

**JEREMY ARKES**

Copyrighted Material



**SECOND EDITION**

# **REGRESSION ANALYSIS**

**A Practical Introduction**



Copyrighted Material

# **Regression Analysis**

## **A Practical Introduction**

SECOND EDITION

**Jeremy Arkes**

 **Routledge**  
Taylor & Francis Group  
LONDON AND NEW YORK

# Contents

<i>List of figures</i>	xi
<i>List of tables</i>	xiii
<i>About the author</i>	xv
<i>Preface</i>	xvi
<i>Acknowledgments</i>	xviii
<i>List of abbreviations</i>	xix
1 Introduction	1
1.1 The problem	2
1.2 The purpose of research	3
1.3 What causes problems in the research process?	4
1.4 About this book	7
1.5 Quantitative vs. qualitative research	10
1.6 Stata and R code	10
1.7 Chapter summary	11
2 Regression analysis basics	12
2.1 What is a regression?	13
2.2 The four main objectives for regression analysis	15
2.3 The Simple Regression Model	17
2.4 How are regression lines determined?	21
2.5 The explanatory power of the regression	26
2.6 What contributes to slopes of regression lines?	28
2.7 Using residuals to gauge relative performance	30
2.8 Correlation vs. causation	32
2.9 The Multiple Regression Model	33
2.10 Assumptions of regression models	36
2.11 Everyone has their own effect	38
2.12 Causal effects can change over time	39
2.13 Why regression results might be wrong: inaccuracy and imprecision	40

2.14	The use of regression flowcharts	42
2.15	The underlying Linear Algebra in regression equations	43
2.16	Definitions and key concepts	45
2.17	Chapter summary	47
3	Essential tools for regression analysis	51
3.1	Using dummy (binary) variables	51
3.2	Non-linear functional forms using Ordinary Least Squares	54
3.3	Weighted regression models	62
3.4	Calculating standardized coefficient estimates to allow comparisons	63
3.5	Chapter summary	64
4	What does “holding other factors constant” mean?	67
4.1	Why do we want to “hold other factors constant”?	68
4.2	Operative-vs-“held constant” and good-vs-bad variation in a key-explanatory variable	68
4.3	How “holding other factors constant” works when done cleanly	72
4.4	Why is it difficult to “hold a factor constant”?	78
4.5	When you do <i>not</i> want to hold a factor constant	81
4.6	Proper terminology for controlling for a variable	88
4.7	Chapter summary	88
5	Standard errors, hypothesis tests, p-values, and aliens	90
5.1	Standard errors	91
5.2	How the standard error determines the likelihood of various values of the true coefficient	97
5.3	Hypothesis testing in regression analysis	99
5.4	Problems with standard errors (multicollinearity, heteroskedasticity, and clustering) and how to fix them	113
5.5	The Bayesian critique of p-values (and statistical significance)	119
5.6	What model diagnostics should you do?	122
5.7	What the research on the hot hand in basketball tells us about the existence of other life in the universe	123
5.8	What does an insignificant estimate tell you?	124
5.9	Statistical significance is not the goal	126
5.10	Why I believe we should scrap hypothesis tests	127
5.11	Chapter summary	128
6	What could go wrong when estimating causal effects?	132
6.1	Setting up the problem for estimating a causal effect	135
6.2	Good variation vs. bad variation in the key-explanatory variable	137
6.3	An introduction to the PITFALLS	140
6.4	PITFALL #1: Reverse causality	141
6.5	PITFALL #2: Omitted-factors bias	146
6.6	PITFALL #3: Self-selection bias	157

6.7	PITFALL #4: Measurement error	162
6.8	PITFALL #5: Using mediating factors or outcomes as control variables	168
6.9	PITFALL #6: Improper reference groups	176
6.10	PITFALL #7: Over-weighting groups (when using fixed effects or dummy variables)	182
6.11	How to choose the best set of control variables (model selection)	190
6.12	What could affect the validity of the sample?	196
6.13	Applying the PITFALLS to studies on estimating divorce effects on children	198
6.14	Applying the PITFALLS to nutritional studies	200
6.15	Chapter summary	201
7	Strategies for other regression objectives	208
7.1	Strategies and PITFALLS for forecasting/predicting an outcome	209
7.2	Strategies and PITFALLS for determining predictors of an outcome	213
7.3	Strategies and PITFALLS for adjusting outcomes for various factors and anomaly detection	217
7.4	Summary of the strategies and PITFALLS for each regression objective	222
8	Methods to address biases	225
8.1	Fixed effects	227
8.2	Correcting for over-weighted groups (PITFALL #7) using fixed effects	238
8.3	Random effects	240
8.4	First-differences	242
8.5	Difference-in-differences	246
8.6	Two-stage least squares (instrumental-variables)	251
8.7	Regression discontinuities	257
8.8	Knowing when to punt	260
8.9	Summary	261
9	Other methods besides Ordinary Least Squares	266
9.1	Types of outcome variables	267
9.2	Dichotomous outcomes	268
9.3	Ordinal outcomes – ordered models	274
9.4	Categorical outcomes – Multinomial Logit Model	276
9.5	Censored outcomes – Tobit models	279
9.6	Count variables – Negative Binomial and Poisson models	280
9.7	Duration models	282
9.8	Summary	285
10	Time-series models	287
10.1	The components of a time-series variable	288
10.2	Autocorrelation	289
10.3	Autoregressive models	291
10.4	Distributed-lag models	297
10.5	Consequences of and tests for autocorrelation	299
10.6	Stationarity	302

10.7 Vector Autoregression	307
10.8 Forecasting with time series	308
10.9 Summary	313
11 Some really interesting research	315
11.1 Can discrimination be a self-fulfilling prophecy?	315
11.2 Does Medicaid participation improve health outcomes?	321
11.3 Estimating peer effects on academic outcomes	322
11.4 How much does a GED improve labor-market outcomes?	325
11.5 How female integration in the Norwegian military affects gender attitudes among males	327
12 How to conduct a research project	331
12.1 Choosing a topic	332
12.2 Conducting the empirical part of the study	334
12.3 Writing the report	336
13 The ethics of regression analysis	343
13.1 What do we hope <i>to see</i> and <i>not to see</i> in others' research?	344
13.2 The incentives that could lead to unethical practices	344
13.3 P-hacking and other unethical practices	345
13.4 How to be ethical in your research	347
13.5 Examples of how studies could have been improved under the ethical guidelines I describe	349
13.6 Summary	351
14 Summarizing thoughts	352
14.1 Be aware of your cognitive biases	352
14.2 What betrays trust in published studies	354
14.3 How to do a referee report responsibly	359
14.4 Summary of the most important points and interpretations	360
14.5 Final words of wisdom (and one final Yogi quote)	362
Appendix of background statistical tools	364
A.1 Random variables and probability distributions	365
A.2 The normal distribution and other important distributions	371
A.3 Sampling distributions	373
A.4 Desired properties of estimators	377
<i>Glossary</i>	379
<i>Index</i>	389