Regression Models for Epidemiology

Contents

Preface

This web-book is derived from my lecture slides for Epidemiology 204: "Quantitative Epidemiology III: Statistical Models", at UC Davis.

I have drawn these materials from many sources, including but not limited to:

- David Rocke[1]'s materials from the 2021 edition of this course[2]
- Hua Zhou[3]'s materials from the 2020 edition of Biostat 200C at UCLA[4]
- Vittinghoff et al. (2012)
- Dobson and Barnett (2018)
- Harrell (2015)

> ❗ Important
>
> I do not claim any of this content as my own original intellectual work. I have attempted to provide more detailed disclaimers for specific sections that are heavily derivative of, or even copied directly from, external sources.
> Please see also the list of contributors on GitHub: https://github.com/d-morrison/rme/graphs/contributors

Using these lecture notes

This website provides lecture notes for Epidemiology 204: Quantitative Epidemiology III (Statistical Models) at UC Davis.

The notes are available online at https://d-morrison.github.io/rme/ and are searchable and continuously updated[5].

[1] https://dmrocke.ucdavis.edu/

[2] https://dmrocke.ucdavis.edu/Class/EPI204-Spring-2021/EPI204-Spring-2021.html

[3] https://hua-zhou.github.io/

[4] https://ucla-biostat-200c-2020spring.github.io/schedule/schedule.html

[5] see the source file repository for recent changes: https://github.com/d-morrison/rme

Multiple Format Options

Each chapter is available in three formats:

1. **HTML (Website)**: Browse chapters online with navigation and search
2. **RevealJS Slides**: Presentation slides for teaching (e.g., `logistic-regression-slides.html`)
3. **PDF Handouts**: Printable documents for each chapter (e.g., `logistic-regression-handout.pdf`)

Compiling chapters locally

To compile chapters from source:

1. Install Quarto[6]

2. Clone the repository:

```
git clone --recurse-submodules https://github.com/d-morrison/rme.git
cd rme
```

3. Install R package dependencies:

```
library(devtools)
devtools::install_deps()
```

4. Render in your desired format:

HTML website:

```
quarto render --profile=website
```

RevealJS slides:

```
quarto render logistic-regression.qmd --profile=revealjs
# Or render all slides:
quarto render --profile=revealjs
```

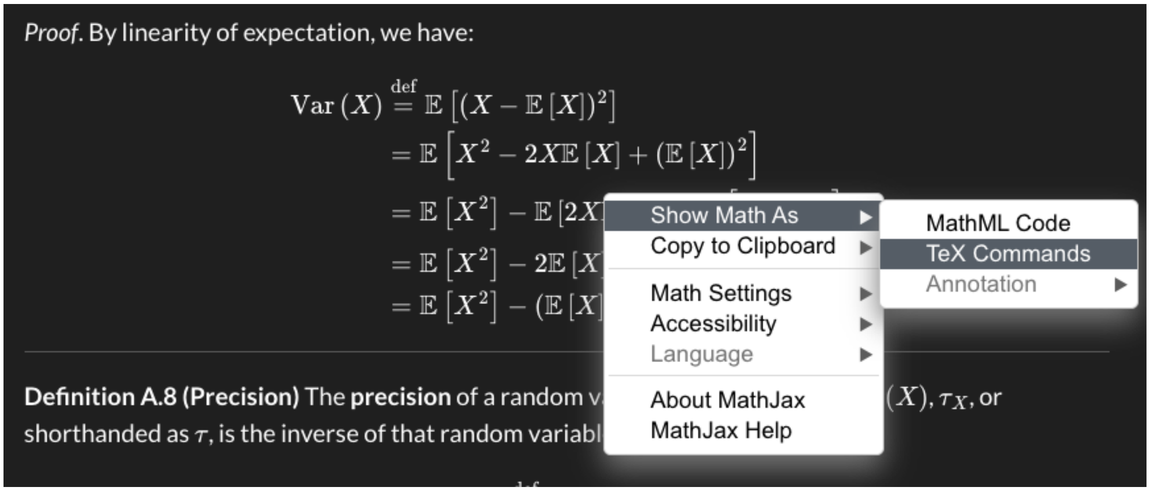
PDF handouts:

```
quarto render logistic-regression.qmd --profile=handout
# Or render all handouts:
quarto render --profile=handout
```

Extracting LaTeX commands from the online version of the notes

If you want to extract the LaTeX commands for any math expressions in the online lecture notes, you should be able to right-click and get this pop-up menu:

[6] https://quarto.org/docs/get-started/



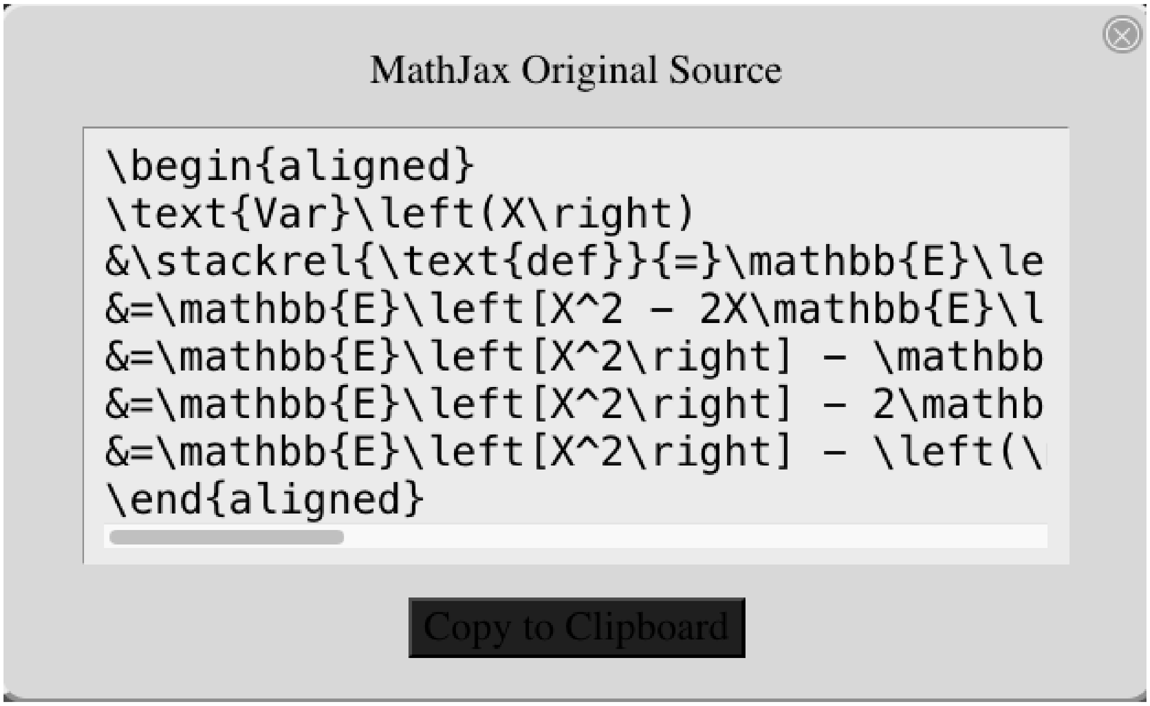
Proof. By linearity of expectation, we have:

$$\begin{aligned}\text{Var}\left(X\right) &\overset{\text{def}}{=} \mathbb{E}\left[(X-\mathbb{E}\left[X\right])^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}\left[X\right] + (\mathbb{E}\left[X\right])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[2X\right. \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}\left[X\right. \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right.\end{aligned}$$

Definition A.8 (Precision) The **precision** of a random v (X), τ_X, or shorthanded as τ, is the inverse of that random variabl

Figure 1: Pop-up menu produced by right-clicking on math in online notes

If you select "TeX commands", you will get a window with LaTeX code.[7]



```
\begin{aligned}
\text{Var}\left(X\right)
&\stackrel{\text{def}}{=}\mathbb{E}\le
&=\mathbb{E}\left[X^2 - 2X\mathbb{E}\l
&=\mathbb{E}\left[X^2\right] - \mathbb
&=\mathbb{E}\left[X^2\right] - 2\mathb
&=\mathbb{E}\left[X^2\right] - \left(\
\end{aligned}
```

Figure 2: LaTeX source code window

You can also grab the TeX commands from the quarto source files on github, but those files use custom macros (defined in https://github.com/d-morrison/rme/blob/main/macros.qmd), so it's a little harder to reuse code from the source files.

Dark Mode

The online notes have two color palette themes: light and dark. You can toggle between them using the oval button near the top-left corner:

[7] MathJax[8] is more or less a dialect of LaTeX

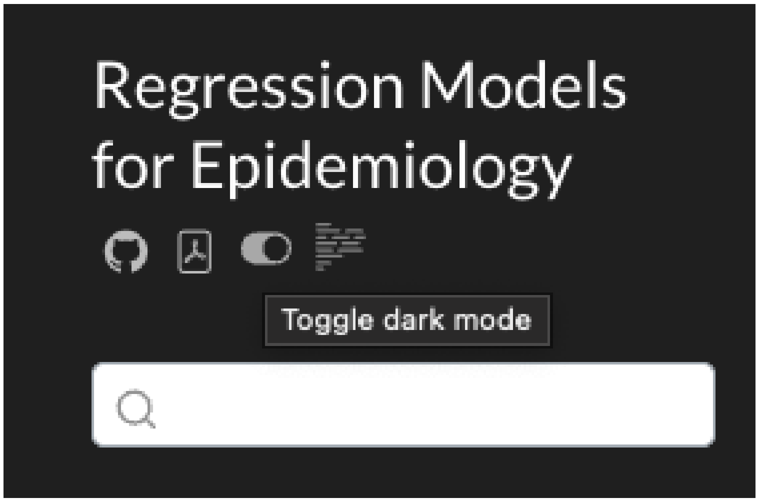


Figure 3: Palette toggle

Other resources

These notes represent my still-developing perspective on regression models in epidemiology. Many other statisticians and epidemiologists have published their own perspectives, and I encourage you to explore your many options and find ones that resonate with you. I have attempted to cite my sources throughout these notes.

Here are some additional resources that I've come across; I haven't had time to read some of them thoroughly yet, but they're all on my to-do list. I'll add my thoughts on them over time.

- Dobson and Barnett (2018) is a classic textbook on GLMs. It was used in UCLA Biostatistics's MS-level GLMs course (Biostat 200C) when I took it, and it helped me a lot. It is fairly mathematically rigorous and concise, bordering on terse. It covers GLMs in detail, and survival analysis briefly, and it also has helpful chapters on Bayesian methods. I have adapted examples and explanations from it extensively in these notes.
- Wakefield (2013) covers GLMs and hierarchical models using both Bayesian and frequentist inference;
  - statistics PhD level
  - author: UW biostatistics professor Jon Wakefield[9]
  - used in UCLA Biostat 250C[10]
- Hosmer, Lemeshow, and Sturdivant (2013) is a classic text on logistic regression. I haven't read it yet.
- Agresti (2012) is another classic text for GLMs. I haven't read it yet.
- Agresti (2018) appears to be a more applied version of Agresti (2012). I haven't read it yet. There are extra exercises[11] and other resources available on the Student Companion Site[12]
- Agresti (2015) has "More than 400 exercises for readers to practice and extend the theory, methods, and data analysis"; might be more theoretical?
- Agresti (2010) is specifically about ordinal data.
- Dunn and Smyth (2018) is a recent textbook on GLMs. It doesn't cover time-to-event models, and it doesn't use the modern `tidyverse`[13] packages (`ggplot2`[14], `dplyr`[15], etc.), but otherwise

[9] https://www.biostat.washington.edu/people/jon-wakefield
[10] https://donatello-telesca.com/biostatistics-251-
[11] https://bcs.wiley.com/he-bcs/Books?action=resource&bcsId=11293&itemId=1119405262&resourceId=44770
[12] https://bcs.wiley.com/he-bcs/Books?action=index&itemId=1119405262&bcsId=11293
[13] https://tidyverse.org/
[14] https://ggplot2.tidyverse.org/
[15] https://dplyr.tidyverse.org/

it seems great. Edelmann (2019) reviews this book formally.

- Moore (2016) is a recent textbook on survival analysis. It also doesn't use the `tidyverse`, but otherwise seems great.
- Klein and Moeschberger (2003) is a classic text for survival analysis. I read most of it in grad school, and it was very helpful. Examples and explanations from it are borrowed extensively in the second half of these notes (partially filtered through David Rocke's course notes.)
- Kalbfleisch and Prentice (2011) is another classic survival analysis text; I haven't read it yet.
- David G. Kleinbaum and Klein (2010) is a mostly applied-level "self-learning" text for logistic regression; I read it cover-to-cover before grad school, and found it very helpful.
- David G. Kleinbaum and Klein (2012) is the corresponding "self-learning" text for survival analysis; I read it cover-to-cover before grad school, and found it very helpful.
- David G. Kleinbaum, Kupper, and Morgenstern (1982), by the same authors, has a solutions manual (David G. Kleinbaum, Kupper, and Morgenstern (1983))
- David G. Kleinbaum et al. (2014) is also by the same group, in a similar style
- Harrell (2015) is another popular textbook. It uses `ggplot2`[16] but not `dplyr`[17], and covers logistic regression and survival analysis (no Poisson or NB models?). An abbreviated but continuously updated version with audio clips is available at https://hbiostat.org/rmsc/.
- Fox (2015) is another standard text. [18]
- McCullagh and Nelder (1989) is a classic, theoretical textbook on GLMs [19]
- Dalgaard (2008) covers GLMs and survival analysis at an applied level, using base R
- Vittinghoff et al. (2012) covers GLMs, survival analysis, and causal inference, using Stata. The authors are UCSF professors, and it is used for the core Epi PhD courses there. I read this book nearly cover-to-cover before grad school, and it was hugely helpful for me, both for statistical modeling and for causal inference (I think it provided my first exposure to DAGs).
- McCulloch, Searle, and Neuhaus (2008) is also by UCSF professors
- Faraway (2016) has GLMs but not survival analysis
- Selvin (2001) provides worked-out examples of applications for a wide range of statistical analysis techniques. The Author[20] is a retired UC Berkeley Biostatistics professor; he used it in a graduate-level biostat/epi course.
- Selvin (2004) is by the same author
  - recommended by Jewell (2003) for Poisson regression
- Jewell (2003) is by another UC Berkeley professor[21]; it mostly covers logistic regression, with one chapter on survival analysis.
- https://ucla-biostat-200c-2020spring.github.io/schedule/schedule.html provides course notes for "Biostat 200C - Methods in Biostatistics C" at UCLA, which is at the Biostatistics MS level.
- https://online.stat.psu.edu/stat504/book/ provides course notes for "STAT 504 - Analysis of Discrete Data" at Penn State University. It includes logistic regression and Poisson regression, as well as 2-way tables and other related topics, and includes SAS code.
- Nahhas (2024) is currently in-development

[16] https://ggplot2.tidyverse.org/

[17] https://dplyr.tidyverse.org/

[18] I don't have anything to say about this book, because I haven't opened it yet, but I've heard it's great!

[19] haven't opened it either

[20] https://publichealth.berkeley.edu/people/steve-selvin

[21] https://publichealth.berkeley.edu/people/nicholas-jewell

- Clayton and Hills (2013) covers binary regression, count regression, and survival analysis. Haven't started it yet.
- https://thomaselove.github.io/2020-432-book/index.html is another set of lecture notes.
- Woodward (2013) covers GLMs and survival; haven't read it yet, but it looks comprehensive.
- Roback and Legler (2021) is recent and uses the `tidyverse`; doesn't appear to cover survival analysis.
- Wood (2017) is about generalized *additive* models but includes a detailed summary of GLMs.
- Kutoyants (2023) appears to be a complete book on Poisson models.
- Hardin and Hilbe (2018) uses Stata.
- Andrews and Herzberg (2012) is a classic "learn-by-example" book with many datasets amenable to GLMs
- Cannell and Livingston (2024) is another open-source, online textbook like this one; it is primarily about statistical programming, but it includes full chapters on linear regression[22], logistic regression[23], and Poisson regression[24]. There is currently (2024/06) a placeholder chapter for survival analysis[25].
- Gelman and Hill (2007) covers GLMs as well as hierarchical extensions of GLMs. No survival models?
- In-development new Gelman et al book: https://bookdown.org/jl5522/MRP-case-studies/
- Soch (2023) is a collection of proofs for results in probability, statistics, and related computational sciences.
- Suárez et al. (2017) covers GLMs but not survival analysis
- Greenland (2014) is a lengthy chapter from the Handbook of Epidemiology
- Rothman et al. (2021) contains several chapters on regression analyses in epidemiology
- Rawlings, Pantula, and Dickey (1998) is used in PLS 206[26]
- Bolker (2008) is used in PLS 207[27]
- Ken Rice[28]'s slides from Stat/Biostat 570 at University of Washington are also useful: https://drive.google.com/file/d/1VwosGvHtRtKnC7P3ja7RAUawvvudgc9T/view

Other similar courses at UC Davis:

- MPM 202[29], 203[30], 204[31] "Medical Statistics I-III"
- PHR 266/SPH 266[32] "Applied Analytic Epidemiology"
  - covers similar content; that course was designed for professional Master's students (e.g., MPVM, MPH) and does not assume a knowledge of mathematical statistics.
- PLS 206[33] "Applied Multivariate Modeling in Agricultural & Environmental Sciences"
- STA 101[34] "Advanced Applied Statistics for the Biological Sciences"
- STA 138[35] "Analysis of Categorical Data"

[22] https://www.r4epi.com/linear-regression
[23] https://www.r4epi.com/linear-regression-1
[24] https://www.r4epi.com/poisson-regression
[25] https://www.r4epi.com/cox-proportional-hazards-regression
[26] https://catalog.ucdavis.edu/search/?q=PLS+206
[27] https://catalog.ucdavis.edu/search/?q=PLS+207
[28] https://www.biostat.washington.edu/people/ken-rice
[29] https://catalog.ucdavis.edu/search/?q=MPM+202
[30] https://catalog.ucdavis.edu/search/?q=MPM+203
[31] https://catalog.ucdavis.edu/search/?q=MPM+204
[32] https://catalog.ucdavis.edu/search/?q=PHR+266
[33] https://catalog.ucdavis.edu/search/?q=PLS+206
[34] https://catalog.ucdavis.edu/search/?q=STA+101
[35] https://catalog.ucdavis.edu/search/?q=STA+138

    - emphasizes methods for analyzing categorical outcomes and predictors (i.e. contingency tables).
- STA 207[36] "Statistical Methods for Research II"

License

Agresti, Alan. 2010. *Analysis of Ordinal Categorical Data*. Vol. 656. John Wiley & Sons.

———. 2012. *Categorical Data Analysis*. Vol. 792. John Wiley & Sons. https://www.wiley.com/en-us/Categorical+Data+Analysis%2C+3rd+Edition-p-9780470463635.

———. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons. https://www.wiley.com/en-us/Foundations+of+Linear+and+Generalized+Linear+Models-p-9781118730034.

———. 2018. *An Introduction to Categorical Data Analysis*. John Wiley & Sons. https://www.wiley.com/en-us/An+Introduction+to+Categorical+Data+Analysis%2C+3rd+Edition-p-9781119405283.

Andrews, David F, and Agnes M Herzberg. 2012. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer Science & Business Media. https://link.springer.com/book/10.1007/978-1-4612-5098-2.

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. 1st ed. Princeton: Princeton University Press.

Cannell, Brad, and Melvin Livingston. 2024. *R for Epidemiology*. Online. https://www.r4epi.com/.

Clayton, David, and Michael Hills. 2013. *Statistical Models in Epidemiology*. Oxford University Press. https://global.oup.com/academic/product/statistical-models-in-epidemiology-9780199671182.

Dalgaard, Peter. 2008. *Introductory Statistics with r*. New York, NY: Springer New York. https://link.springer.com/book/10.1007/978-0-387-79054-1.

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

Dunn, Peter K, and Gordon K Smyth. 2018. *Generalized Linear Models with Examples in R*. Vol. 53. Springer. https://link.springer.com/book/10.1007/978-1-4419-0118-7.

Edelmann, Dominic. 2019. "Generalized Linear Models with Examples in r. Peter k.dunnand Gordon k.smyth (2018). Berlin, Germany: Springer Science+business Media, Pp. 562 Pages, ISBN: 978-1-4419-0118-7." *Biometrical Journal* 62 (1): 253–53. https://doi.org/10.1002/bimj.201900264.

Faraway, Julian J. 2016. *Extending the Linear Model with r: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. 2nd ed. Chapman; Hall/CRC. https://doi.org/10.1201/9781315382722.

Fox, John. 2015. *Applied Regression Analysis and Generalized Linear Models*. Sage publications.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge, MA: Cambridge University Press.

Greenland, Sander. 2014. "Regression Methods for Epidemiological Analysis." In *Handbook of Epidemiology*, edited by Wolfgang Ahrens and Iris Pigeot, 1087–1159. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-09834-0_17.

Hardin, James W, and Joseph M Hilbe. 2018. *Generalized Linear Models and Extensions*. 4th ed. Stata Press.

Harrell, Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd ed. Springer. https://doi.org/10.1007/978-3-319-19425-7.

Hosmer, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied Logistic Regression*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387.

[36] https://catalog.ucdavis.edu/search/?q=STA+207

[37] http://creativecommons.org/licenses/by-nc-nd/4.0/

[38] https://creativecommons.org/publicdomain/zero/1.0/

Jewell, Nicholas P. 2003. *Statistics for Epidemiology*. Oxford, UK: Chapman; Hall/CRC. https://www.routledge.com/Statistics-for-Epidemiology/Jewell/p/book/9781584884330.

Kalbfleisch, John D, and Ross L Prentice. 2011. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Klein, John P, and Melvin L Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. Vol. 1230. Springer. https://link.springer.com/book/10.1007/b97377.

Kleinbaum, David G, and Mitchel Klein. 2010. *Logistic Regression: A Self-Learning Text*. 3rd ed. Springer. https://link.springer.com/book/10.1007/978-1-4419-1742-3.

———. 2012. *Survival Analysis: A Self-Learning Text*. 3rd ed. Springer. https://link.springer.com/book/10.1007/978-1-4419-6646-9.

Kleinbaum, David G., Lawrence L. Kupper, and Hal Morgenstern. 1982. *Epidemiologic Research : Principles and Quantitative Methods*. Belmont, Calif: Lifetime Learning Publications.

———. 1983. *Solutions Manual for Epidemiologic Research : Principles and Quantitative Methods*. Belmont, Calif: Lifetime Learning Publications.

Kleinbaum, David G, Lawrence L Kupper, Azhar Nizam, K Muller, and ES Rosenberg. 2014. *Applied Regression Analysis and Other Multivariable Methods*. 5th ed. Cengage Learning. https://www.cengage.com/c/applied-regression-analysis-and-other-multivariable-methods-5e-kleinbaum/9781285051086/.

Kutoyants, Yury A. 2023. *Introduction to the Statistics of Poisson Processes and Applications*. Springer Nature. https://link.springer.com/book/10.1007/978-3-031-37054-0.

McCullagh, Peter, and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Routledge. https://www.utstat.toronto.edu/~brunner/oldclass/2201s11/readings/glmbook.pdf.

McCulloch, Charles E, Searle Shayle R, and John M Neuhaus. 2008. *Generalized, Linear, and Mixed Models*. 2nd ed. Vol. 651. John Wiley & Sons.

Moore, Dirk F. 2016. *Applied Survival Analysis Using R*. Vol. 473. Springer. https://doi.org/10.1007/978-3-319-31245-3.

Nahhas, Ramzi W. 2024. *Introduction to Regression Methods for Public Health Using R*. CRC Press. https://www.bookdown.org/rwnahhas/RMPH/.

Rawlings, John O., Sastry G. Pantula, and David A. Dickey. 1998. *Applied Regression Analysis : A Research Tool*. 2nd ed. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/b98890.

Roback, Paul, and Julie Legler. 2021. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in r*. Chapman; Hall/CRC. https://bookdown.org/roback/bookdown-BeyondMLR/.

Rothman, Kenneth J., Timothy L. Lash, Tyler J. VanderWeele, and Sebastien Haneuse. 2021. *Modern Epidemiology*. Fourth edition. Philadelphia: Wolters Kluwer.

Selvin, Steve. 2001. *Epidemiologic Analysis: A Case-Oriented Approach*. Oxford University Press.

———. 2004. *Statistical Analysis of Epidemiologic Data*. 3rd ed. Monographs in Epidemiology and Biostatistics ; v. 35. Oxford ; Oxford University Press.

Soch, Joram, ed. 2023. *The Book of Statistical Proofs*. Zenodo. https://doi.org/10.5281/ZENODO.4305949.

Suárez, Erick, Cynthia M Pérez, Roberto Rivera, and Melissa N Martínez. 2017. *Applications of Regression Models in Epidemiology*. John Wiley & Sons.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.

Wakefield, Jon. 2013. *Bayesian and Frequentist Regression Methods*. 1st ed. 2013. Springer Series in Statistics. New York, NY: Springer New York.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with r*. chapman; hall/CRC.

Woodward, Mark. 2013. *Epidemiology: Study Design and Data Analysis*. CRC press. https://www.routledge.com/Epidemiology-Study-Design-and-Data-Analysis-Third-Edition/Woodward/p/book/9781439839706.