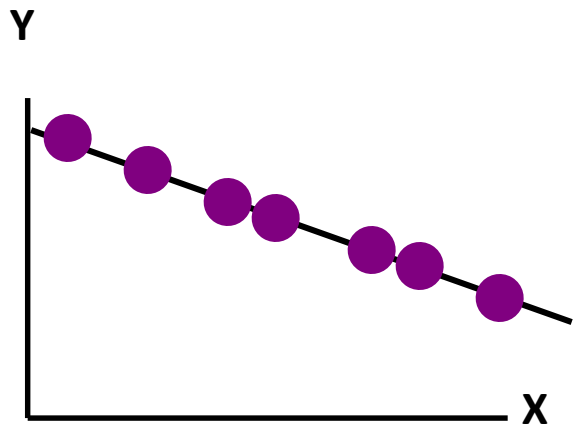
Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

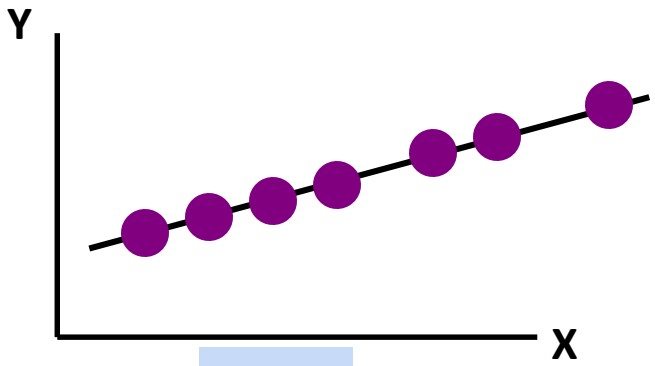
$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note: $0 \le R^2 \le 1$

Examples of Approximate r^2 Values



$R^2 = 1$

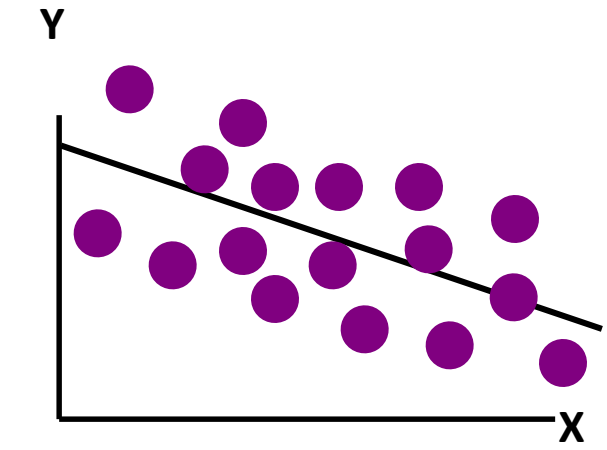


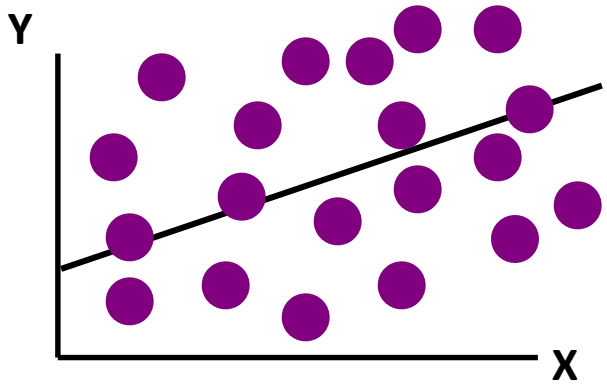
$R^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values



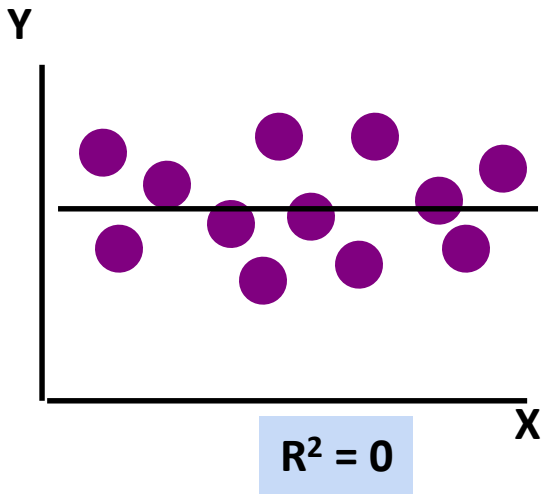


$0 < R^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values



$R^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Correlation and R^2

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$R^2 = r^2$$

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

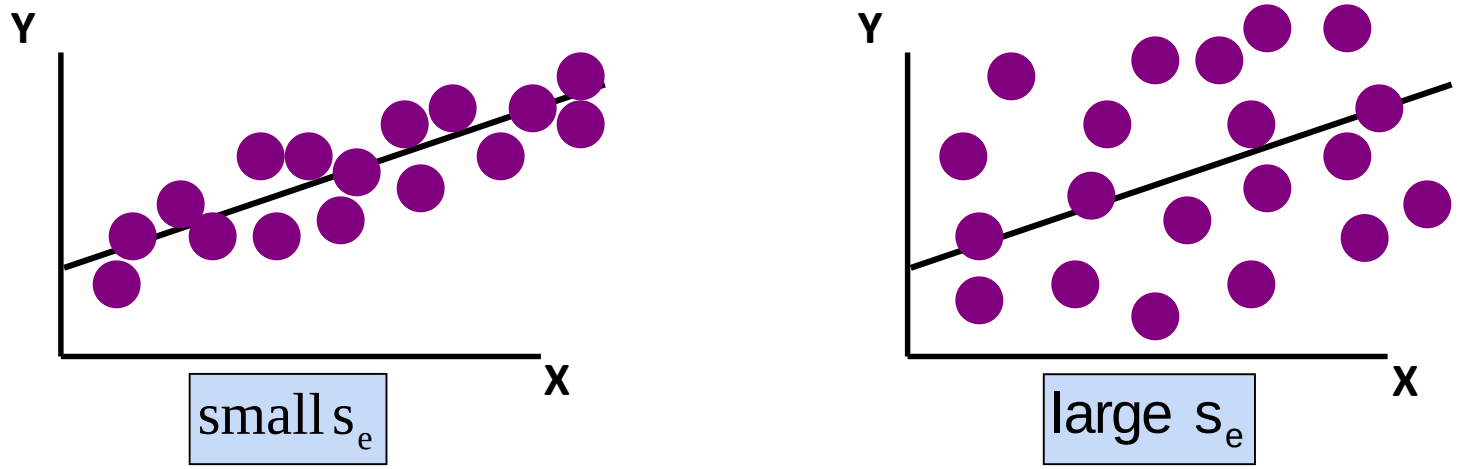
$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

- Division by n – 2 instead of n – 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$s_e = \sqrt{s_e^2}$ is called the standard error of the estimate

Comparing Standard Errors

s_e is a measure of the variation of observed y values from the regression line



The magnitude of s_e should always be judged relative to the size of the y values in the sample data

Statistical Inference: Hypothesis Tests and Confidence Intervals

- The variance of the regression slope coefficient (b_1) is estimated by

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}$$

where:

S_{b_1} = Estimate of the standard error of the least squares slope

$$s_e = \sqrt{\frac{SSE}{n-2}}$$ = Standard error of the estimate

Hypothesis Test for Population Slope Using the F Distribution

- F Test statistic:

where

$$F = \frac{MSR}{MSE}$$

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

where F follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

Hypothesis Test for Population Slope Using the F Distribution

(continued)

- An alternate test for the hypothesis that the slope is zero:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

- Use the F statistic

$$F = \frac{MSR}{MSE} = \frac{SSR}{s_e^2}$$

- The decision rule is

reject H_0 if $F \geq F_{1,n-2,\alpha}$

F-Test for Significance

(continued)

$H_0: \beta_1 = 0$

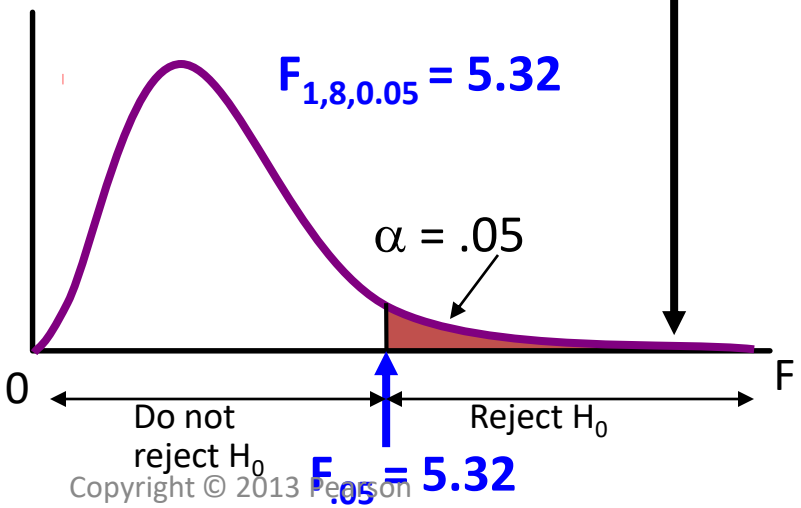
$H_1: \beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \quad df_2 = 8$

Critical Value:

$F_{1,8,0.05} = 5.32$



Test Statistic:

$$F = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

F (Fisher) Distribution

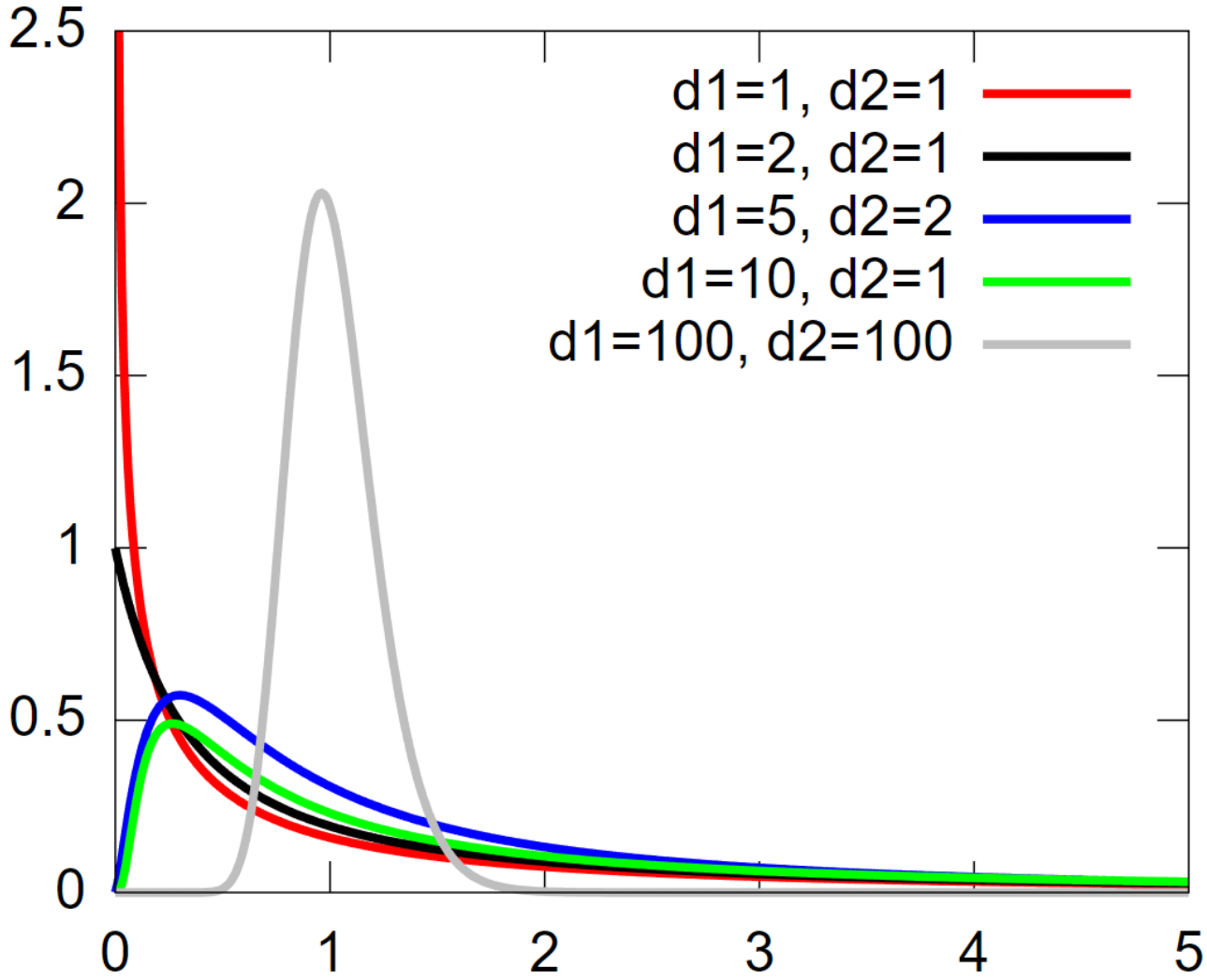
The F Distribution can be used to test whether the variances of 2 populations are equal:

- Taking 2 distributions with equal variance, both normally distributed.
- Draw 2 independent random samples, one from each population, of size $\mathbf{n_1}$ **and** $\mathbf{n_2}$ respectively
- Estimate the variances of the two populations from the samples: $\mathbf{s_1^2}$ **and** $\mathbf{s_2^2}$ respectively
- The ratio of the 2 random variables is call “F” and is the ratio of Chi Squared Distribution divided by respective degrees of freedom,
  - $\chi 2(n_1-1)/(n_1-1) \,/\chi 2(n_2-1)/(n_2-1) = \mathbf{s_1^2 \,/\, s_2^2 = F}$
- A random variable formed by two independent chi-squared variables, each divided by it’s degrees of freedom, is an F-ratio, and has an F Distribution.
- https://www.easycalculation.com/statistics/f-test-p-value.php
- http://vassarstats.net/tabs.html

F Distribution Assumptions

- Both parent populations are normally distributed
- Both parent populations have the same variance
- The samples are independent

F Distribution



https://en.wikipedia.org

Example Results

- Least Squares Equation: $Y = -4.087 + 0.118X$
- Correlation $r = 0.504$
- Test $p = 0$,
  - $t = 7.1$ (p 94), $\rho < 0.0001$
  - t-test can be one or two sided

ANOVA Table (p. 96)

Source of variation	Sums of squares	df	Mean square	F
Regression	$\sum(\hat{Y}-\bar{Y})^2$	1	$SS_{reg}/1$	MS_{reg}/MS_{res}
Residual	$\sum(Y-\hat{Y})^2$	$N-2$	$SS_{res}/(N-2)$	
Total	$\sum(Y-\bar{Y})^2$	$N-1$		

Source of variation	Sums of squares	df	Mean square	F
Regression	16.0532	1	16.0532	50.50
Residual	47.0451	148	0.3179	
Total	63.0983	149		

Test $\beta_1 = 0$

- From ANOVA table:
  - $F > F_{\alpha,1,n-2}$
  - percentile of the F distribution corresponding to a cumulative probability of $(1- \alpha)$
- In this example:
  - $F = 50.5$ gives 2-sided test, p-value < 0.0001
- One sided test is: $t = F^{1/2} = 7.1$
- Same as test for $\rho = 0$

Residual Analysis

- Residual = $e = Y - \hat{Y}$
- Studentized residual = $\dfrac{\text{Residual}}{\text{estimate of residual standard deviation}} = \dfrac{e}{S\sqrt{1-h}}$
  - where *h* called "leverage"
- Leverage: shows how far away is an observation from others

Outliers

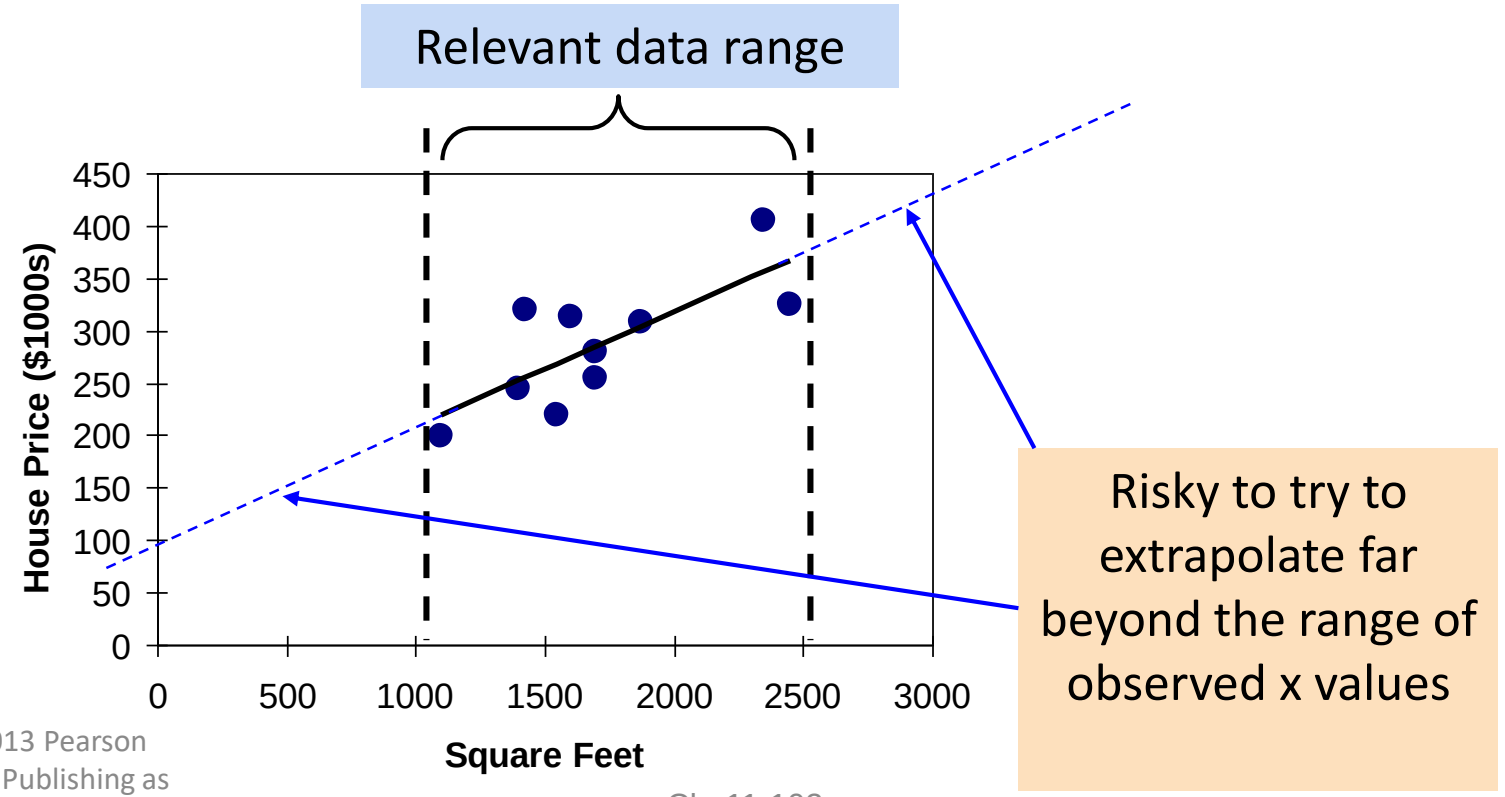
- Outlier in Y is studentized (or deleted studentized) residual >2

- Leverage = h = $\frac{1}{N} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}$
  - X's far from the mean of X have large leverage (h)
  - Observations with large leverage have large effect on the slope of the line.

- Outlier in X if h > 4/N

Relevant Data Range

- When using a regression model for prediction, only predict within the relevant range of data



Effect of Outliers (p 102)

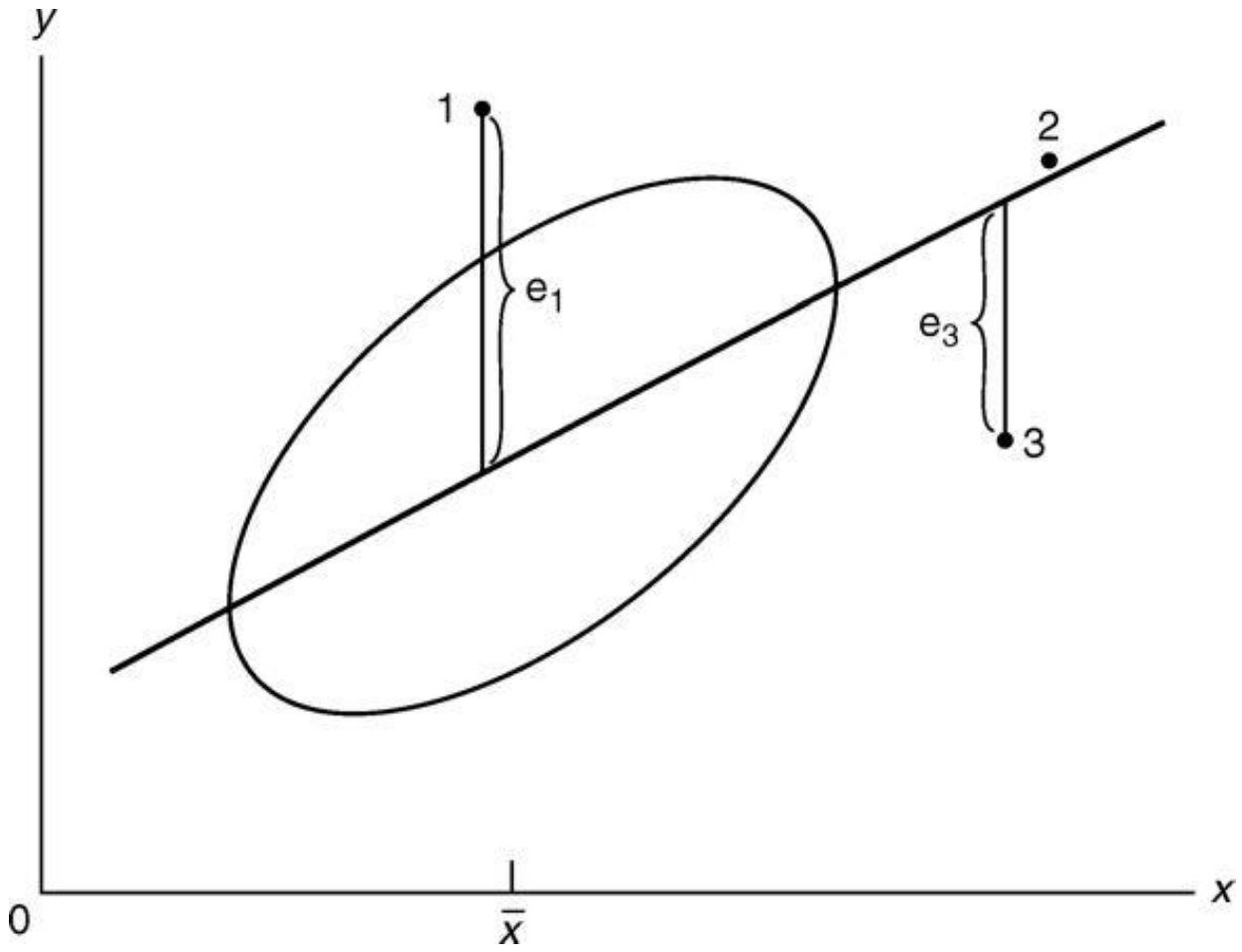


Figure 6.8 Illustration of the Effect of Outliers

Observations

- Point 1 is an outlier in Y with low leverage
  - impacts estimate of intercept but not slope
  - Tends to increase the estimates of S & SE of B
- Point 2 has high leverage; not an outlier in Y
  - doesn't impact estimate of B or A
- Point 3 has high leverage and is an outlier in Y
  - impacts the values of B, A, and S

Influential observations

An observation is influential if:

- It is an outlier in X and Y
- Cook's distance > $F_{0.5}(2,N-2)$
- DFFITS > $\frac{2\sqrt{p}}{\sqrt{N-2}}$

- Here, the number of parameters (p) = 2 and N is the number of points

Try analysis with and without influential observations and compare results.

Assumptions

- Homogeneity of variance (same σ^2)
  - Not extremely serious
  - Can be achieved through transformations if necessary
- Normal residuals
  - Slight departures ok
  - Can use transformations to achieve it
- Randomness
  - Serious
  - Can use hierarchical models for clustered samples

Checking Assumptions

- Plot residuals *vs* X or *vs* the predicted Y to check linearity and homogeneity of variance
- Create normal probability plots of residuals to check for normality

Residual Plots (p 98)

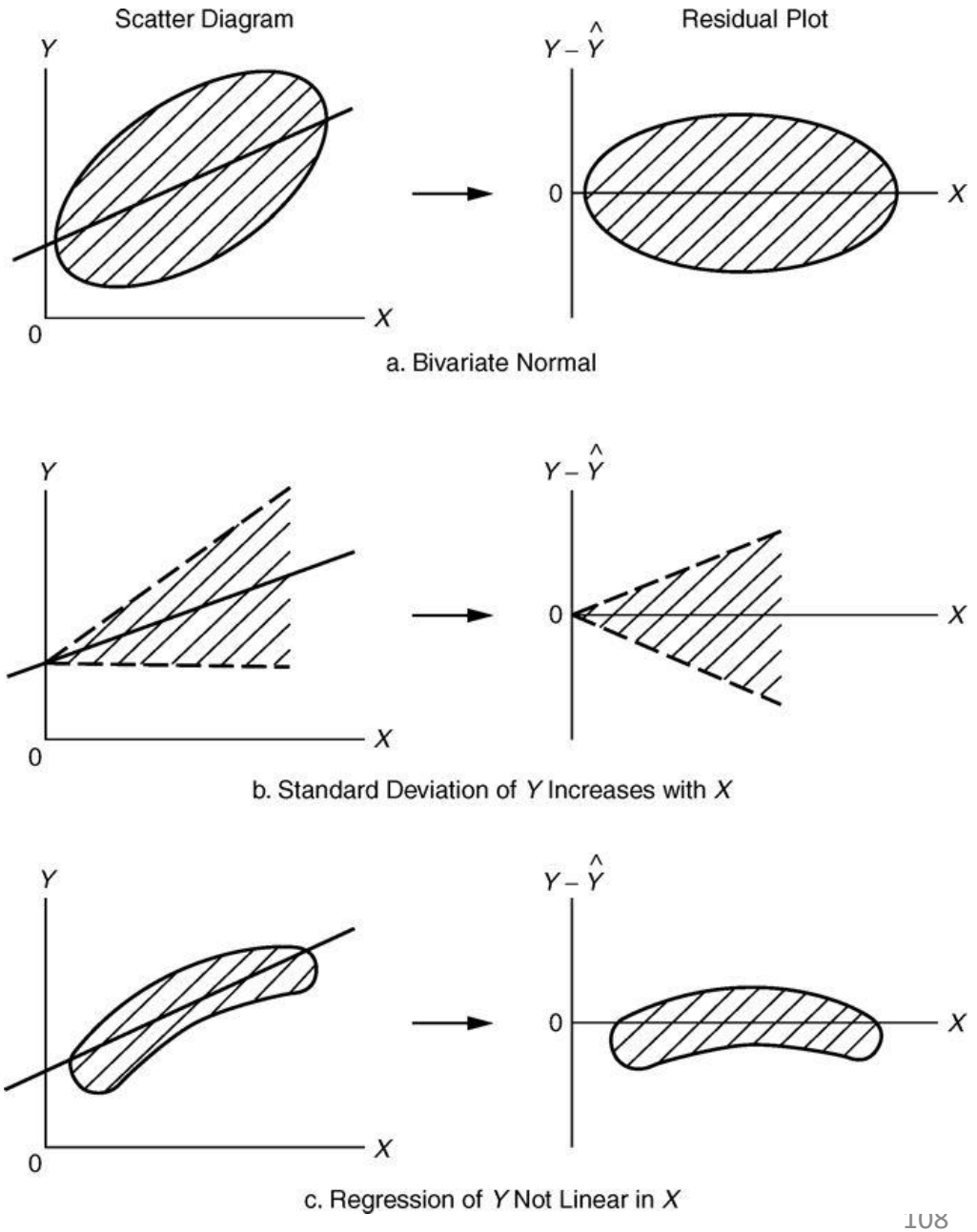


Figure 6.7 Hypothetical Scatter Plots and Correspondin g Residual Plots

Weighted Regression

- If σ^2 are not equal, use weight for each residual in the sum of squares used in Least Squares process.
- Weight = $1/\sigma^2$
- Gives unbiased estimate with smaller variance

Weighted Regression - Caveat

- In SAS, wrong answer possible
- Solution,, standardize weight (w) to add up to the sample size (N)
  - e.g. N = 5, w = 4,1,8,2,4, sum of w = 19
  - define standardized weight (sw) = w*5/19
  - sum of sw = 5
  - = 1.05 + .26 + 2.11 + .53 + 1.05 = 5

What to watch for

- Need representative sample
- Range of prediction should match observed range in X in sample
- Use of nominal or ordinal, rather than interval or ration data
- Errors in variables
- Correlation does not imply causation
- Violation of assumptions
- Influential points
- Appropriate model

Python3 code for simple regression

- For fit a linear model we usually use Ordinary Least squares(OLS)

[7]:
```
#package for simple regression
from statsmodels.formula.api import ols
```

[8]:
```
model = ols("y ~ x", data).fit()
```

[9]:
```
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.804
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     74.03
Date:                Fri, 08 Feb 2019   Prob (F-statistic):           8.56e-08
Time:                        21:21:16   Log-Likelihood:                -57.988
No. Observations:                  20   AIC:                             120.0
Df Residuals:                      18   BIC:                             122.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.5335      1.036     -5.342      0.000      -7.710      -3.357
x              2.9369      0.341      8.604      0.000       2.220       3.654
==============================================================================
Omnibus:                        0.100   Durbin-Watson:                   2.956
Prob(Omnibus):                  0.951   Jarque-Bera (JB):                0.322
Skew:                          -0.058   Prob(JB):                        0.851
Kurtosis:                       2.390   Cond. No.                         3.03
==============================================================================

Warnings:
```

Python code for ANOVA

- After you fit the model you can use anova_lm obtain a Anova table

```
In [14]: #ANOVA TEST
         import statsmodels.api as sm

In [16]: aov_table = sm.stats.anova_lm(model, typ=2)
         print (aov_table)
```

	sum_sq	df	F	PR(>F)
x	1588.873443	1.0	74.029383	8.560649e-08
Residual	386.329330	18.0	NaN	NaN

Python3 code: find outlier

- Firstly, we usually use visualization tool to detect the outlier

In [20]:
```
import seaborn as sns
sns.boxplot(x=data['y'])
```

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x1f19b57f8d0>

- After we spot a outlier, we can use IQR or Z-score to find this outlier and remove it.

In [57]:
```
import numpy as np

quartile_1, quartile_3 = np.percentile(data, [25, 75])
iqr = quartile_3 - quartile_1
lower_bound = quartile_1 - (iqr * 1.5)
upper_bound = quartile_3 + (iqr * 1.5)
outlier = np.where((data > upper_bound) | (data < lower_bound))
print(outlier)
```

```
(array([1, 2, 3, 5], dtype=int64), array([1, 1, 1, 1], dtype=int64))
```

Python3 code: split dataset and correlation coefficient

- Split dataset

```
print(outlier)
```

```
(array([1, 2, 3, 5], dtype=int64), array([1, 1, 1, 1], dtype=int64))
```

```
In [61]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(data, y, test_size=0.2# split data by 80% and 20%)
```

- Correlation coefficient: there are three different methods which are pearson, kendall, spearman (data type = pandas.df)

```
In [64]: data.corr(method = 'kendall')
```

Out[64]:

	x	y
x	1.000000	0.726316
y	0.726316	1.000000

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

stevens.edu

Amir H Gandomi; PhD
Assistant Professor of Analytics & Information Systems
a.h.gandomi@stevens.edu