

Introduction to Business Statistics: A Modeling Approach

Viswa Viswanathan

2025-12-04

Table of contents

Introduction to Business Statistics	5
Preface	6
I Module 1: Data, Visualization, and Patterns	7
Module 1: Data, Visualization & Patterns	8
Learning Goals	8
Structure of This Module	8
1 Data Frames	9
2 Point Plots	10
3 Relationships	11
II Module 2: Best Guess and Means	12
Module 2: Mean as the best guess	13
Learning Goals	13
Structure of This Module	13
4 Means	14
5 Category Means	15
6 Residuals	16
III Module 3: What Is a Model?	17
Module 3: Models	18
Learning Goals	18
Structure of This Module	18
7 Concept of model	19

IV Module 4: Regression Without Confidence Bands	20
Module 4: Linear Regression Models	21
Learning Goals	21
Structure of This Module	21
8 Line Models	22
9 Interpreting coefficients	23
V Module 5: Populations and Samples	24
Module 5: Population and Samples	25
Learning Goals	25
Structure of This Module	25
10 Population vs Sample	26
11 Sampling Variability	27
VI Module 6: Probability as Variation	28
Module 6: Probability	29
Learning Goals	29
Structure of This Module	29
12 Randomness	30
13 Variability	31
14 Distributions	32
VII Module 7: Sampling Variation and Estimators	33
Module 7: Sampling Variation	34
Learning Goals	34
Structure of This Module	34
15 SE Estimators	35
16 Sampling Distributions Slopes	36
VIII Module 8: Confidence Intervals and Bands	37
Module 8: Confidence Bands	38
Learning Goals	38
Structure of This Module	38

17 CI Slope	39
18 CI Bands	40
IX Module 9: Signal, Noise, and R-Squared	41
Module 9: Signal and Noise	42
Learning Goals	42
Structure of This Module	42
19 Variance Decomposition	43
20 R-Squared	44
X Module 10: Hypothesis Testing via Regression	45
Module 10: Hypothesis Testing	46
Learning Goals	46
Structure of This Module	46
21 t-tests for Regression	47
22 Interpretation	48
XI Module 11: Business Applications of Regression	49
Module 11: Business Applications	50
Learning Goals	50
Structure of This Module	50
23 Marketing	51
24 Finance	52
25 Human Resources	53
26 Operations	54
XII Module 12: Projects and Datasets	55
Module 12: Project	56
Learning Goals	56
Structure of This Module	56
27 Project Guidelines	57
28 Datasets	58

Introduction to Business Statistics

A Modeling and Regression-Based Approach

This text presents a modern, modeling-first approach to business statistics. Rather than a traditional probability-first sequence, we begin with data, relationships, patterns, models, and regression — the tools most relevant to business decision-making in marketing, finance, economics, HR, and analytics.

You will learn:

- how to visualize data
- how to describe relationships
- how to build regression models
- how to interpret uncertainty
- how to make data-driven decisions

Let's get started.

Preface

This book accompanies the course *Introduction to Business Statistics* at Stillman School of Business. It takes inspiration from *Lessons in Statistical Thinking* and builds on two core teaching principles:

1. **Students understand structure before uncertainty.**
2. **Regression is the central statistical tool for business.**

The book is organized into 12 modules reflecting this storyline.

Part I

Module 1: Data, Visualization, and Patterns

Module 1: Data, Visualization & Patterns

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

1 Data Frames

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

2 Point Plots

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

3 Relationships

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part II

Module 2: Best Guess and Means

Module 2: Mean as the best guess

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

4 Means

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

5 Category Means

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

6 Residuals

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part III

Module 3: What Is a Model?

Module 3: Models

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

7 Concept of model

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part IV

Module 4: Regression Without Confidence Bands

Module 4: Linear Regression Models

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

8 Line Models

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

9 Interpreting coefficients

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part V

Module 5: Populations and Samples

Module 5: Population and Samples

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

10 Population vs Sample

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

11 Sampling Variability

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part VI

Module 6: Probability as Variation

Module 6: Probability

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

12 Randomness

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

13 Variability

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

14 Distributions

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part VII

Module 7: Sampling Variation and Estimators

Module 7: Sampling Variation

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

15 SE Estimators

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

16 Sampling Distributions Slopes

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part VIII

Module 8: Confidence Intervals and Bands

Module 8: Confidence Bands

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

17 CI Slope

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

18 CI Bands

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part IX

Module 9: Signal, Noise, and R-Squared

Module 9: Signal and Noise

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

19 Variance Decomposition

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

20 R-Squared

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part X

Module 10: Hypothesis Testing via Regression

Module 10: Hypothesis Testing

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

21 t-tests for Regression

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

22 Interpretation

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part XI

Module 11: Business Applications of Regression

Module 11: Business Applications

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

23 Marketing

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

24 Finance

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

25 Human Resources

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

26 Operations

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

Part XII

Module 12: Projects and Datasets

Module 12: Project

This module introduces the basic building blocks of statistical thinking: data frames, variables, and point plots. We begin with visualization because patterns reveal what models will later formalize.

Learning Goals

- Understand data frames, variables, and instances
- Create and interpret point plots
- Recognize patterns: direction, form, strength

Structure of This Module

- Data frames
- Point plots
- Relationships between variables

27 Project Guidelines

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```

28 Datasets

The mean of a numerical variable is the “best guess” of its value when no other information is available — in the specific sense that it minimizes the sum of squared errors.

```
library(ggplot2) mean(mtcars$mpg)
```