Generalized Linear Models

Course Description

EDF 5401: Generalized Linear Models

Who am I?

Russell Almond

- http://ralmond.net/
- Email: ralmond@fsu.edu
- Office Hours, By appointment (appointment link: https://doodle.com/mm/russellalmond/book-a-time)
- https://fsu.zoom.us/my/ralmond
- Coffee Hour: TBD (Vote for time at: https://doodle.com/poll/nc3qvaf2i2rs4qti?utm_source=poll&utm_r
- Tea Time, TTH 4:30–5:00
- STB 3204-J
- Has ADHD

Questions welcome during lectures (speak up if on Zoom)

Who Am I?

Russell Almond (thee, thou)

Figure 1: Russell Almond

How to find me

- http://ralmond.net/
- Email: ralmond@fsu.edu
- Office Hours, By appointment appointment link
- Zoom: https://fsu.zoom.us/my/ralmond
- Coffee Hour: MW 10:30–11:00
- Tea Time, MW 4:30–5:00
- STB 3204-J
- Has ADHD

Questions welcome during lectures (speak up if on Zoom)

Textbooks

Required

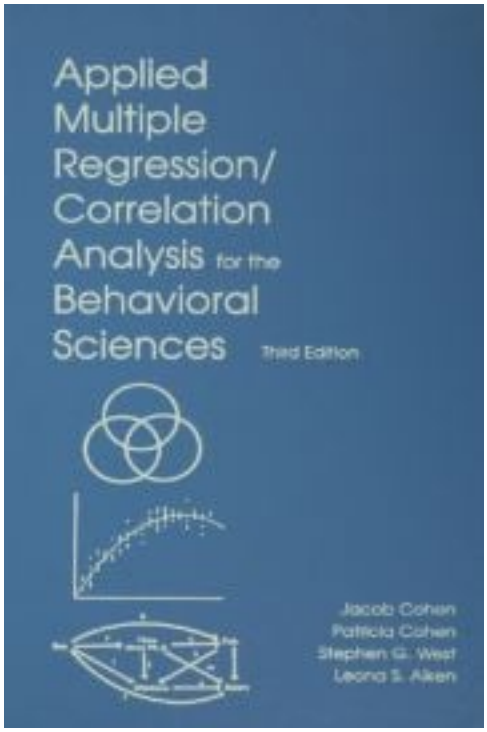


Figure 2: CCWA

Cohen, J., Cohen, P., West, S. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed).* Lawrence Erlbaum Associates, Inc. ISBN

9780203774441

This is called CCWA in the readings list.

Recommended

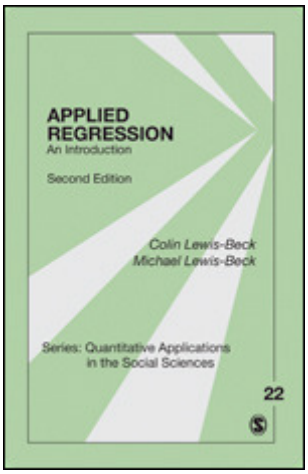


Figure 3: LB

Lewis-Beck, C. and Lewis-Beck, M. S. (1980). *Applied regression: An introduction.* Sage Publications. (ISBN e-book: 9781483381480, Paperback: 9781483381473)

Called LB in the reading lists.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables. Sage Publications. (ISBN e-book: 9781452210834, paperback: 9780761929895)

More suggestions

All of these books are available online.

Gelman, A., Hill, J. & Vehtari, A. (2020). *Regression and Other Stories.* Cambridge University Press. URL: https://avehtari.github.io/ROS-Examples/ (This text is used for EDF 5484.)

Navaro, DJ (2018). Learning Statistics with R: A tutorial for psychology students and other beginners. Version 0.6.1 URL: https://learningstatisticswithr-bookdown.netlify.app/

Wickham, H., Cetinkaya-Rundel, M., & Grolemund, G. (2023). R for Data Science. O'Reilly. ISBN 978-1492097402, URL: https://r4ds.hadley.nz

Leemis, L. M. (2023) Statistical Modeling: Regression, Survival Analysis, and Time Series AnalysisLinks to an external site., 2023, Lightning Source, ISBN: 978-0-9829174-3-5.

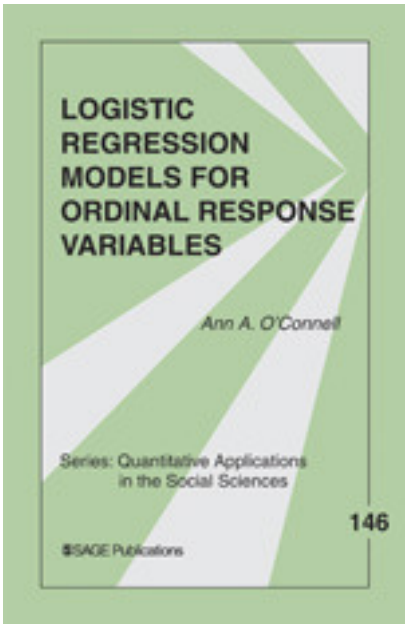


Figure 4: O'Connell

- Chapter 1: Simple Linear Regression
- Chapter 2: Inference in Simple Linear Regression
- Chapter 3: Topics in Regression

Additional Readings

Andy Gelman's Blog: https://andrewgelman.com

Software

Will do most work through Posit Studio:

https://posit.cloud/spaces/404203/join?access_code=pvsD1nGmC6ryEb0Nzhy8hEyQiDgakCGsgYys_Mop

Free Student Account.

Other Tools:

R

The R software is available through CRAN (Comprehensive R Archive Network), https://cloud.r-project.org/.

R is a programming language and command line tool. Most people use the RStudio integrated development environment. This is also available for free at https://posit.co/products/open-source/rstudio/.

Installing R

R and R Studio

SPSS

SPSS is rather expensive and is only available with time-limited licenses. SPSS is available on every computer in labs in College of Education. Students may also access to SPSS through FSU virtual lab. The FSU virtual lab can be accessed via: https://its.fsu.edu/service-catalog/end-point-computing/myfsuvlab

For help, search on YouTube "SPSS Regresssion".

Flipped Class

1) View lecture videos before class.
2) Ask questions in class or on discussion board
3) Will work on Case Studies in class.

Video Lectures in SPSS, will use R (through Posit Studio) in class.

Course Grading

- Participation (10%)
  - Self-reflections (5%)
    - On Qualtrics (link on Canvas)
    - For you to monitor your own learning, and provide feedback to me.
    - If you are having difficulty with a concept: Ask.
  - Case Studies (5%)
    - Done in class through Posit cloud.
    - There will be an SPSS alternative.
    - Post note in canvas about where your solution is.

- Homework (40%)
  - Four Problem Sets
  - May work in collaboration (up to 3 people).
  - Is officially late after grading is complete.
- Exams (50%)
- Take home
- 3 day window
- Will require access to R or SPSS
- No collaboration

Due dates and late work

Due dates are posted on Canvas.

When changes are necessary (usually extending deadline), these will be posted on canvas.

Participation activities can be completed up to the end of the class; but keeping up is helpful for me and you.

Homework late homework is accepted, however, if it is turned in after the answer key is posted on Canvas, the score will be reduced by 25%.

** Last Day for All Homework, Friday December 8, 2023.**

Exams are due on the day posted on Canvas. Students who are unable to complete the exam by the due date must contact the instructor to make arrangements before the due date, or they will receive no credit.

How's My Teaching?

Dial: ~~644-5203~~

Email: ralmond@fsu.edu

- Speed Up
- Slow Down
- I have a question

People on Zoom, feel free to unmute and ask questions.

Post on Discussion Forum

Come to coffee hours & online reviews.

Let me know if you have problems reading material

- Problems distinguishing black and red
- Let me know if you need better closed captions on videos

What to do if you are lost

Confusion is a part of learning

- But it should be temporary

Post questions in class forums!

Come to Tea and Coffee Hours

Make an appointment

Be specific about what is confusing you

Generalized Linear Models

What is a Statistic?

A *statistic* is an operator which summarizes a data set. This could be a numerical summary or a graphical summary.

Statistics is the study of statiscs.

What are Data?

Data (a plural noun) *are* a collection of observations about a collection of subjects.

Usually put into a spreadsheet (called data frame or tibble in R).

- Rows, i, represent individuals (unit on which measurement was made)
  - Total number of rows is N (population) or n (sample).
- Columns, j, represent variables–different measurements.
  - Total number of columns in J

If the data set is $\mathbf{X}$, then the value of Variable j for Individual i is $x_{i,j}$.

- In R, use the expression `X[i,j]` to extract values from a matrix, data frame, or tibble.

[Matrixes are bold upper case, values in the matrix are lower case]

The vector $x_i = (x_{i,1}, \ldots, x_{i,J})$ is the set of all measurements on Individual i. (Vectors are in bold face).

- In R, use `X[i,]` to extract an observation vector.

The response to an unstated individual on Variable j is written Y_j if it is a random variable, or y_j is if is the value of a random variable.

- In R, can select all values on a particular value using `X[,j]` or `X$j`.
  - The dollar sign notation expects a name, not a number, and only works for data frames and tibbles, but not for matrixes.
  - For tibbles, the dollar sign notation is usually better as it extracts just the variable.

Sum notation

$$\sum_{i=1}^{N} x_i = \sum_{i} x_i = x_1 + x_2 + \cdots x_n$$

This can be used sum a column (summation index i) or row (summation index j)

In R, this is a `for` loop:

```
x <- 1:5
x
```

```
[1] 1 2 3 4 5
```

```
xsum <- 0
for (i in 1:length(x)) {
  ## Code in braces will be repeated for every value of i
  xsum <- xsum + x[i]
}
xsum
```

```
[1] 15
```

The `sum()` function in R does this implicitly. Thus `sum(x)` is faster, to both code and run. Many statistic functions in R work the same way, e.g., `mean()`, `median()`, `sd()`.

- Beware of `NA`s, missing data, which are contagious. So `mean(x)` is `NA` if any element of `x` is `NA`.

- Use `mean(x,na.rm=TRUE)` to get the mean of the non-missing cases.

Using a matrix or a data frame, we usually are talking about row sums and column sums.

- Row sums: `sum(X[i,])`, `rowSums(X)`, `apply(X,1,sum)`
- Row means: `mean(X[i,])`, `rowMeans(X)`, `apply(X,1,mean,na.rm=TRUE)`
- Column sums: `sum(X[,j])`, `colSums(X)`, `apply(X,2,sum)`
- Column means: `mean(X[,m])`, `colMeans(X)`, `apply(X,2,mean,na.rm=TRUE)`

R only provides built-in `rowXXX` function for mean and sums, but the other two methods work with any function.

Distributions

$F(x) = \Pr(X \leq x)$

Also write $ X F( )$

Often a distribution has parameters (usually written as Greek letters).

- $F(x|\theta)$ – Bayesian style
- $F(x;\theta)$ – Frequentist style
- $X \sim N(\mu, \sigma)$ – X is normally distributed with mean μ and standard deviation σ. (Check carefully, authors sometimes uses variance, σ^2 or precision σ^{-2} instead of s.d.)

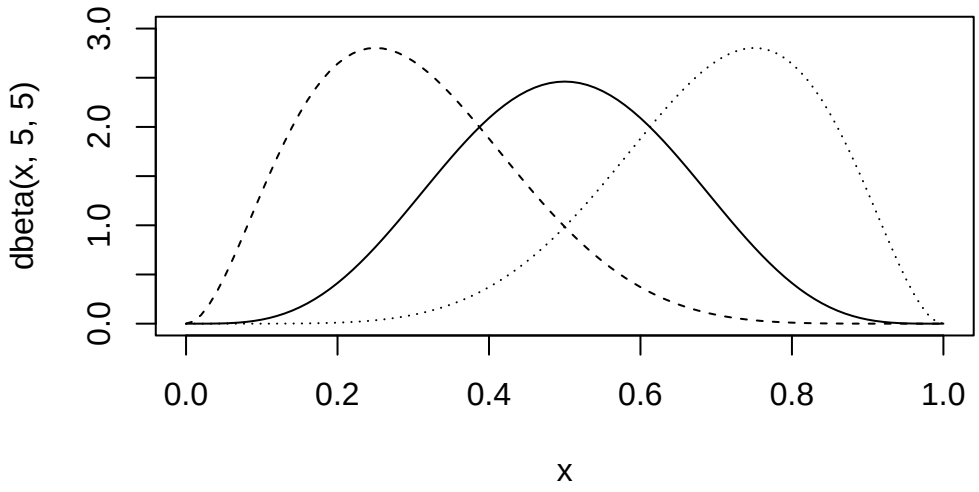
A joint distribution is over two or more variables:

$F(x, y) = \Pr(X \leq x \wedge Y \leq y)$, where $\wedge$ means *and*.

$X \sim N(\mu, \Sigma)$ ## 4 Moments of a Distribution

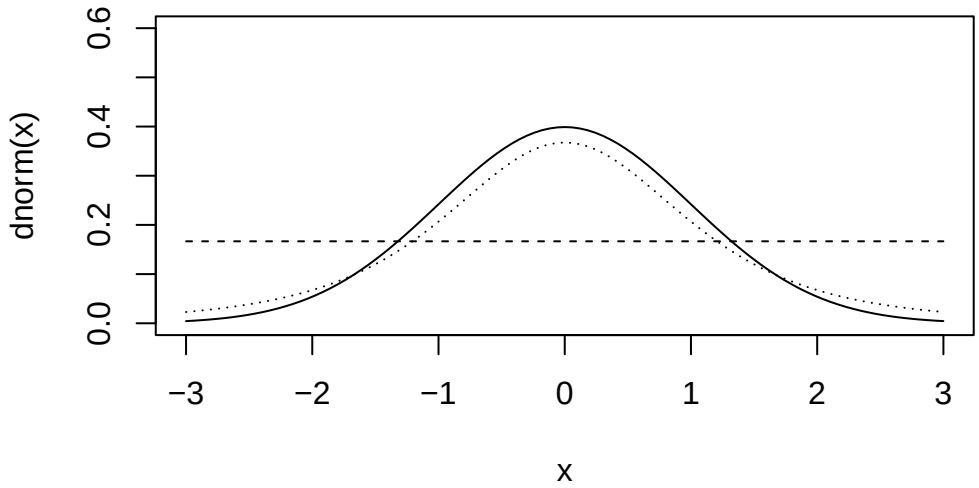
- Center: Mean, $E[X]$, $\bar{X}$ `mean(x)`
  - Median, `median(x)`; trimmed mean `mean(x,trim=0.05)`
- Spread: Variance, $\text{Var}(X)$, σ_X^2, s_X^2, `var(X)`
  - Standard Deviation, σ_X, s_X, `sd(X)`; Precision, σ_X^{-2}
  - Interquartile range (IQR) `IQR(x)`; median absolute deviation (MAD), `mad(x)`
- Skewness, `DescTools::Skew()`
  - Positive skewness has long tail to the right, negative to the left.

```
curve(dbeta(x,5,5),xlim=c(0,1),ylim=c(0,3.0),lty=1)
curve(dbeta(x,3,7),lty=2,add=TRUE)
curve(dbeta(x,7,3),lty=3,add=TRUE)
```



- Kurtosis, `DescTools::Kurt()`
  - normal is like normal distribution
  - platykurtic (flat) is usually not a problem
  - leptokurtic (heavy tails) is a problem because of lots of outliers

```
curve(dnorm(x),xlim=c(-3,3),ylim=c(0, .6),lty=1)
curve(dunif((x+3)/6)/6,lty=2,add=TRUE)
curve(dt(x,df=3),lty=3,add=TRUE)
```



Samples

A sample is a subset of the individuals in a population.

Under certain conditions, sample statistic converges to population statistic as the size of the sample increases (Law of Large Numbers).

So if $t(\mathbf{X})$ is a statistic, write $\hat{t}$ or $\tilde{t}$ for estimate.

Types of Samples

- Random Samples
  - Simple Random Sample – All individuals have an equal probability of being chosen
  - Stratified sample – Break population into groups (strata) and then do a SRS in each strata.
  - Cluster sample – Individuals come in clusters (e.g., students in a classroom). Randomly sample clusters, but take everybody in the cluster.
  - Sample weights – generally related ot the probability of being selected; used with more complex sampling methods.
- Non-random Samples
  - Systematic Sample – Individuals chosen accoridng to some deterministic rule.
  - Convenience Sample – Pick readily available individuals
  - Quota samples – Pick the first n volunteers (often in strata)
- Missing Data
  - *Missing Completely at Random* – complete cases are a SRS of all data.
    - Complete case analysis works.
  - *Missing at Random* – complete cases are a stratified sample of all data
    - Can use various techniques to adjust for missingness
  - *Non-ignorable Missingness* – need a more complex model for missingness.

Properties of an estimate

A statistic T from a sample is a random variable. Let T^* be the value of the statistic in the population.

- bias – a systematic error, $E[T - T^*]$
- standard error – standard deviation of the statistic (if we could take the sample over and over again), $s_T = \sqrt{\text{Var}(T)}$
- robustness/resistance – how sensitive the statistic is to outliers or other model assumptions
- efficiency – how good a standard error can be achieve given the sample size.

Representative Sample

Fundamental equation of statistics

$$\text{Estimate} = \text{Estimand} + \text{bias} + \text{residual error}$$

The residual error had mean zero, and is characterized by its standard error.

Generally, the residual error is easier to deal with than bias.

A *representative sample* is one which will produce an unbiased estimate.

Mean Square Error includes both bias and random error.

$$\text{MSE}(T) = \text{bias}(T)^2 + s_T^2$$

What is a model?

A model is a function we can use to make predictions.

Y is the *dependent variable*, the target of prediction.

$\mathbf{X}$ is a set of variables which will be used to predict Y, *independent variables.*

Coefficients of a model

Let $f(x) = b_0 + b_1 x$

$$E[Y|b_0, b_1] = b_0 + b_1 X$$

This model is a line; b_0 is the intercept and b_1 is the slope. In general, $f(x) = b_0 + b_1 x_1 + b_2 x_2 + \cdots$ is a *linear predictor*.

The *fitted value* or *predicted value* for Y_i is $f(x_i)$, often written $\hat{Y}_i$ or $\tilde{Y}_i$.

Error model and prediction error

To get a full model, we need to express how Y might deviate from its prediction.

Simplest model is normally distributed errors

$$Y_i \sim N(f(X_i), \sigma_e)$$

This can be rewritten

$$Y_i = f(X_i) + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma_e)$$

The difference between the prediction, $f(X_i)$ and Y_i is the residual, ϵ_i.

This is a **linear model**. The parameters are the coefficients (slopes and intercepts) plus the residual standard error.

A simple model

Line Parameters

Estimating parameters

Given a set of data Y and $\mathbf{X}$

Prediction error is $Y_i - \tilde{Y}_i$

Can define a *loss function*, $L(Y, f(X|\beta))$.

Least squares:

$$L(Y, f(\mathbf{X}, \beta)) = \sum_i (Y_i - f(X_i|\beta))^2$$

The values of β which minimize the square error, are the least square estimate, $\hat{\beta}$.

In machine learning, this is called *learning* or *training*.

Interpolation versus extrapolation

We want the training sample to be representative of the population of interest.

Predictions of new values inside the range of the training data are *interpolation.*

Predictions outside of the range are *extrapolation*.

Extrapolation has a big assumption; the data outside the training sample will behave like the data inside the sample.

Models as a decription of reality

If we think the variables in $\mathbf{X}$ influence Y, then the prediction should be better by including those variables.

Can't prove causality: - Can provide evidence for suspected causes - Can identify candidate causes for later testing

Box's Maxim

All models are wrong, but some are useful
– George Box [@box1976]

Models are an approximation of the truth, they always leave something out.

Often simplified version of the truth allows us to see certain relationships more clearly.

Review of EDF 5400

- EDF 5400 Review Quiz on Canvas

Some Review Tools:

- Normal Distribution
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/NormalParams.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/NormalCalculator.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/NormalZ-scores.Rmd
- Histograms and Boxplots
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/SkewnessPractice.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/SkewnessBoxplot.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/KurtosisPractice.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/KurtosisBoxplots.Rmd
- Probability
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/ConditionalProbability.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/Independence.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/LawOfLargeNumbers.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/LawOfLargeNumbersAnimated.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/CentralLimitTheorem.Rmd
- Inference
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/EffectSize.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/ConfidenceInterval.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/TestCI.Rmd
  - https://pluto.coe.fsu.edu/rdemos/IntroStats/VaccineCI.Rmd