International Series on Actuarial Science

Regression Modeling with Actuarial and Financial Applications

Edward W. Frees

CAMBRIDGE

CERTUM EX INCERTIS

AD FINEM FIDELIS

Regression Modeling with Actuarial and Financial Applications

EDWARD W. FREES
University of Wisconsin, Madison

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521135962

First published 2010

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Frees, Edward W.
Regression modeling with actuarial and financial applications / Edward W. (Jed) Frees.
p. cm.
Includes index.
ISBN 978-0-521-76011-9 (hardback : alk. paper) 1. Insurance – Statistical methods. 2. Finance –
Statistical methods. 3. Regression analysis. I. Title.
HG8781.F67 2010
519.5′36–dc22 2009032791

ISBN 978-0-521-76011-9 Hardback
ISBN 978-0-521-13596-2 Paperback

Additional resources for this publication at http://research.bus.wisc.edu/RegActuaries

Regression Modeling with Actuarial and Financial Applications

Statistical techniques can be used to address new situations. This is important in a rapidly evolving risk management and financial world. Analysts with a strong statistical background understand that a large data set can represent a treasure trove of information to be mined and can yield a strong competitive advantage.

This book provides budding actuaries and financial analysts with a foundation in multiple regression and time series. Readers will learn about these statistical techniques using data on the demand for insurance, lottery sales, foreign exchange rates, and other applications. Although no specific knowledge of risk management or finance is presumed, the approach introduces applications in which statistical techniques can be used to analyze real data of interest. In addition to the fundamentals, this book describes several advanced statistical topics that are particularly relevant to actuarial and financial practice, including the analysis of longitudinal, two-part (frequency/severity), and fat-tailed data.

Datasets with detailed descriptions, sample statistical software scripts in R and SAS, and tips on writing a statistical report, including sample projects, can be found on the book's Web site: http://research.bus.wisc.edu/RegActuaries.

Christopher Daykin, Independent Consultant and Actuary
Angus Macdonald, Heriot-Watt University

The International Series on Actuarial Science, published by Cambridge University Press in conjunction with the Institute of Actuaries and the Faculty of Actuaries, will contain textbooks for students taking courses in or related to actuarial science, as well as more advanced works designed for continuing professional development or for describing and synthesizing research. The series will be a vehicle for publishing books that reflect changes and developments in the curriculum, that encourage the introduction of courses on actuarial science in universities, and that show how actuarial science can be used in all areas in which there is long-term financial risk.

There is an old saying, attributed to Sir Issac Newton:

"If I have seen far, it is by standing on the shoulders of giants."

I dedicate this book to the memory of two giants who helped me, and everyone who knew them, see farther and live better lives:

James C. Hickman
and
Joseph P. Sullivan.

Preface

Actuaries and other financial analysts quantify situations using data – we are "numbers" people. Many of our approaches and models are stylized, based on years of experience and investigations performed by legions of analysts. However, the financial and risk management world evolves rapidly. Many analysts are confronted with new situations in which tried-and-true methods simply do not work. This is where a toolkit like regression analysis comes in.

Regression is the study of relationships among variables. It is a generic statistics discipline that is not restricted to the financial world – it has applications in the fields of social, biological, and physical sciences. You can use regression techniques to investigate large and complex data sets. To familiarize you with regression, this book explores many examples and data sets based on actuarial and financial applications. This is not to say that you will not encounter applications outside of the financial world (e.g., an actuary may need to understand the latest scientific evidence on genetic testing for underwriting purposes). However, as you become acquainted with this toolkit, you will see how regression can be applied in many (and sometimes new) situations.

Who Is This Book For?

This book is written for financial analysts who face uncertain events and wish to quantify the events using empirical information. No industry knowledge is assumed, although readers will find the reading much easier if they have an interest in the applications discussed here! This book is designed for students who are just being introduced to the field as well as industry analysts who would like to brush up on old techniques and (for the later chapters) get an introduction to new developments.

To read this book, I assume knowledge comparable to a one-semester introduction to probability and statistics – Appendix A1 provides a brief review to brush up if you are rusty. Actuarial students in North America will have a one-year introduction to probability and statistics – this type of introduction will help readers grasp concepts more quickly than a one-semester background. Finally, readers will find matrix, or linear, algebra helpful though not a prerequisite for reading this text.

Different readers are interested in understanding statistics at different levels. This book is written to accommodate the "armchair actuary," that is, one who passively reads and does not get involved by attempting the exercises in the text. Consider an analogy to football or any other game. Just like the armchair quarterback of football, there is a great deal that you can learn about the game just by watching. However, if you want to sharpen your skills, you have to go out and play the game. If you do the exercises or reproduce the statistical analyses in the text, you will become a better player. Still, this text interweaves examples with the basic principles. Thus, even the armchair actuary can obtain a solid understanding of regression techniques through this text.

What Is This Book About?

The table of contents provides an overview of the topics covered, which are organized into four parts. The first part introduces linear regression. This is the core material of the book, with refreshers on mathematical statistics, distributions, and matrix algebra woven in as needed.

The second part is devoted to topics in time series. Why integrate time series topics into a regression book? The reasons are simple, yet compelling: most accounting, financial, and economic data become available over time. Although cross-sectional inferences are useful, business decisions need to be made in real time with currently available data. Chapters 7–10 introduce time series techniques that can be readily accomplished using regression tools (and there are many).

Nonlinear regression is the subject of the third part. Many modern-day predictive modeling tools are based on nonlinear regression – these are the workhorses of statistical shops in the financial and risk management industry.

The fourth part concerns actuarial applications, topics that I have found relevant in my research and consulting work in financial risk management. The first four chapters of this part consist of variations of regression models that are particularly useful in risk management. The last two chapters focus on communications, specifically report writing and designing graphs. Communicating information is an important aspect of every technical discipline, and statistics is certainly no exception.

How Does This Book Deliver Its Message?

Chapter Development

Each chapter has several examples interwoven with theory. In chapters in which a model is introduced, I begin with an example and discuss the data analysis without regard to the theory. This analysis is presented at an intuitive level, without reference to a specific model. This is straightforward, because it amounts to little more than curve fitting. The goal is to have students summarize data sensibly without having the notion of a model obscure good data analysis. Then, an introduction to the theory is provided in the context of the introductory example. One or more additional examples follow that reinforce the theory already introduced and provide a context for explaining additional theory. In Chapters 5 and 6, which do introduce not models but rather techniques for analysis, I begin with an introduction of the technique. This introduction is then followed by an example that reinforces the explanation. In this way, the data analysis can be easily omitted without loss of continuity, if time is a concern.

Real Data

Many of the exercises ask the reader to work with real data. The need for working with real data is well documented; for example, see Hogg (1972) or Singer and Willett (1990). Some criteria of Singer and Willett for judging a good data set include authenticity, availability of background information, interest and relevance to substantive learning, and availability of elements with which readers can identify. Of course, there are some important disadvantages to working with real data. Data sets can quickly become outdated. Further, the ideal data set for illustrating a specific statistical issue is difficult to find. This is because with real data, almost by definition, several issues occur simultaneously. This makes it difficult to isolate a specific aspect. I particularly enjoy working with large data sets. The larger the data set, the greater the need for statistics to summarize the information content.

The larger the data set, the greater the need for statistics to summarize the information content.

Statistical Software and Data

My goal in writing this text is to reach a broad group of students and industry analysts. Thus, to avoid excluding large segments, I chose not to integrate any specific statistical software package into the text. Nonetheless, because of the applications orientation, it is critical that the methodology presented be easily accomplished using readily available packages. For the course taught at the University of Wisconsin, I use the statistical packages SAS and R. On the book's Web site, at

http://research.bus.wisc.edu/RegActuaries,

users will find scripts written in SAS and R for the analyses presented in the text. The data are available in text format, allowing readers to employ any statistical packages that they wish. When you see a display such as this in the margin, you will also be able to find this data set (*TermLife*) on the book's Web site.

Ⓡ EMPIRICAL
Filename is "TermLife"

Technical Supplements

The technical supplements reinforce and extend the results in the main body of the text by giving a more formal, mathematical treatment of the material. This treatment is, in fact, a supplement because the applications and examples are described in the main body of the text. For readers with sufficient mathematical background, the supplements provide additional material that is useful in communicating to technical audiences. The technical supplements provide a deeper, and broader, coverage of applied regression analysis.

I believe that analysts should have an idea of "what is going on under the hood," or "how the engine works." Most of these topics will be omitted from the first reading of the material. However, as you work with regression, you will be confronted with questions such as, "Why?" and you will need to get into the details to see exactly how a certain technique works. Further, the technical supplements provide a menu of optional items that an instructor may wish to cover.

Suggested Courses

There is a wide variety of topics that can go into a regression course. Here are some suggested courses. The course that I teach at the University of Wisconsin is the first on the list in the following table.

Audience	Nature of Course	Suggested Chapters
One-year background in probability and statistics	Survey of regression and time series models	Chapters 1–8, 11–13, 20–21, main body of text only
One-year background in probability and statistics	Regression and time series models	Chapters 1–8, 20–21, selected portions of technical supplements
One-year background in probability and statistics	Regression modeling	Chapters 1–6, 11–13, 20–21, selected portions of technical supplements
Background in statistics and linear regression	Actuarial regression models	Chapters 10–21, selected portions of technical supplements

In addition to the previously suggested courses, this book is designed as supplemental reading for a time series course as well as a reference book for industry analysts. My hope is that college students who use the beginning parts of the book in their university courses will find the later chapters helpful in their industry positions. In this way I hope to promote lifelong learning!

Acknowledgments

It is appropriate to begin the acknowledgment section by thanking the students in the actuarial program here at the University of Wisconsin; students are important partners in the knowledge creation and dissemination business at universities. Through their questions and feedback, I have learned a tremendous amount over the years. I have also benefited from excellent assistance from those who have helped me pull together all the pieces for this book, specifically Missy Pinney, Peng Shi, Yunjie (Winnie) Sun, and Ziyan Xie.

I have enjoyed working with several former students and colleagues on regression problems in recent years, including Katrien Antonio, Jie Gao, Paul Johnson, Margie Rosenberg, Jiafeng Sun, Emil Valdez, and Ping Wang. Their contributions are reflected indirectly throughout the text. Because of my long association with the University of Wisconsin–Madison, I am reluctant to go back further in time and provide a longer list, for fear of missing important individuals. I have also been fortunate to have a more recent association with the Insurance Services Office (ISO). Colleagues at ISO have provided me with important insights into applications.

Through this text that features applications of regression into actuarial and financial industry problems, I hope to encourage the fostering of additional partnerships between academia and industry.

I am pleased to acknowledge detailed reviews that I received from my colleagues Tim Welnetz and Margie Rosenberg. I also wish to thank Bob Miller for permission to include our joint work on designing effective graphs in Chapter 21. Bob has taught me a lot about regression over the years.

Moreover, I am happy to acknowledge financial support through the Assurant Health Professorship in Actuarial Science at the University of Wisconsin–Madison.

Saving the most important for last, I thank my family for their support. Ten thousand thanks to my mother, Mary; my brothers Randy, Guy, and Joe; my wife, Deirdre; and our sons, Nathan and Adam.

Contents

1

Regression and the Normal Distribution

Chapter Preview. Regression analysis is a statistical method that is widely used in many fields of study, with actuarial science being no exception. This chapter provides an introduction to the role of the normal distribution in regression, the use of logarithmic transformations in specifying regression relationships, and the sampling basis that is critical for inferring regression results to broad populations of interest.

1.1 What Is Regression Analysis?

Statistics is about data. As a discipline, it is about the collection, summarization, and analysis of data to make statements about the real world. When analysts collect data, they are really collecting information that is quantified, that is, transformed to a numerical scale. There are easy, well-understood rules for reducing the data, through either numerical or graphical summary measures. These summary measures can then be linked to a theoretical representation, or model, of the data. With a model that is calibrated by data, statements about the world can be made.

Statistics is about the collection, summarization, and analysis of data to make statements about the real world.

Statistical methods have had a major impact on several fields of study:

- In the area of data collection, the careful design of *sample surveys* is crucial to market research groups and to the auditing procedures of accounting firms.
- *Experimental design* is a subdiscipline devoted to data collection. The focus of experimental design is on constructing methods of data collection that will extract information in the most efficient way possible. This is especially important in fields such as agriculture and engineering where each observation is expensive, possibly costing millions of dollars.
- Other applied statistical methods focus on managing and predicting data. *Process control* deals with monitoring a process over time and deciding when intervention is most fruitful. Process control helps manage the quality of goods produced by manufacturers.
- *Forecasting* is about extrapolating a process into the future, whether it be sales of a product or movements of an interest rate.

Regression analysis is a statistical method used to analyze data. As we will see, the distinguishing feature of this method is the ability to make statements about