Statistical Modeling: A Fresh Approach

Daniel T Kaplan

E-book version of Second Edition ©2017

Contents

Preface to this electronic version **9**

1 Introduction **11**

Example: Applying to Law School 11

Example: Nitrogen Fixing 11

Example: Sex Discrimination 11

1.1 Models and their Purposes 11

1.2 Observation and Knowledge 11

1.3 The Main Points of this Book 11

Reading Questions 11

2 Data: Cases, Variables, Samples **13**

2.1 Kinds of Variables 13

2.2 Data Frames and the Unit of Analysis 13

2.3 Populations and Samples 13

2.4 Longitudinal and Cross-Sectional Samples 13

3 Describing Variation **15**

3.1 Coverage Intervals 15

Aside: What's Normal? 15

3.2 The Variance and Standard Deviation 15

3.3 Displaying Variation 15

3.4 Shapes of Distributions 15

4 Groupwise Models **17**

4.1 Grand and Group-wise Models 17

4.2 Accounting for Variation 17

Aside: The geometry of partitioning 17

4.3 Group-wise Proportions 17

4.4 What's the Precision? 17

4.5 Misleading Group-wise Models 17

5 Confidence Intervals **19**

5.1 The Sampling Distribution 19

Aside: Precision and Sample Size 19
5.2 The Resampling Distribution & Bootstrapping 19
5.3 Re-sampling 19
5.4 The Re-Sampling Distribution 19
Example: The Precision of Grades 19
5.5 The Confidence Level 19
5.6 Interpreting Confidence Intervals 19
5.7 Confidence Intervals from Census Data 19

6 Language of Models 21
6.1 Models as Functions 22
6.2 Model Functions with Multiple Explanatory Variables 22
6.3 Reading a Model 22
6.4 Choices in Model Design 22
The Data 22
The Response Variable 22
Explanatory Variables 22
6.5 Model Terms 22
The Intercept Term (and no other terms) 22
Intercept and Main Terms 22
Interaction Terms 22
Transformation Terms 22
Aside: Are swimmers slowing down? 22
Main Effects without the Intercept 22
6.6 Notation for Describing Model Design 22

7 Model Formulas and Coefficients 23
7.1 The Linear Model Formula 24
7.2 Linear Models with Multiple Terms 24
Aside: Interpreting Interaction Terms 24
7.3 Formulas with Categorical Variables 24
7.4 Coefficients and Relationships 24
7.5 Model Values and Residuals 24
7.6 Coefficients of Basic Model Designs 24
7.7 Coefficients have Units 24
Aside: Comparing Coefficients 24
7.8 Untangling Explanatory Variables 24
Aside: Interaction terms and partial derivatives 24
7.9 Why Linear Models? 24

8 Fitting Models to Data 25
8.1 The Least Squares Criterion 25
Aside: Why Square Residuals? 25
8.2 Partitioning Variation 25
Aside: Partitioning and the Sum of Squares 25
8.3 The Geometry of Least Squares Fitting 25

8.4 Redundancy 25
Example: Almost redundant 25
Aside: Redundancy – (Almost) Anything Goes 25

9 Correlation and Partitioning of Variance 27
9.1 Properties of R^2 28
Example: Quantifying the capture of variation 28
9.2 Simple Correlation 28
Example: Relationships without Correlation 28
Example: R versus R^2 28
9.3 The Geometry of Correlation 28
9.4 Nested Models 28
Example: R^2 Out of the Headlines 28
9.5 The Geometry of R^2 28

10 Total and Partial Change 29
10.1 Total and Partial Relationships 30
10.2 Example: Covariates and Death 30
Example: Used Car Prices 30
10.3 Models and Partial Relationships 30
Aside: Partial change and partial derivatives 30
10.4 Adjustment 30
10.5 Simpson's Paradox 30
Example: Cancer Rates Increasing? 30
10.6 Explicitly Holding Covariates Constant 30
Example: SAT Scores and School Spending 30
Aside: Divide and Be Conquered! 30
10.7 Adjustment and Truth 30
10.8 The Geometry of Covariates and Adjustment 30
Aside: Interaction terms and partial derivatives 30

11 Modeling Randomness 31
11.1 Describing Pure Randomness 33
Example: Rolling a Die 33
Example: The Chance of Rain 33
Example: Flipping two coins. 33
11.2 Settings for Probability Models 33
11.3 Models of Counts 33
Example: Multiple Coin Flips 33
Example: Houses for Sale 33
The Poisson Model 33
Example: The Rate of Highway Accidents 33
11.4 Common Probability Calculations 33
Example: A Political Poll 33
Example: A Normal Year? 33
11.5 Models of Continuous Outcomes 33

The Uniform Model 33
Example: Equally Likely 33
The Normal Model 33
Example: IQ Test Scores 33
The Log-normal Model 33
Example: High Income 33
The Exponential Model 33
Example: Times between Earthquakes 33
The Normal Distribution 33
Example: IQ Percentiles from Scores. 33
Example: IQ Scores from Percentiles. 33
Example: Bad Luck at Roulette? 33

12 Confidence in Models 35
12.1 The Sampling Distribution & Model Coefficients 36
12.2 Standard Errors and the Regression Report 36
12.3 Confidence Intervals 36
Example: Wage discrimination in trucking? 36
Example: SAT Scores and Spending, revisited 36
12.4 Confidence in Predictions 36
Example: Catastrophe in Grand Forks 36
12.5 A Formula for the Standard Error 36
12.6 Confidence and Collinearity 36
Aside: Redundancy and Multi-collinearity 36
12.7 Confidence and Bias 36

13 The Logic of Hypothesis Testing 37
13.1 Example: Ups and downs in the stock market 38
13.2 An Example of a Hypothesis Test 38
13.3 Inductive and Deductive Reasoning 38
Deductive Reasoning 38
Inductive Reasoning 38
13.4 The Null Hypothesis 38
13.5 The p-value 38
13.6 Rejecting by Mistake 38
13.7 Failing to Reject 38
Aside: Calculating a Power 38
13.8 A Glossary of Hypothesis Testing 38
13.9 Update on Stock Prices 38

14 Hypothesis Testing on Whole Models 39
14.1 The Permutation Test 40
14.2 R^2 and the F Statistic 40
Example: Marriage and Astrology 40
14.3 The ANOVA Report 40
Aside: F and R^2 40

Example: Is height genetically determined? 40
Aside: The shape of F 40
Example: F and Astrology 40
14.4 Interpreting the p-value 40
Multiple Comparisons 40
Example: Multiple Jeopardy 40
Significance vs Substance 40
Example: The Significance of Finger Lengths 40

15 Hypothesis Testing on Parts of Models 41
15.1 The Term-by-Term ANOVA Table 42
15.2 Covariates Soak Up Variance 42
Example: Wages and Race 42
15.3 Measuring the Sum of Squares 42
15.4 ANOVA, Collinearity, and Multi-Collinearity 42
Example: Height and Siblings 42
15.5 Choosing the Order of Terms in ANOVA 42
Example: Wages and Race: Part 2 42
Example: Wages and Race: Part 3 42
Example: Testing Universities 42
15.6 Non-Parametric Statistics 42

16 Models of Yes/No Variables 43
16.1 The 0-1 Encoding 43
16.2 Inference on Logistic Models 43
16.3 Model Probabilities 43
Example: Log-odds ratios of Prostate Cancer 43

17 Causation 45
17.1 Interpreting Models Causally 46
Example: Greenhouse Gases and Global Warming 46
17.2 Causation and Correlation 46
17.3 Hypothetical Causal Networks 46
17.4 Networks and Covariates 46
17.5 Pathways 46
17.6 Pathways and the Choice of Covariates 46
Example: Learning about Learning 46
17.7 Sampling Variables 46
17.8 Disagreements about Networks 46
Example: Sex Discrimination in Salary 46
Update on Global Warming 46

18 Experiment 47
18.1 Experiments 49
18.2 Experimental Variables and Experimental Units 49
Example: Virtues of Doing Surgery while Blind 49

Example: Oops! An accidental correlation! 49
18.3 Choosing levels for the experimental variables 49
18.4 Replication 49
18.5 Experiments vs Observations 49
18.6 Creating Orthogonality 49
Blocking in Experimental Assignment 49
18.7 When Experiments are Impossible 49
18.8 Intent to Treat 49
18.9 Destroying Associations 49
Instrumental Variables 49
Matched Sampling 49
Example: Returning to Campaign Spending 49
18.10Conclusion 49

Preface to this electronic version

Placeholder

Chapter 1

Introduction

Placeholder

Example: Applying to Law School

Example: Nitrogen Fixing

Example: Sex Discrimination

1.1 Models and their Purposes

1.2 Observation and Knowledge

1.3 The Main Points of this Book

Reading Questions

Chapter 2

Data: Cases, Variables, Samples

Placeholder

2.1 Kinds of Variables

2.2 Data Frames and the Unit of Analysis

2.3 Populations and Samples

2.4 Longitudinal and Cross-Sectional Samples

Chapter 3

Describing Variation

Placeholder

3.1 Coverage Intervals

Aside: What's Normal?

3.2 The Variance and Standard Deviation

3.3 Displaying Variation

3.4 Shapes of Distributions

3.4.1 Categorical Variables

3.4.2 Quantifying Categorical Variation

Chapter 4

Groupwise Models

Placeholder

4.1 Grand and Group-wise Models

4.2 Accounting for Variation

Aside: The geometry of partitioning

4.3 Group-wise Proportions

4.4 What's the Precision?

4.5 Misleading Group-wise Models

Chapter 5

Confidence Intervals

Placeholder

5.1 The Sampling Distribution

Aside: Precision and Sample Size

5.2 The Resampling Distribution & Bootstrapping

5.3 Re-sampling

5.4 The Re-Sampling Distribution

Example: The Precision of Grades

5.5 The Confidence Level

5.6 Interpreting Confidence Intervals

5.7 Confidence Intervals from Census Data

Chapter 6

Language of Models

Placeholder

6.1 Models as Functions

6.2 Model Functions with Multiple Explanatory Variables

6.3 Reading a Model

6.4 Choices in Model Design

The Data

The Response Variable

Explanatory Variables

6.5 Model Terms

The Intercept Term (and no other terms)

Intercept and Main Terms

Interaction Terms

Transformation Terms

Aside: Are swimmers slowing down?

Main Effects without the Intercept

6.6 Notation for Describing Model Design

Chapter 7

Model Formulas and Coefficients

Placeholder

7.1 The Linear Model Formula

7.2 Linear Models with Multiple Terms

Aside: Interpreting Interaction Terms

7.3 Formulas with Categorical Variables

7.4 Coefficients and Relationships

7.5 Model Values and Residuals

7.6 Coefficients of Basic Model Designs

7.7 Coefficients have Units

Aside: Comparing Coefficients

7.8 Untangling Explanatory Variables

Aside: Interaction terms and partial derivatives

7.9 Why Linear Models?

Chapter 8

Fitting Models to Data

Placeholder

8.1 The Least Squares Criterion

Aside: Why Square Residuals?

8.2 Partitioning Variation

Aside: Partitioning and the Sum of Squares

8.3 The Geometry of Least Squares Fitting

8.4 Redundancy

Example: Almost redundant

Aside: Redundancy – (Almost) Anything Goes

Chapter 9

Correlation and Partitioning of Variance

Placeholder

9.1 Properties of R²

Example: Quantifying the capture of variation

9.2 Simple Correlation

Example: Relationships without Correlation

Example: R versus R²

9.3 The Geometry of Correlation

9.4 Nested Models

Example: R² Out of the Headlines

9.5 The Geometry of R²

Chapter 10

Total and Partial Change

Placeholder

10.1 Total and Partial Relationships

10.2 Example: Covariates and Death

Example: Used Car Prices

10.3 Models and Partial Relationships

Aside: Partial change and partial derivatives

10.4 Adjustment

10.5 Simpson's Paradox

Example: Cancer Rates Increasing?

10.6 Explicitly Holding Covariates Constant

Example: SAT Scores and School Spending

Aside: Divide and Be Conquered!

10.7 Adjustment and Truth

10.8 The Geometry of Covariates and Adjustment

Aside: Interaction terms and partial derivatives

Chapter 11

Modeling Randomness

Placeholder

11.1 Describing Pure Randomness

Example: Rolling a Die

Example: The Chance of Rain

Example: Flipping two coins.

11.2 Settings for Probability Models

11.3 Models of Counts

The Binomial Model

Example: Multiple Coin Flips

Example: Houses for Sale

The Poisson Model

Example: The Rate of Highway Accidents

11.4 Common Probability Calculations

Example: A Political Poll

Example: A Normal Year?

11.5 Models of Continuous Outcomes

The Uniform Model

Example: Equally Likely

The Normal Model

Example: IQ Test Scores

The Log-normal Model

Chapter 12

Confidence in Models

Placeholder

12.1 The Sampling Distribution & Model Coefficients

12.2 Standard Errors and the Regression Report

12.3 Confidence Intervals

Example: Wage discrimination in trucking?

Example: SAT Scores and Spending, revisited

12.4 Confidence in Predictions

Example: Catastrophe in Grand Forks

12.5 A Formula for the Standard Error

12.6 Confidence and Collinearity

Aside: Redundancy and Multi-collinearity

12.7 Confidence and Bias

Chapter 13

The Logic of Hypothesis Testing

Placeholder

13.1 Example: Ups and downs in the stock market

13.2 An Example of a Hypothesis Test

13.3 Inductive and Deductive Reasoning

Deductive Reasoning

Inductive Reasoning

13.4 The Null Hypothesis

13.5 The p-value

13.6 Rejecting by Mistake

13.7 Failing to Reject

Aside: Calculating a Power

13.8 A Glossary of Hypothesis Testing

13.9 Update on Stock Prices

Chapter 14

Hypothesis Testing on Whole Models

Placeholder

14.1 The Permutation Test

14.2 R^2 and the F Statistic

Example: Marriage and Astrology

14.3 The ANOVA Report

Aside: F and R^2

Example: Is height genetically determined?

Aside: The shape of F

Example: F and Astrology

14.4 Interpreting the p-value

Multiple Comparisons

Example: Multiple Jeopardy

Significance vs Substance

Example: The Significance of Finger Lengths

Chapter 15

Hypothesis Testing on Parts of Models

Placeholder

15.1 The Term-by-Term ANOVA Table

15.2 Covariates Soak Up Variance

Example: Wages and Race

15.3 Measuring the Sum of Squares

15.4 ANOVA, Collinearity, and Multi-Collinearity

Example: Height and Siblings

15.5 Choosing the Order of Terms in ANOVA

Example: Wages and Race: Part 2

Example: Wages and Race: Part 3

Example: Testing Universities

15.6 Non-Parametric Statistics

Chapter 16

Models of Yes/No Variables

Placeholder

16.1 The 0-1 Encoding

16.2 Inference on Logistic Models

16.3 Model Probabilities

Example: Log-odds ratios of Prostate Cancer

Chapter 17

Causation

Placeholder

17.1 Interpreting Models Causally

Example: Greenhouse Gases and Global Warming

17.2 Causation and Correlation

17.3 Hypothetical Causal Networks

17.4 Networks and Covariates

17.5 Pathways

17.6 Pathways and the Choice of Covariates

Example: Learning about Learning

17.7 Sampling Variables

17.8 Disagreements about Networks

Example: Sex Discrimination in Salary

Update on Global Warming

Chapter 18

Experiment

Placeholder

18.1 Experiments

18.2 Experimental Variables and Experimental Units

Example: Virtues of Doing Surgery while Blind

Example: Oops! An accidental correlation!

18.3 Choosing levels for the experimental variables

18.4 Replication

18.5 Experiments vs Observations

18.6 Creating Orthogonality

18.6.1 Random Assignment

Blocking in Experimental Assignment

18.7 When Experiments are Impossible

18.8 Intent to Treat

18.9 Destroying Associations

Instrumental Variables

Matched Sampling

Example: Returning to Campaign Spending ...

18.10 Conclusion