Ezra Hauer

The Art of Regression Modeling in Road Safety

Springer

The Art of Regression Modeling in Road Safety

Ezra Hauer

The Art of Regression Modeling in Road Safety

Springer

Ezra Hauer
Professor Emeritus
University of Toronto
Toronto, ON, Canada

Additional material to this book can be downloaded from http://extras.springer.com

ISBN 978-3-319-12528-2 ISBN 978-3-319-12529-9 (eBook)
DOI 10.1007/978-3-319-12529-9
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014953294

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Gordon Newell, mentor, in gratitude.

Preface

This book has two objectives. The first is to teach how to fit a multivariable statistical model to cross-sectional safety data using a simple spreadsheet. The second is to promote the understanding that is at the core of good modeling.

These twin objectives determine the flow and the structure of the book. After some preliminaries, a real data set is introduced. From here on and all the way to the last chapter, the same data are used to gradually build up a regression model. Along the way there are the "what and how" sections: what is an exploratory data analysis, how to use pivot tables, what is a curve-fitting spreadsheet and how to build one, how to use the Solver for parameter estimation, how to examine the quality of a fit by a CURE plot, what functions look like, etc. These are in support of the first objective, that of teaching how to fit a model to data. Interspersed between these are sections of a reflective nature. These support the "understanding" objective and speak about the "why, when, and whether" issues of modeling: why do we need to curve-fit, whether models be used in a cause–effect manner, when should a variable be added to the model equation, why it is important to know what function hides behind the data, etc.

Data about accidents, about the road, and about the traffic on it are routinely collected and maintained in databases. We know where reported accidents occurred and much about their circumstances. We also have information about many features of the road (grade, curve radius, lane width, speed limit, parking control, etc.) and of its traffic (daily volumes, percent of trucks, etc.). Inasmuch as these data pertain to what is observed to exist on a cross section of "units" (road segments, intersections, etc.) they are called "observational cross-sectional" data.

To be of use for evidence-based road safety management, the data need to be summarized and cast into the form of statistical models. Models serve three purposes:

1. To estimate how (un)safe are certain road segments, intersections, ramps, crossings, etc. thereby determining the size of the road safety problem that could perhaps be altered by interventions and design changes.

2. To estimate by how much safety has changed following an intervention or a design change.
3. To estimate by how much safety might be altered due to a change in some variable value.

The book focuses on the development of regression models for purposes 1 and 2. While the use of regression models for purpose 3 is commonplace, the trustworthiness of the result is in doubt. Still, it is possible that even this vexed purpose will be well served by the approach to modeling that is advocated here.

Who will use this book? Perhaps they will be graduate students with interest in road safety, perhaps professionals with responsibilities in data analysis, or perhaps others. I do not know how much math, probability, and statistics I can rely on. Some such knowledge is required by the very nature of the subject matter. There will be parts which some readers may find taxing. The hope is that judicious skipping will make the book accessible and useful for a variety of audiences. I tried to make the narration succinct; diversions, elaborations, and detail were relegated to footnotes. A glossary is provided to serve as a refresher of notation and acronyms; the index at the end is for finding the page on which a topic or concept is first introduced.

The book evolved from lecture notes used in a series of hands-on workshops and is richly laced with data-based illustrations, tables, and figures. The supporting materials are available for downloading. To access and download them, go to http://extras.springer.com/ and enter the ISBN of this book. The ISBN (International Standard Book Number) is found just after the title page. The information is in five folders: Data, PowerPoint presentations, Problems, Solutions, and Spreadsheets. Together with the book these materials will be of interest to the reader, student, and instructor.

The book, of course, can be read as is. However, unlike textbooks in the past, it can also be used actively and creatively. The reader is invited to use the downloadable materials to see how results were obtained; to modify, expand, and enrich the analysis; and to use the same spreadsheets with other data.

Modeling in this book is built around the use of Excel spreadsheets and readers are assumed to have facility in their use. Information about less commonly known spreadsheet functionalities is provided where needed. Of course there are specialized and sophisticated statistical software packages which, once acquired and mastered, will do a good, perhaps a superior, job of model development. However, I find that the spreadsheet provides all the essentials; it makes for intimate contact with the data, it has adequate and flexible visualization, the "pivot tables" serve for exploration, an optimization tool does the curve-fitting, and it is a hospitable environment for writing custom pieces of code.

Model development is often presented as if it was a nearly algorithmic sequence of steps, an ordered progression of activities from "Start" to "End." In my opinion such an approach tends to produce inferior results. It is better to think of model development as detective work with clues embedded in data. Like in a game of snakes and ladders, there are advances and setbacks whereby the modeler gradually moves towards a satisfactory outcome. Such work is well served by the atmosphere

of a spreadsheet. For all these reasons, the spreadsheet is my environment of choice for both instruction and creative modeling.

The modeling approach described in these pages may be thought old-fashioned, perhaps unsophisticated. Emphasis is on what is of essence. Papers describing novel modeling ideas and newfangled statistical techniques are being published daily and I make no attempt to capture the latest. In defense I only say that the quality of a meal depends more on the skill of the cook and the time spent on its preparation than on the modernity of the food processor. As has been said: "... second-rate minds grappling with first-rate problems can teach you more than first-rate minds lost in the shrubbery." (Lilla, 2013, p. xii).

Toronto, ON, Canada Ezra Hauer

Reference

Lilla M in Foreword for Berlin I (2013) Against the current, 2nd ed. Princeton University Press, Princeton

Contents

Glossary

Notational Conventions

$\sigma\{.\}$	Standard deviation of what the dot stands for
^	A caret stands for "estimate of" the letter below
$E\{.\}$	Expected value of what the dot stands for
ln	Natural logarithm
$P(K=k)$ or $P(k)$	Probability that the random variable K takes on the value k

Acronyms

AADT	Annual average daily traffic
AIC	Akaike information criterion
BIC	Bayesian information criterion
C-F	Curve-fitting
CMF	Crash modification factor or function
EDA	Exploratory data analysis
F&I	Fatal and injury
ML	Maximum likelihood
NB	Negative binomial
NM	Negative multinomial
N-W	Nadaraya-Watson
OVB	Omitted variable bias
pdf	Probability density function
PDF	Probability distribution function
PDO	Property damage only
RHR	Roadside hazard rating
SD	Squared differences (residuals)
SPF	Safety performance function

SSD	Sum of squared differences (residuals)
TAB	Total accumulated bias
VBA	Visual basic for applications
VIEDA	Variable introduction exploratory data analysis
vpd	Vehicles per day

Greek

α, β	Parameters
β_0	Scale parameter
β_i	Regression parameters $i = 1, 2, 3, \ldots$
ε_{f,X_i}	Elasticity of function "f" with respect to change in variable X_i
θ	Gamma distributed random variable with mean = 1 and variance = $1/b$
λ	Mean number of reported accidents per unit of time
μ_i	Expected number of accidents of unit i
$\sigma\{\mu\}$	Standard deviation of the μ's in a population of units
$\hat{\sigma}^2(i)$	Sum of sorted squared residuals from 1 to i
$\pm\hat{\sigma}'(i)$	Standard error for CURE plot at index = i

Latin

a, b	Parameters of the gamma and the negative binomial distributions
$A, B, \ldots$	Traits
c	Vehicle "concentration"—the (average) number of vehicles/(lane-km)
$E\{\mu\}$	Expected value of the μ's in a population of units
f	The expected number of reported accidents per lane per second
$f()$	The function linking predictor variables and parameters
h	Bandwidth in the N-W nonparametric regression
$\overline{h}$	Average headway
i	Counter (index) of units
j	Counter (index) of time periods
$K(.)$	Kernel function in the Nadaraya-Watson nonparametric regression
K, k	Count of accidents and a certain value of that count
$\mathcal{L}(.)$	Likelihood. The dot in the parenthesis is a placeholder for parameters.
L_i	Length of road segment i
$\ln(\mathcal{L}^*)$	Abridged log-likelihood
$_m$	As subscript denotes "multivehicle"

N	Number of bins or levels of a trait
n	Number of units or, occasionally, of accident counts
NB	Negative binomial
NM	Negative multinomial
p	Probability of a vehicle to be in a crash in the next second
$P(.)$	Probability of the event that the dot stands for
r	The probability of the crash to be reported
s	Standard error or, occasionally, the sum of accident counts
$_s$	As subscript denotes "single-vehicle"
T	Duration of a time period
$V\{.\}$	Variance of the random variable that the dot stands for
$\overline{v}$	Average speed
v_0	Average free-flow speed
v_b	Average speed at which a bottleneck begins to form
$X_1, X_2, \ldots$	Variables in model equation
Δt	Small time interval

What Is What

1

Abstract

Statistical models that use data to express the safety of populations of units (road segments, intersections, grade crossings, etc.) as a function their traits (traffic, geometry, operation) are nowadays called Safety Performance Functions; SPFs for short. To make this notion precise one has to say what is meant by "safety," "unit," "population," and "trait." Most importantly, one has to be clear about what exactly is the information that a SPF provides and what are its practical uses. These are then illustrated by a series of examples. The practical uses of SPFs call for an approach to modeling which is somewhat different from what is usually done.

1.1 Units and Their Safety Property

What should one call the safety of Bobcaygeon Road between the Scotch Line and Plantation Roads in Ontario or the safety of the intersection of Eglinton and Don Mills roads in Toronto? These questions refer to safety as a property of some elements of the real world to be called "units." A road segment, an intersection, a vehicle, or a person is a "unit." A key feature of a "unit" is that it may be involved in accidents[1] (crashes) or that crashes (accidents) may occur on it. While the count of

[1] Those who prefer to use "crash" instead think that "accident" has connotations of being unavoidable, without cause, and thus unpreventable. This, they fear, might weaken the resolve to reduce crashes and their harm. Since the very purpose of studying road safety is to assist in the task of managing the frequency and severity of accidents, such an interpretation makes no sense in this book. In this book "crash" and "accident" describe the same event. One reason for not shunning the term "accident" is that it is the common currency in the community of transportation professionals. Another reason is that the word "accident" provides the proper associations for the randomness inherent in accident counts. The editor of the Canadian Oxford Dictionary (Barber 1999) says that: "No dictionary that I know of uses the word "unpreventable" in any of its definitions of the word "accident." Were they to do so, the definition would be inaccurate and

E. Hauer, *The Art of Regression Modeling in Road Safety*,
DOI 10.1007/978-3-319-12529-9_1

accidents is an indicator of the safety of a unit, it is not identical to it. The safety property of a unit is defined to be *the number of accidents by type and severity, expected to occur on or to it in a specified period of time*.[2]

Two elements of this definition need clarification. First, the word "expected" usually corresponds to "average in the long run." This works well for the idealized devices used in the instruction of probability theory: coin tosses, decks of unmarked cards, urns with balls, and fair dies. It works well because all these devices can be plausibly assumed not to change with repeated use. In contrast, the safety of real units changes with time.

Accordingly, one has to interpret term "expected" by a conditional and counterfactual statement as "what the limit of the long-term average would be if[3] it was possible to freeze all the safety-related traits and circumstances of the unit."

The second element in the definition of safety which requires clarification is the phrase "by type and severity." This means that, generally, the safety property of a unit is not a single number. Thus, for example, the safety of the intersection of High and Main streets in the 2-year period 2012 and 2013 could be described by the array in Table 1.1.

From such an array one can obtain the row sums (5.00 rear-end, 2.40 angle, 0.42 single-vehicle, and 0.08 pedestrian accidents), column sums (4.80 PDO, 2.75 injury, and 0.35 fatal accident), and the total (7.90 accidents). Each accident may involve one or more drivers and their vehicles. Therefore for multi-vehicle categories such as rear-end and angle accidents the additional categorization by number of involved vehicles may be needed.[4]

---

not reflect the actual usage of the word. The defining terms that dominate are "unexpected," "unforeseen," "unintentional," and "undesirable . . . ." Most people recognize that the things we refer to as "accidents" do indeed have causes, whether it be an unplanned pregnancy, slipping on a banana peel or the dog peeing on the rug and that accidents are preventable." In some cases, such as when one speaks of Crash Modification Functions (CMFs) the use of "crash" is so well established that its use already seems natural. However, the exclusive use of "crash" is often a sign of advocacy, an impression I want to avoid. This is why both terms will be used interchangeably.

[2] For a more detailed discussion of how safety is defined see Chap. 3 in Hauer (1997).

[3] As if the notion of "average in the long run" was not enough of an obstacle, the "would be if" phrase further distances the safety property from what is observable. Not only is safety not the same as the number of accidents, it is now something that can be imagined and perhaps estimated but can never be observed.

[4] Alternatively, one may want to speak not of the number of accidents but of the number of accident-involved vehicles or drivers. Similarly, an injury accident is one in which one or more persons were injured. Thus one may wish to describe the safety of a unit not by the number of injury accidents but by the number of injured persons. To convert from "number of accidents" to "number of involved vehicles or drivers" one needs to know the ratio "involved vehicles/accident." For the US from 1988 to 2005 the ratio remained steady at 1.724 ± 0.015. To convert from "fatal accidents" to "persons killed" knowledge of the ratio "persons killed/fatal accident" is needed. For the US from 1988 to 2005 the ratio for fatalities/fatal crash was 1.113 ± 0.004 which is similar to that in Michigan (1.104 ± 0.013 and 1.114 ± 0.019 for those who had been drinking), in Ohio (1.088 ± 0.013), Wisconsin (1.130 ± 0.022). It is also similar to the ratio pedestrian fatalities/Pedestrian Fatal accidents which in Michigan was 1.080 ± 0.023.