Multilevel Models for Crowdsourced Route Ratings

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Will Jones

May 2016

Approved for the Division
(Mathematics)

_____

Andrew Bray

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Dedication

You can have a dedication here if you wish.

Introduction

Knock Software's *Ride Report* app combines a simple thumbs-up/thumbs-down rating system with GPS traces of bicycle rides to compile a crowdsourced data set of which routes are and are not stressful for urban bicyclists.

The app that collects the data is simple: *Ride Report* automatically detects when a user start riding their bike, records the GPS trace of the route, and then prompts the user at the end of the ride to give either a thumbs-up or thumbs-down rating. From this, they were able to create a crude "stress map" of Portland, OR, which simply shows the average ride rating of rides going through each discretized ride segment.

The app privileges reducing barriers to response to increase sample size over ensuring quality and consistent responses. This presents the first problem: how can we analyze ratings when riders are likely rating rides inconsistently?

At the same time, we have another challenge. We have ratings associated with routes, but we would like to know the effect of particular road segments, for both inference (what effect does this road segment have on the rating?) and prediction (given a route, what do we expect the rating to be?) purposes.

0.1 Accounting for Rider Rating Variance

For ratings we are interested in modeling variance between riders (as we might expect different rides to rate differently on average) and within riders (as riders may not rate the same route and conditions the same every time). To model this, we propose using multilevel regression, with random effects from each rider. This approach has been used in similar situations, in one case to model sexual attraction[1].

In a multilevel model, we fit a regression where a slope of intercept term is a random variable whose distribution is unique to all the groupings. For example, if we let r_i be the rating of the ith ride, X_i be the ride-level variables, then we can fit a regression:

$$\mathbb{P}(r_i = 1) = \text{logit}^{-1}\left(\alpha_{j[i]} + \beta \cdot X_i\right),$$

where α_j is the contribution of the j rider:

$$\alpha_j \sim N(\mu_\alpha, \sigma_j^2).$$

[1] Mackaronis, Strassberg, Cundiff, & Cann (2013)

We explore multilevel model further in Section 2.1 and multilevel models for riders in Section 4.2.

0.2 Addressing Road Segments as a Level

We examine multiple approaches to modeling road segments. In the first, we regard road segments as groups rides belong to, with the catch that rides can belong to multiple of these groups.

Chapter 1

Data Sources

Here is a brief introduction into using *R Markdown*. *Markdown* is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. *R Markdown* provides the flexibility of *Markdown* with the implementation of **R** input and output. For more details on using *R Markdown* see `http://rmarkdown.rstudio.com`.

Be careful with your spacing in *Markdown* documents. While whitespace largely is ignored, it does at times give *Markdown* signals as to how to proceed. As a habit, try to keep everything left aligned whenever possible, especially as you type a new paragraph. In other words, there is no need to indent basic text in the Rmd document (in fact, it might cause your text to do funny things if you do).

1.1 Ride Report

It's easy to create a list. It can be unordered like

- Item 1
- Item 2

or it can be ordered like

1. Item 1
2. Item 2

Notice that I intentionally mislabeled Item 2 as number 4. *Markdown* automatically figures this out! You can put any numbers in the list and it will create the list. Check it out below.

To create a sublist, just indent the values a bit (at least four spaces or a tab). (Here's one case where indentation is key!)

1. Item 1
2. Item 2
3. Item 3
    - Item 3a
    - Item 3b

1.2 Weather Data

Make sure to add white space between lines if you'd like to start a new paragraph. Look at what happens below in the outputted document if you don't:

Here is the first sentence. Here is another sentence. Here is the last sentence to end the paragraph. This should be a new paragraph.

Now for the correct way:

Here is the first sentence. Here is another sentence. Here is the last sentence to end the paragraph.

This should be a new paragraph.

1.3 Road Data

When you click the **Knit** button above a document will be generated that includes both content as well as the output of any embedded **R** code chunks within the document. You can embed an **R** code chunk like this (`cars` is a built-in **R** dataset):

Chapter 2

Data Transformation

2.1 Working in Road Networks

2.2 Using Nearest Neighbor Search for Map Matching Data

Chapter 3

Methods

3.1 Logistic Regression

3.2 Multilevel Models

Chapter 4

Model 1: Rides and Riders

4.1 Choosing Ride-Level Parameters

4.2 Adding Random Effects from Riders

4.3 Evaluating the Ride-Level Models

Chapter 5

Model 2: Segments as a New Level

5.1 Choosing Segment-Level Parameters

5.2 Evaluating Segment-Level Models

Chapter 6

Model 3: A Spatial Model

Chapter 7

Comparative Evaluation

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{.unnumbered}` attribute. This has an unintended consequence of the sections being labeled as 3.6 for example though instead of 4.1. The LaTeX commands immediately following the Conclusion declaration get things back on track.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

References

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/Hierarchical models*. The Edinburgh Building, Cambridge CB2 8RU, UK: Cambridge University Press, New York.

Mackaronis, J. E., Strassberg, D. S., Cundiff, J. M., & Cann, D. J. (2013). Beholder and beheld: A multilevel model of perceived sexual appeal. *Archives of Sexual Behavior*.