A Minimal Book Example

Yihui Xie

2019-03-26

Contents

Chapter 1

About the Book

A book of personal notes on different important methods for a comprehensive exam on methodology.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 4.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```r
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```r
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2018) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

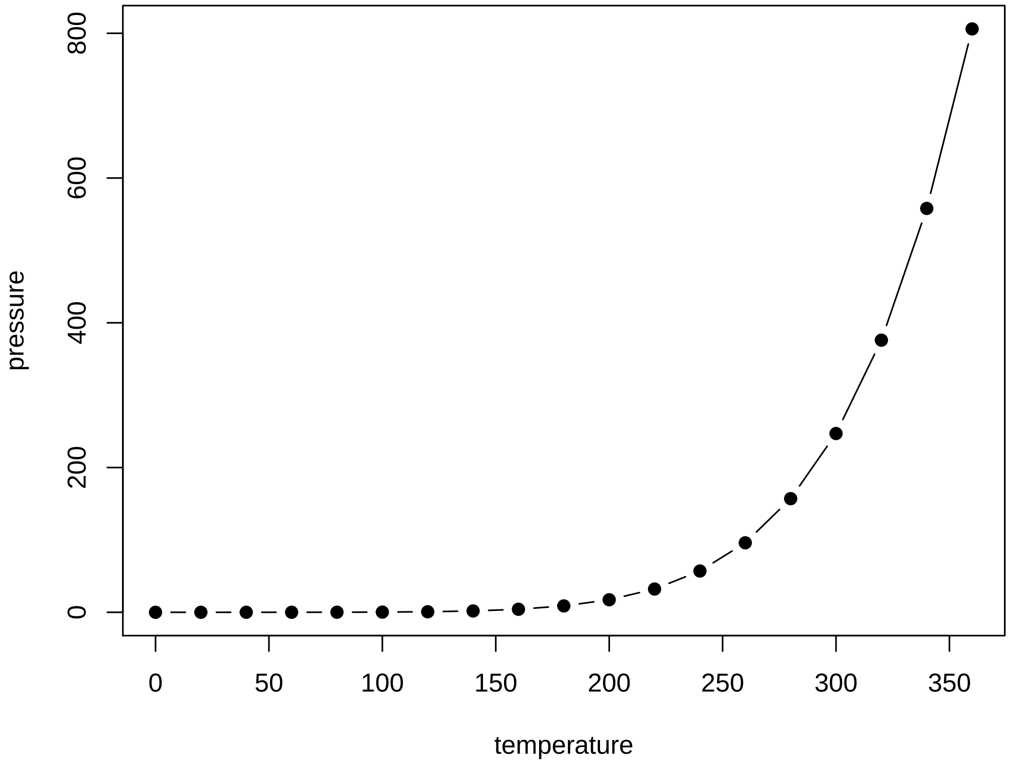

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Multi Level Modeling

3.1 Background

Multilevel data most frequently contains observations that are nested within larger spatial categories or groupings (individuals within a country) (Jones 2010). Can also contain observations that are nested temporally (annual GDP) and may even be nested in larger spatial groupings, across time (individual responses within surveys within years). The beauty of multilevel modeling is that it does not lose the information that will happen when you pool the information or use techniques like fixed effects (Achen 2005).

3.1.1 How to Deal with Multilevel Data

Disaggregate group data to individual level: for example, individual data nested within states, include state level variables at individual level with same values for all individuals in a given state. PROBLEM: all unmodeled contextual information (usually macro effects) ends up in the error term. Individuals within same macro group then have correlated errors (violating OLS assumption).

Pooling: when you assume there is not heterogeneity among the units (so that people in U.S. are similar to people in Canada, Mexico, New Zealand, China, and Nigeria). If you are assuming this, then you can use a garden variety regression (Jones 2010). However, this assumption will likely be incorrect (Franzese 2005). If incorrect, then you likely introduce heteroskedasticity as well as autocorrelation (this mainly occurs because respondents, i, in U.S. will be more alike to each other than respondents in New Zealand (Jones 2010; Priomo ea 2007). This will lead to inefficient and inconsistent standard errors leading to invalid hypothesis testing (Jones 2010). If you are only afraid of non-spherical errors, then you can use what is commonly referred to as the “Huber-White sandwich” estimator (White 1980; Jones 2010; Primo ea 2007). However, much of the information that we care about is lost if we don’t use multilevel modeling.

Solutions:

- Fixed effects (not very common in political science): Essentially adding an additional dummy variable for each macro-level groupings to account for contextual variation. It is shown with the following equation: $y_i = \beta_{j[I]} + \beta_1 x_i + \epsilon_i$.where β_j gives a different intercept for each unit. Prevents correlated error issue, but will be inefficient given that we burn a degree of freedom for every new independent variable we are adding to the model. Additionally, all covariates that are constant within j cannot be estimated (Jones 2010). A variation of the fixed effect is to estimate a model for each unit and then compare the different coefficients, however the unbalanced nature of political science data commonly precludes this kind of analysis or at least makes some estimates will wide variability that has nothing to do with theory and everything to do with sample size (Jones 2010).
- Random effects (not very common in political science): like fixed effects, allows the estimation of different intercepts for each macro-level group. The formula for this one is $y_i = \beta_{j[I]} + \epsilon_i$. The

difference between fixed and random effects is the way $\beta_{j[I]}$ is treated. Under fixed effects, the unit effects are unknown constants that are estimated from the data (Faraway 2006). This approach treats it as a random coefficient by assuming these intercepts are randomly drawn for a given (usually normal) distribution (Gelman and Hill 2007). However, estimates may be biased.

- Clustering (much more common but kind of like a band-aid): essentially a statistical “fix” of the problem by allowing a compound error term that accounts for the macro-level information. It is a variation of the Huber-White robust standard error. Cluster-adjustment works well for procedures including logit and probit assuming that the number of clusters is large enough. Simulations have shown that 50 clusters are more than sufficient (Kezdi 2003).
- Multilevel Modeling (probably theoretically better, but not necessarily methodologically): Also known as Hierarchical Linear Modeling or Mixed Effects Modeling. Commonly used in education research (students within classes, within schools, within school districts, etc). Goal is to predict influences on a dependent variable using independent variables from several contexts (individual and macro). I think the equation looks like this: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_j + \epsilon_i$ where the x is the level 1 data and z is the level 2 data.

3.2 Assumptions

Not so much an assumption, but you should have a lot more observations in the first stage than in the second stage (Achen 2005). However, having a small number of observations at the second level, we open ourselves up to bias and incorrect hypothesis testing (if the distributional assumptions are off (Bowers and Drake 2005)). Bowers and Drake (2005) suggest using graphs of the differences between second level groups to check assumptions. Additionally, hierarchical models are delicate meaning that specification error in the first stage will lead to biases and inconsistencies at the second stage, even if the second stage is properly handled (Achen 2005).

3.3 Estimating Multilevel Modeling

For multilevel modeling:

Level 1: $y_{ij} = \alpha_{j[i]} + \beta x_i + \epsilon_i$

Level 2: $\alpha_j \ N(\mu_\alpha, \sigma^2_\alpha)$ or $\alpha_j = \gamma_0 + \gamma_1 u_j + \eta_j$

Where: i = individuals, j = groups

Considerations for Multilevel modeling:

- How many levels are in the data? - Social science generally only has 2 or 3
- How many predictors for each level are needed? - Modeling becomes increasingly complex as these increase (especially for macro-level predictors). Are any cross-level interactions hypothesized? Which parts of the model will include random effects? What structural form will you use? -Varying intercepts only, varying slopes only, varying intercepts and slopes.
  - Varying intercepts: will have same slope, but cross y-axis at different places.

  - Varying slopes: will have same intercept, but different slopes
  - Varying slopes and intercepts: will have both

If you choose varying intercept and slope, you add additional layer of complexity. Level 1 stays the same, but level 2 now has to account for fact that the intercept and slope will vary.

3.4 Additional Considerations for Multilevel Modeling

Number of groups: Some argue that a minimum number of groups is needed for multilevel modeling. However, even with a small number of groups, a multilevel regression will simply reduce to a classical regression. Therefore, the number of groups is a limitation, only in that the estimation of between-group variation will be limited.

Number of observations per group: Another issue that is brought up, but doesn't really exist. With small numbers of observations in some groups, estimates of the α parameters for those groups will be imprecise. Also, if there is significant imbalance there can be issues with random effects estimates.

Chapter 4

Methods

We describe our methods in this chapter.

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr.* Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.9.