My Final College Paper

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

Naomi A. Boss

December 2020

Approved for the Division
(Mathematics)

Andrew P. 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 (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

#Abstract

Dedication

You can have a dedication here if you wish.

Introduction

Welcome to the *R Markdown* thesis template. This template is based on (and in many places copied directly from) the Reed College LaTeX template, but hopefully it will provide a nicer interface for those that have never used TeX or LaTeX before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won't have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of LaTeX in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You'll see more on this later.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

For additional help with bookdown Please visit the free online bookdown reference guide.

Chapter 1

Literature Review

1.1 Survey Methodology

```
   [@thompson_adaptive]When you are trying to make asertain something about a popul
   - Sampling Procedures
(1-9)

Probability sampling is used in surveying.
     - Methods (section likely mixed rather than divided like so)
```

Simple Random Sampling (Lohr, n.d.)Simple reandom sampling is a procedure in which you take n independent samples from a population of size N. There are two ways of executing a simple random sample, with replacement and without replacement. If you survey with replacement, every sample has probabiliy 1/N of being drawn. The advantage of surveying without replacement is that you do not have repeated observations in the sample. In finite population sampling a duplicate observation does not provide new information, so it is more usefule to sample without replacement. The samples in the population are still considered indepdent samples; however, since the population decreases by one every draw, the probability becomes $\frac{n!\text{N-n!}}{N!}$. (Lohr, n.d.)

(Thompson & Seber, n.d.)When you draw n distinct units of observation(unit) from a population of size N without replacement, it is called *Simple Random Sampling*. Each unit is indepdent of one another with equal probabilty of being drawn. The probability of drawing a specific unit from the population is 1/N. You can use a random number generator on a computer to determine which units to include in your sample. Since each unit has equal probability of being drawn, the population mean is the average of the observations:

$$\mu = \frac{1}{N}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

with a population variance of

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}\left(x_i - \mu\right)$$

Similarly, sample mean and variance are:

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_1 - \overline{x}\right)^2$$

You can also calculate the variance of the estimators themselves. If you use the population variance, you derive the variance of the estimator:

$$\operatorname{var}\left(\overline{x}\right) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$$

If you use the sample variance, then you get an unbiased estimator of the variance:

$$\widehat{\operatorname{var}}\left(\overline{x}\right) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

A variation of simple random sampling takes samples withoutreplacement. The draws are still independent of one another, but repeated observations is now allowed. This is used in situations where you do not want repeated observations. Now, instead of each draw being equal proability, we have that every sequence of samples is of equal probability. This slightly alters the sample mean

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

with a variance of

$$\operatorname{var}\left(\overline{x}_n\right) = \frac{1}{nN}\sum_{i=1}^{N}\left(x_i = \mu\right)^2 = \frac{N-1}{nN}\sigma^2$$

The unbiased estimator of the variance of the sample mean is

$$\widehat{\operatorname{var}}\left(\overline{x}_n\right) = \frac{s^2}{n}$$

(Thompson & Seber, n.d.) **Stratified:** (Lohr, n.d.)Stratified random sampling is a variation of sampling in which you partition the population into disjoint subpopulations that you sample from. The overall outcome of the survey is the combination of independent probability samples taken from the disjoint groups. There are several reasons why stratified random sampling can be advantageous. Stratified sampling can provide more precise estimates, especially if your aim involves comparing subgroups. Moreover, it allows flexibility in the method of conducting the surveye by customizing for each strata.(Lohr, n.d.)

(Thompson & Seber, n.d.)Another method of sampling is called stratified sampling, where you separate the population into disjoint subpopulations algorithmically. Then you can sample from each strata independenty. You can combine the mean and variations of each strata to caluclate the total variance and mean of the sample. One

goal when designing the partition, is to minimize the variance within each strata. If the sampling procedure of each strata is simple random sampling, then the method is referred to as stratified random sampling. Both the sample mean and the ith observation are random variables Here we can define elements and probabilities for stratified sampling. Let h refer to a specific strata, with the population partitioned into L stratas. Let us define the terms and variables associated with stratified sampling. Let h be a designator of the strata. The total population size is the sum of the sizes of each strata, N_h. The sample size is the sum of the samples taken from each strata, denoted n_h. We define τ as the sum of the unit values in the population, and τ_h to be the sum of the unit values in strata h. The population mean is μ/N and the mean of each strata is $\mu_h = \tau_h/N_h$. Let $\hat{\tau}_h$ be the unbiased estimator of τ_h. The unbiased estimator of the population is

$$\hat{\tau}_{st} = \sum_{h=1}^{L} \hat{\tau}_h$$

The variance of this estimator is the sum of the variances of the strata estimators:

$$\text{var}\left(\hat{\tau}_{st}\right) = \sum_{h=1}^{L} \text{var}\left(\hat{\tau}_h\right)$$

with an unbiased variance is similarly equal to the sum of the unbiased variances of each strata. Regardless of sampling method, we can define the stratified estimator of the mean to be

$$\hat{\mu}_{st} = \hat{\tau}_{st}/N$$

The unbiased variance of the estimator, given that the sample selection from each strata are indepdent, is

$$\widehat{\text{var}}\left(\hat{\mu}_{st}\right) = \frac{1}{N^2} \sum_{h=1}^{L} \hat{\text{var}}\left(\hat{\tau}_{st}\right)$$

(Thompson & Seber, n.d.) **Cluster:** M: total number of seconary units N: the number of primary units M_i total number of secondary units in the ith primary unit x_i: the total of the values in the ith primary unit x_{ij}: the jth secondary unit in the ith primary unit τ: population total of secondary units μ: population mean per secondary unit μ_1: population mean per primary unit $\hat{\tau}$: unbiased estimator of population total var($\hat{\tau}$): variance of the unbiased population total estimator σ_u^2: variance of the primary unit totals $\widehat{\text{var}}(\hat{\tau})$: unbiased estimate of the variance s_u^2: sample variance of the primary unit totals $\hat{\tau}_r$: ratio estimator of population total r: sample ratio p_i: probability of draw proportional to ith primary unit size.

(Lohr, n.d.)Cluster sampling is similar to stratified sampling because both methods involve segmenting the primary population; however, the choice for how the population is split is different. A unit is included in the sample if it falls within the partition of the cluster. When calculating probabilities, you need to consider how many clusters there are, as well as how many units are within each cluster. A cluster sample can be advantageous because it is cheaper to conduct than other methods. Moreover, populations are often distributed in natural clusters such as comparing universities. (Lohr, n.d.)

(Thompson & Seber, n.d.)The structure of cluster sampling and systemic sampling is the same. First the population is partitioned into units called *primary units*. Within each unit was have *secondary units*, or the units of observations. The primary unit of a cluster sample is determined by spacing proximity. On the other hand, the primary unit of a systemic sample is determined by equally spaced secondary units. The key aspect of cluster sampling and systematic sampling is that you choose the sample by selecting primary units, therefore, all secondary units in a primary unit will be included in the sample. There is variability in characterisitics of the primary units in cluster sampling, such as size and shape, which can provide supplementary information but can affect the efficiency of the sample.

The selection of primary units is conducted by other sampling procedures.You can use simple random sampling to select a primary unit. Calculating the unbiased estimators using simple random sampling without replacement: Estimator of population total

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} = N\overline{y}$$

where $\overline{y} = (1/n) \sum_{i=1}^{n} y_i$ is the sample mean of the primary unit totals. Let σ_u^2 be the finite population variance of the primary unit totals

$$\sigma_u^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_i)^2$$

Then the variance of the population total estimator is

$$\operatorname{var}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n}$$

and an unbiased variance is

$$\widehat{\operatorname{var}}\left(\tau^2\right) = N(N-n)\frac{s_u^2}{n}$$

s_u^2 is the sample variance of the primary unit totals, calculated by

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

The ratio estimator based on size can be more efficient than simple random sampling. The primary unit totals for cluster sampling do not have to be the same size, so the unit total is highly correlated to unit size. The sample ratio r is defined as

$$r = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i}$$

Then the ratio estimator of population total is

$$\widehat{\tau}_r = rM$$

You can approximate the variance or mean squared error of the ratio estimator

$$\operatorname{var}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (x_i - M_i\mu)^2$$

which has an estimator

$$\widehat{\operatorname{var}}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{n} (x_i - rM_i)^2$$

The population mean per primary unit is calculated by dividing the ratio estimator by the number of primary units. (Thompson & Seber, n.d.)

(Lohr, n.d.) Systemactic sampling is a specific kind of cluster sampling. First decide on the number of samples n. Choose a random number between 1 and the number of clusters, this will (Lohr, n.d.)

(Thompson & Seber, n.d.) The Hansen-Hurwits estimator is a cluster sampling method which proportionalizes the probability of drawing the primary unit based on the size. The unbiased Hansen-Hurwitz estimator is

$$\hat{\tau}_p = \frac{M}{n} \sum_{i=1}^{n} \frac{x_i}{M_i}$$

```
    - Advantages/Disadvantages
    pg 255
    Unequal sampling - chp 6
    pg 49
    113-125
- Data Collection
    - Designing survey (sample unit, sample size, hypothesis/motivation role)
    pg 247
    pg 258
    31-33
    - Survey method (brief list e.g. interview, questionnaire)
- Data Analysis
    - Hypothesis testing
    - Inference
```

1.2 Non-Probability Sampling

```
- Sampling Procedures
    - Methods
        - Convenience sampling
        - Consecutive sampling
        - Judgmental Sampling
        -Hansen-Hurwitz
```

```
          46-49
     ##Advantages/Disadvantages
- Data Collection
     - Designing Survey (sample unit, sample size, hypothesis/motivation role)
     - Survey method (brief list e.g. interview, questionnaire)
- Data Analysis
67
94-99
     - Hypothesis testing
     - Inference
```

1.3 Election Polling

```
- Purpose
- Design
     - Population
          - Sample selection
     - Methods
     - Sample size
     - Question design
          - Wording impacting response
- Data collection
     - Implementation
     - Errors
          - Sampling Error
          - Non-Sampling Error
               - Questionaire design
               - Interviewer effects
               - Responder effects
- Data analysis
```

Unclear Questions: Simple random sample implies that an unit has a 1/N probability of being drawn. What is the distinction from the probability than an ith unit of the popultion is included, which is n/N

3) “Design an Inference in Finite Population Sampling” A. S. Hedayat, Bikas K. Sinha hed91

Chapter 2

Replication of Variance

Chapter 3

Tables, Graphics, References, and Labels

3.1 Tables

In addition to the tables that can be automatically generated from a data frame in **R** that you saw in [R Markdown Basics] using the `kable` function, you can also create tables using *pandoc*. (More information is available at `http://pandoc.org/README.html#tables`.) This might be useful if you don't have values specifically stored in **R**, but you'd like to display them in table form. Below is an example. Pay careful attention to the alignment in the table and hyphens to create the rows and columns.

Table 3.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following: Table 3.1. If you go back to [Loading and exploring data] and look at the `kable` table, we can create a reference to this max delays table too: Table **??**. The addition of the `(\#tab:inher)` option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

3.2 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `reed.jpg` in our main directory. We then give it the caption of "Reed logo", the label of "reedlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

```
include_graphics(path = "figure/reed.jpg")
```

Figure 3.1: Reed logo

Here is a reference to the Reed logo: Figure 3.1. Note the use of the `fig:` code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter **??**. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

```
flights %>% group_by(carrier) %>%
  summarize(mean_dep_delay = mean(dep_delay)) %>%
  ggplot(aes(x = carrier, y = mean_dep_delay)) +
  geom_bar(position = "identity", stat = "identity", fill = "red")
```

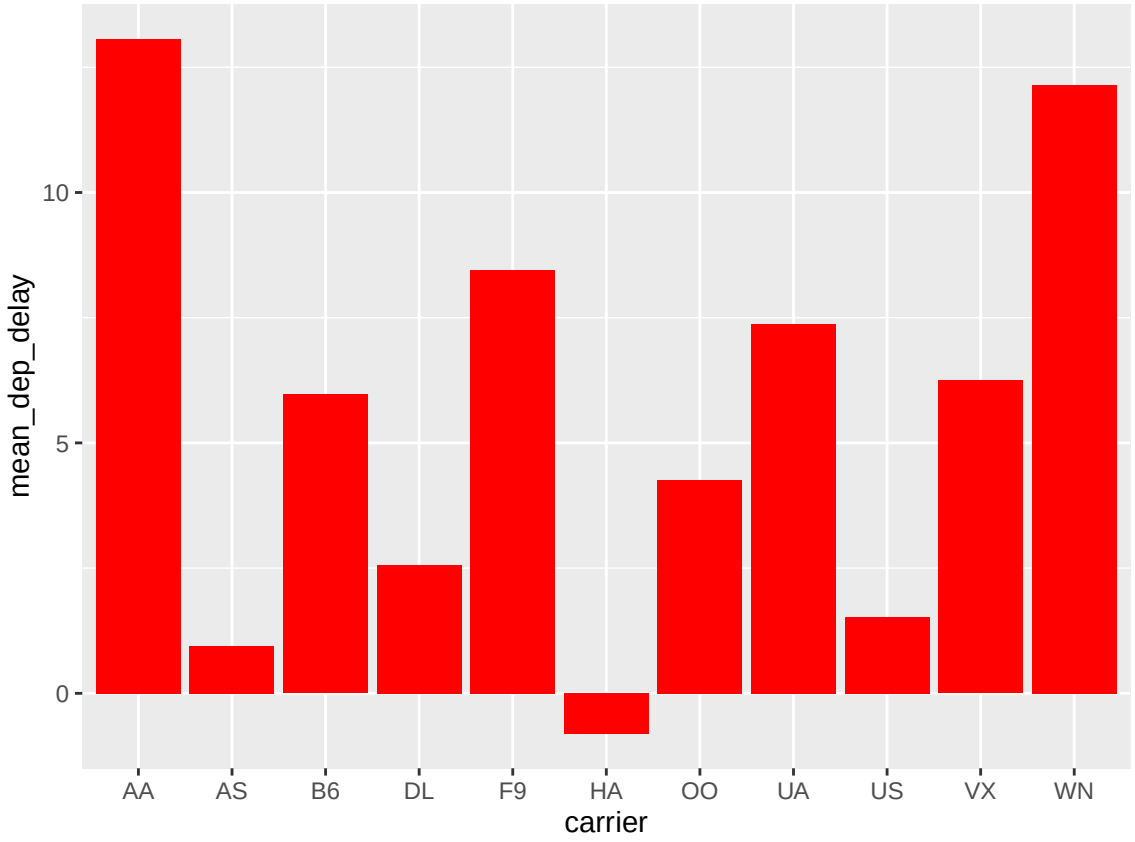

Figure 3.2: Mean Delays by Airline

Here is a reference to this image: Figure 3.2.

A table linking these carrier codes to airline names is available at `https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv`.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

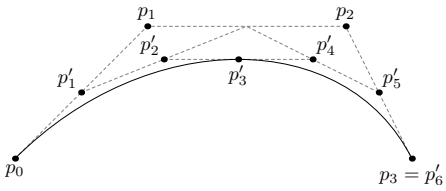

Figure 3.3: Subdiv. graph

Here is a reference to this image: Figure 3.3. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)

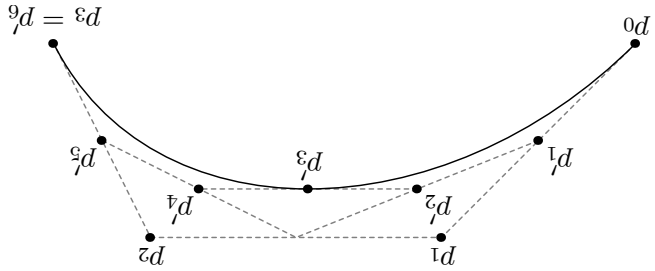

Figure 3.4: A Larger Figure, Flipped Upside Down

As another example, here is a reference: Figure 3.4.

3.3 Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to `data@reed.edu`.

3.4 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/`

[1] footnote text

`citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (**???**). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

[2] (**???**)
[3] (**???**)

3.5 Anything else?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

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.

Appendix A

The First Appendix

Partial Pooling Multilevel Regression

Multilevel linear regression is a method of linear regression where the coefficients of the parameters vary at multiple levels of the data. We often use this method when the data we have is clustered or nested. One way to think about multilevel modeling is as a compromise between aggregated and disaggregated regression, otherwise known as complete pooling and no-pooling. We will use the following notation for single level grouping: * j = 1,. . . ,J groups * j[i] identifies the group that the i^{th} obsrvation belongs in. (e.g. j[19]= 3 means that observation 19 belongs in group 3) * k predictors * Vector β of coefficients * varying intercept model: $y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$ or $y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$ * varying intercept, varying slope model: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$ or $y_i \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$ * multiple predictors: $y_i = X_i + B + \epsilon_i$ or $y_i \sim N(X_i B, \sigma_y^2)$ where B is a matrix of coefficients.

Consider the following example. Say that we are investigating the number of sourdough loaves baked per bakery in J=6 cities. The city of Lilliput has 12 bakeries, Gotham City has 2 bakeries, (3) has 7 bakers, Arkham has 15 bakeries, and Seymour's Bay has 10 bakeries. One attempt would be to find the average loaves per bakery across all of the cities. This is known as complete-pooling, where we do not consider the categorical predictor in the model (i.e. ignoring which city the bakery is in). This method ignores the differences between each city. For example, let's say that in Gotham City, the citizens ate more bread than those in (3). Then it would make sense that the bakeries in Arkham would, on average, bake more bread than those in (3) but since we average across all groups we would be underestimating loaves per bakery in Arkham and overestimating the loaves per bakery in (3). This can run the risk of falling into the ecological fallacy. After sampling, we find calculate $\overline{y} = 215$.

On the otherhand, we could fit an estimate for each city that we are sampling from, which would result in 6 different estimates. This is called no-pooling. This method can result in bad estimates for groups with small sample sizes. Using the no-pooling method, we find that $\overline{y}_1 = 180$, $\overline{y}_2 = 250$, $\overline{y}_3 = 190$, $\overline{y}_4 = 250$, and $\overline{y}_5 = 205$. Since we only sampled from 2 bakeries in Gotham City, can we really say that all of their bakeries bake that much? This makes it seem like the bakeries from city to city are more different than they actually are.

Instead, let us use multilevel modeling, otherwise known as partial pooling, to

make estimates for the bakeries. We can consider this a compromise between complete pooling and no-pooling. Roughly speaking, the multilevel estimate of the average loaves per bakery is the weighted average of the mean within the city and the mean across all of the cities. We weight the means with their respective variances. The amount of information we have for each city is seen in the weighted averages. If we have a small sample, then the partial pooled mean will be pulled closer to the complete pooled mean instead of the no-pooled mean. We can see this in the following example.

Let's compare the estimates for Gotham City and Arkham. Let's assume that the within county variance, $\sigma_y^2 = 3$ and the overall variance between bakeries be $\sigma_\alpha^2 = 4$. The partial pooled estimate of Gotham City would be:

$$\frac{\frac{2}{3}250 + \frac{1}{4}200}{\frac{2}{3} + \frac{1}{4}} \approx 236$$

Let's compare this estimate to the partial pooled estimate of Arkham

$$\frac{\frac{10}{3}250 + \frac{1}{4}200}{\frac{10}{3} + \frac{1}{4}} \approx 247$$

We can see that the partial pooled estimate of Gotham City which only sampled from 2 bakeries is much closer to the complete pooled estimate as compared to the partial pooled estimate of Arkham.

Now let us consider partial pooling with predictor variables. For our scenario, the predictors are $\mathbf{B}$ = (which cities these bakeries are in, the number of patrons per day, the average income of the city, and the average household size of the city). In a regression setting, complete pooling with predictors fits a model onto all of the data, ignoring the group it resides in. In this case we would estimate a single coefficient for each predictor based on all of the data, assuming that each bakery is independent of one another. The complete pooling regression can be written as $y_i = \alpha + BX_i + \epsilon)i$. On the otherhand we have the no-pooling model. An extreme definition of no-pooling is when you fit a regression model onto each group, resulting in k models. However a more common use of the term no-pooling is when when we include an variable indicating the group the observation belongs to. We then run a regression model including the factor variable, which has it's own coefficient. This model assigns a different intercept to each group. A no-pooling regression can be written as $y_i = \alpha_{j[i]} + BX_i + epsilon_i$ with j[i] indicating the group. As before, the complete pooling method ignores the variation between the cities and the no-pooling analysis can still overfit the data.

This brings us to multilevel regression analysis. The simplest partial pooled regression, the varying-intercept regression, can be written as $y_i \sim N(\alpha_{j[i]} + BX_i, \sigma_y^2)$. The key difference between the varying-intercept model and no-pooling is the α_js. In no-pooling regression, $\alpha_{j[i]}$ is determined by using the method of least squares for the intercept of each county. In the multilevel model, the α_js are assigned a probability distribution, $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$ for j=1,. . . ,J which places a constraint on the intercept. Here σ_α^2 is the county level variance and μ_α is the estimated average regression for all of the data. In otherwords, the partial pooled regression model pulls the intercept closer

to the overall average intercept as the group level standard deviation gets smaller. We can express the partial pooling estimate for α_j as

$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}(\overline{y}_i - B\overline{X}_j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}\mu_\alpha$$

References

Lohr, S. (n.d.). *Sampling: Design and Analysis*. Duxbury Press.

Thompson, S., & Seber, G. (n.d.). *Adaptive Sampling*. Wiley.