# Comments on Advanced Epidemiological Methods

## General comments:

* I feel there is a slight disconnect between how this book is initially presented as openly accessible and relevant for everyone regardless of whether or not they are taking the course, and the language in the actual book. There is often language around submitting assignments or how much a question in an assignment is worth. These are not going to be relevant to those not taking the course. I would either market it as primarily for those who are taking the course, and open to those who are not, though there may be some parts that are not relevant to those not taking the course. Or get rid of language that is course-specific and instead include all of this information only in the canvas course.
* Organization looks good, I made a few comments here and there about things that would make the transitions between sections smoother.
* I think the lists of new R functions adds a lot of value. Perhaps the availability of these pages could be highlighted in the R basics section on built-in functions.
* Could consider reducing the amount of commented results that are given in the code chunks. I know it is important to show results for the student to understand what is happening, but sometimes the wall of commented code can become overwhelming enough that I think they would likely just skip it entirely. Instead could write a commented line of code and describe what it does (e.g. “Use str(allImputations) to visualize the structure of the imputed data now that we have formatted it as a list”). Just a suggestion.
* See note about video in Missing data analysis – imputation in NHANES

## Title Page:

### The Project

* “Place” sounds a bit odd in the first sentence, I would maybe replace with “website.”
* Third sentence maybe flip: “If advanced statistical methods seem daunting, whether or not you are familiar with health research, you are in the right place”
* Write out CDC?
* Perhaps in here somewhere it could be explicitly stated what exactly is the “unique gap in the health research world” this website is bridging.
* “raise” rather than “uplift” in the last sentence?

### What we aim to achieve

No comments

### Dive into our modules

Module 5 in the table – Issues should be capitalized. Perhaps also consider calling “predictive modeling” rather than “predictive issues”

### How our content is presented

No comments

### Grant applicants

* Write out acronyms?
* Last sentence: “**This** knowledge **has now been** transformed…”

### Acknowledgments

* “including” not include

# Data wrangling

### Background

No comments

### Overview of tutorials

* The difference between “Importing Dataset” and “Import External Data” is not super clear from the summaries presented on this page.

## R basics

### Basic syntax

* I would suggest either ending section 4 (creating functions) with an example of how to call the function we just created, or starting section 5 with calling the function we just created. I feel this would be the natural step to follow building a function
* Perhaps the work with the rnorm function and setting the seed could be moved into a new section on producing reproducible code

### Package management

* Are there any resources that would help someone understand how to find available packages for what they want to do? I typically just google what I want to do but I’m wondering if there are any other resources for common packages or just a list of available ones that would be helpful to those new to R. Could be that this does not exist

## Data types

### Matrix

* Maybe worth mentioning creating an empty matrix is also possible? I.e. matrix(nrow=5, ncol=5)

### Data frame

* Suggest introducing that columns can be accessed using $ in section 2 (extract elements), before getting to section 3 (modifying values)

## Automating tasks

### Repeating tasks

* The way the first for-loop example is described, it sounds like the printing happens after the 10 cycles are done. But I think when I run this in R it prints as it goes, doesn’t it?

### Functions

* It feels like this repeats some information from a previous section… didn’t we already learn functions in the basics?

## Importing dataset

### Importing .txt files

* Not sure if overkill, but I think it would be helpful to have screenshots of the heads of .txt datasets that contain or don’t contain variable names so the students can see what is meant by this. Either that or potentially consider adding the language “column headers” or “column names” or something like that

## Data manipulation

### Subset

* The last sentence in the introduction to dropping variables confused me – I was expecting 2 examples using no package and 1 using dplyr. I would suggest clarifying this – maybe “Below are two examples – one without using packages and one with the dplyr package” or something similar.
* In the section on keeping observations, Tip for option 3 should say “use filter in dplyr” not select.
* For the tip on common logical tests, I found the (=) for < and > a bit confusing and would suggest having separate rows for <, <=, >, >=. I know this takes up more space but would be clearer.
* Converting characters to factors: change “factor function can be also used to combine factors. If the user want to combine multiple factors to one factors” to “The factor function can also be used to combine multiple factors into one factor”

## R Markdown

### Working with RMD

* The portion about omitting the code (echo=FALSE) is repeated in two consecutive sentences.

## Exercise (W)

### Problem Statement

* First sentence in second paragraph does not make sense: Take out the “a paper” at the start.
* Will this book be used by people who are not taking the 604 course? If so, the instruction to submit the knitted file may be confusing. But if this is mainly meant for people taking the course then perhaps it is fine.

# Accessing Data

## Title Page

### Overview of tutorials

* In the note on the bottom of the page, the link for “Roles of Variables” leads to a page titled “Causal Roles” and the link for “Prediction Models” leads to a page titled “Prediction ideas” – I would suggest making sure the links & titles match to avoid confusion

## Descriptions

### CCHS

* There is no description of the sampling procedure used for the CCHS. There is a mention of updated sampling methods in 2015 but no mention of what the new sampling methods are or what the old methods were.

### NHANES

* Missing data & outliers: typo – Key points on **NHANES** data analysis

## Importing CCHS to R

### Option 1

* Step 3 should say “if you **want** to view…”

### Option 3

* Again the language here sounds as if everyone using textbook is in the course “Let me know if that is not the case” – not sure if this is intended or remains from taking material from the course, but since it is marketed as available to everyone, perhaps consider adjusting the language.

## Importing NHIS to R

### Table 1

* The use of psu, cluster, and strata lacks explanation. I would at least link to a chapter where the use of these for survey data is explained

# Research Questions

## Overview of tutorials

* Suggestion to change the language to future tense in the gray note on the side (“In this chapter, we **will** embark on a journey….”) since we have not gone through this chapter yet.

### Causal questions

* Typo: “fo**u**rth” tutorial

## Predictive question-1

### Creating analytic dataset

* I think you mean to say “so that the variables generally match with the way **they were coded** in the original article.”?

## Predictive question-2b

### Imputed data

* I think it would be helpful here to link to the missing data analysis materials that will be covered later

## Causal question-1

### Combine 3 cycle datasets

* I would suggest taking out the statement about your coding preference since there are multiple authors and nobody will know who this is referring to. Also not really needed.

# Causal roles

## Z-bias

* I think this whole section could benefit from additional discussion/explanation of the results adjusting or not adjusting for the instrumental variable and what should or should not be done in an analysis in a scenario as presented in the DAG.

## Collapsibility

### Conditional and crude OR

* I was expecting some conclusions at the end of the section, similarly to the previous section. Something along the lines of the fact that x and y are the same/different means that z is/is not a collapsible measure
* Same as above

# Prediction ideas

## Data splitting

* Typo: “Splitting” in title should be written with two t’s
* Typo: Subsection on loading data title should be “load data an**d** files”
* Typo: Same typo on splitting in subsection title (data splitting to avoid model overfitting)

## Bootstrap

* May be good to briefly state what sampling with replacement means?

### Method boot632

* Typo last sentence of first paragraph should read “…boot632 addresse**s** the bias that is due to sampling with replacement”

# Survey data analysis

## CCHS: Bivariate analysis

### Proportions and Design Effect

* Formulas/notation: hat is off-centre

## NHANES: Blood Pressure

### A note about Predictive models

* Suggest having consistent capitalization in titles throughout
* At the end of the first paragraph, not sure what is meant by “In the tutorial you mentioned” – who mentioned a tutorial? The reader? Where/when was this mentioned?

# Missing data analysis

## Imputation

* I didn’t see it, but perhaps it was mentioned somewhere – I would make sure to mention that the next section will show steps for imputation in survey data

### Convergence & diagnostics

* Could be helpful to have a brief statement on what convergence actually means/shows us – why is it important?
* It would also be helpful to see side-by-side examples of good convergence vs. no convergence so students know what to look for.
* For the diagnostic density plots, would maybe be good to include a legend explaining the blue line is the observed data line, similarly for the box plots since the observed data is labeled as “imputation number 0”, I think.

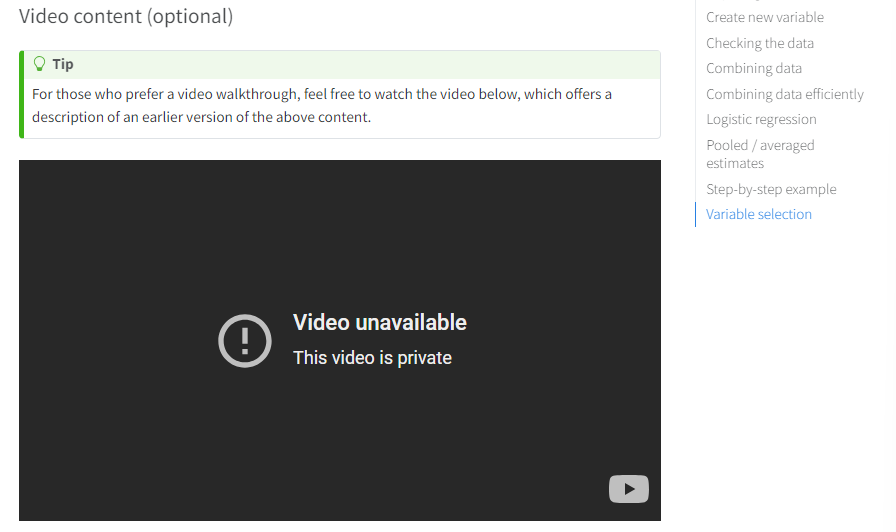
## Imputation in NHANES

### Create new variable

* Perhaps mention that such thresholds should be based on the data (i.e. median, quintiles, etc.) or on the literature (clinically relevant).

### Video content

* This video is private – students will not be able to view!



# Propensity score

* General suggestion to use consistent formatting such as capitalization in subtitles

## Exact Matching (CCHS)

* Typo: “Pr**e**liminary analysis” should be the subheading

# Machine learning

* Same typo as before with data splitting