Introduction to EPIB 704

Mabel Carabali

EBOH, McGill University

EPIB 704

(updated: 2024-08-28)

Outline

1. Presentations

1. Course description

1. Reproducible research

1. Descriptive epidemiology

Who are we?

Mabel

- Social and Infectious Diseases epidemiologist
- ***"Recovered"*** clinician; **ES(T)L**
- ***Twitter/FB/IG-less***
- Website

Rina

Edgar

Who are you?

Course description

Some medical research bingo we will cover

unclear analysis aims causality, prediction or description: it makes a difference	**evidence of absence fallacy** not significant = no effect? if only things were that easy	**data dredging** "look what I found that turned out significant"	**noisy data fallacy** that what doesn't kill it makes it stronger? i wish!
dichotomania categorize everything, chapter 1 of data torture handbook	**table 2 fallacy** interpret each regression coefficient as adjusted for confounding? yeah, that is probably wrong	**regression to the mean** yes, measurements at follow-up are different than at baseline. doesn't mean much	**ignoring dependent observations** the data are probably nested, and that is important
poor reporting yes, there are reporting guidelines for that	**small sample size** gets these confidence intervals nice and wide	**ignoring missing data** can't miss what you never had? think again	**data driven variable selection** post selection inference, yikes!
only apparent predictive performance that is optimistic. could have tried a cross-validation or a bootstrap, perhaps?	**point estimate is the effect** why did you even bother calculating the confidence interval?	**collider** is it a cause? is it a confounder? no, it is a collider here to ruin your inferences!	**multivariate model** what you mean to say is: multi*variable* model

Course description - more specifically

- Emphasis on epidemiologic theory and the estimation of epidemiologic effect measures and uncertainty in different study designs
- Emphasis on causal inference concepts and its distinction from, and yet simultaneously dependency, on statistical models
- Emphasizes the limitations of mainstream null hypothesis statistical significance (NHST) paradigm
- Encourages the recognition of the value of an estimation (Bayesian) paradigm for optimal inferences

Course description (2)

- “Hands-on” approach whereby the various key statistical concepts covered will be illustrated by computer coding, the new calculus for modern epidemiologic methods
- Counting has historically been the essential background to epidemiologic research and remains so in the 21^{st} century but is now often best accomplished by simulation and sampling of posterior probability distributions
- Examples of the required computing code will be extensively provided.

Course description (3)

- Attention will also be given to exploratory data analysis (tabular and graphical), data interpretation, critical examination of assumptions and reproducible research
- Focus mainly on cumulative incidence measures for categorical outcomes, with attention to model checking and screening for confounders and effect measure modifiers
- Regression models covered will include linear, logistic, Poisson, survival, & meta-analytical (hierarchical) models
- Experimental and quasi-experimental designs will be discussed
- Miscellaneous topics will include attributable fractions, selection bias, sensitivity analyses, bootstrapping, matched data, missing data, and misclassified data

Reference textbooks

1) ***What if?*** by Miguel Hernán and James Robins, available here (Chapters 1-11)

2) ***Modern Epidemiology*** by Timothy L. Lash, Tyler J. VanderWeele, Sebastien Haneuse, Kenneth J. Rothman, available at the McGill Library here

3) ***Regression and Other Stories*** by Andrew Gelman, Jennifer Hill, and Aki Vehtari, available here

Optional deeper Bayesian dive consider the excellent

4) ***Statistical rethinking: a Bayesian course with examples in R and Stan*** by Richard McElreath available at the McGill Library here (Chapters 1-10)

- Associated website: https://xcelab.net/rm/statistical-rethinking/ and YouTube lectures here

Other readings will be assigned each week from published journal articles

Computing language

The computing language of choice for the course is `R`.

There are several reasons for this choice including its open source and rich online community which means help is often only a `Google` away.

We believe `R` has become the *lingua franca* for much of the epidemiology/biostatistical universe. Of course, other languages including `Julia` or `Python` will give the same results but unfortunately, we can't supply the necessary support for those languages.

`Stata` is a popular proprietary software, but we consider the scripting and reproducibility offered by `R-Markdown` provides additional advantages for its choice.

R - Related resources

We understand that not everyone is familiar with R as a computing language and therefore **strongly recommend** starting, reviewing and or strengthening your knowledge on the software. This will make your experience much more pleasant and will greatly contribute to the learning experience during this course and throughout your PhD training.

Here we provide a list of resources to access before and throughout the academic term.

In-person: McGill's Computational and Data Systems Initiative R Summer Camp: August 19 to 23, 2024. Registration here.

Online

- R-Studio Education: Independent learning (https://education.rstudio.com/learn/)
- R-Studio/posit: Practice-based learning (https://posit.co/products/enterprise/academy/)
- R-related Books: https://www.rstudio.com/resources/books/
- Princeton site to explore R: https://exploringr.princeton.edu/self-learning-resources-for-r/
- R-resources list: https://thatdatatho.com/r-resources-beginner-advanced/

Course Structure

- Lectures
- Assignments (submitted through *mycourses*)
- Rapid Reviews

Assessment/Evaluation

	%	Rubrics
Attendance & Participation	10%	Presence, attention, active participation 📲 😴 🕳️
Rapid Review	18%	Critical appraisal, time management ⏰ 📝 🔬
Homework Assignments (6x12%)	72%	Accuracy, demonstrated skills, timely delivery 🗒️ ⏰
	100%	

Attendance & Participation

We expect to see everyone **in class in-person** . We consider that this will facilitate the learning experience and will provide everyone the opportunity to interact as colleagues.

We understand that life happens and therefore will enable *Zoom* attendance **only under special circumstances**.

Attendance will be graded for a total of 6 points to make up for the 6% of the total grade. We expect you to attend class, pay attention to the lectures and your classmates' presentations, and to participate actively of the in-class discussion, which is another reason to **favor and encourage in-person attendance** . Doing the suggested lecture **readings BEFORE the lectures** will assure an optimal learning experience.

*To safeguard the well being of everyone, **if you present respiratory-like symptoms or any signs or symptoms of communicable conditions** that may endanger your health, the health of classmates or instructors, **please consult a medical professional and inform the instructors as early as possible** to make the required accommodations.*

Please see McGill's guidelines for attendance: Guidelines and Policies

Rapid Review

Every enrolled student must sign up for one *rapid review* which is a 10-minute presentation of a **methodological critique** about any new paper (**publication date since July 2024**) published in the biomedical literature (i.e., indexed in PUBMED).

- Presenters should comment on study design, validity, precision, causal interpretation, modeling strategy and other concepts from the course.

The comments can be praise or critique but **must** demonstrate thoughtful application of the principles discussed in the lectures as applied to new work in the field.

The rapid review will be graded out of 18 points to make up for the 18% of the total grade.

Guidelines for the *rapid review*

- **Selection of the document:** Identify the published manuscript considering the date of publication, your personal/thesis interest and the methodological approach. **[6/18 total points]**

Consider the **focus on methods**, consider exploring new methodological approaches or novel use of existent epidemiological methods.

- E-pubs or online publications ahead and *pre-prints* are allowed.
- Please provide a pdf of your target paper **at least one week before the presentation** to be shared with the class.
    - **All students** must participate in the review of the paper through the Perusall platform. **Details bellow.**
- You must have received an email to access the EPIB 704 Perusall platform. However, you could also access it using the code: `CARABALI-CRLM6`, or accessing it through this link

Guidelines for the *rapid review* (2)

- **Critical appraisal:** Please read and critically appraise the selected manuscript considering key aspects of the design and analytic plan.

To illustrate your understanding of the selected document, **you must re-write the abstract**, as follows:

- Write a 300-words (maximum) structured abstract: Background, Objectives, Methods, Results, and Conclusion.
- Provide a **50-words** (maximum) section of **strengths and limitations**.
- Finally, provide a brief statement (50-words) indicating if the methodological approach used in the reviewed paper is something that you will do, or something that you would like to do, or something that you would definitely not do, and why.
- Submit the **written document (max 500 words total)** the day of your review through myCourses's rapid review section. **[6/18 total points]**

Guidelines for the *rapid review* (3)

- **In-Class presentation format:** In a maximum total time of **10 minutes** present to the class the results of your review **[6/18 total points]**, considering the following:
  - Power point / slides or any visual aids are welcome, but there **should not be more than 5 slides/pages**.
  - After each presentation there would be a ***brief period for Q&As (approximately 5 minutes)***.
  - Interactive discussions are highly encouraged but a **maximum of 15 minutes** are allocated for each presentation and discussion in every class.

Dates assignment

select an article and a date (last date Nov 28) [1] and **inform the instructors by September 19, 2024** 1st come 1st served.

Date	Name	Article/Manuscript
24 Sep	TAs	TBD
26 Sep		
01 Oct		
03 Oct		
08 Oct		
10 Oct		
22 Oct		
24 Oct		
29 Oct		
31 Oct		
05 Nov		

Homework Assignments

The homework will make 72% of the total grade. There will be 6 homework assignments (each worth 12%) throughout the 13 weeks of the semester, which will be graded by the TAs. The schedule of dates for when these will be assigned and handed in and handed back is as follows:

HW #	Assigned	Due	Returned
1	Sept 10	Sept 17	Sep 24
2	Sept 24	Oct 01	Oct 08
3	Oct 08	Oct 22	Oct 29
4	Oct 24	Oct 31	Nov 07
5	Nov 07	Nov 14	Nov 21
6	Nov 21	Nov 28	Dec 3 (end of term)

Howework Grading

Some collaboration in homework assignments may be beneficial but please use good judgment in preventing your collaboration from becoming detrimental to your learning of the material.

Submitted assignments should be your **individual effort**, even if you consult with other students about your strategy for obtaining these solutions (see plagiarism note below).

To reinforce the concept of scientific **reproducibility** all assignments should be submitted as `R-Markdown` files. Data files for the assignments are available on GitHub and may be installed directly into `R`.

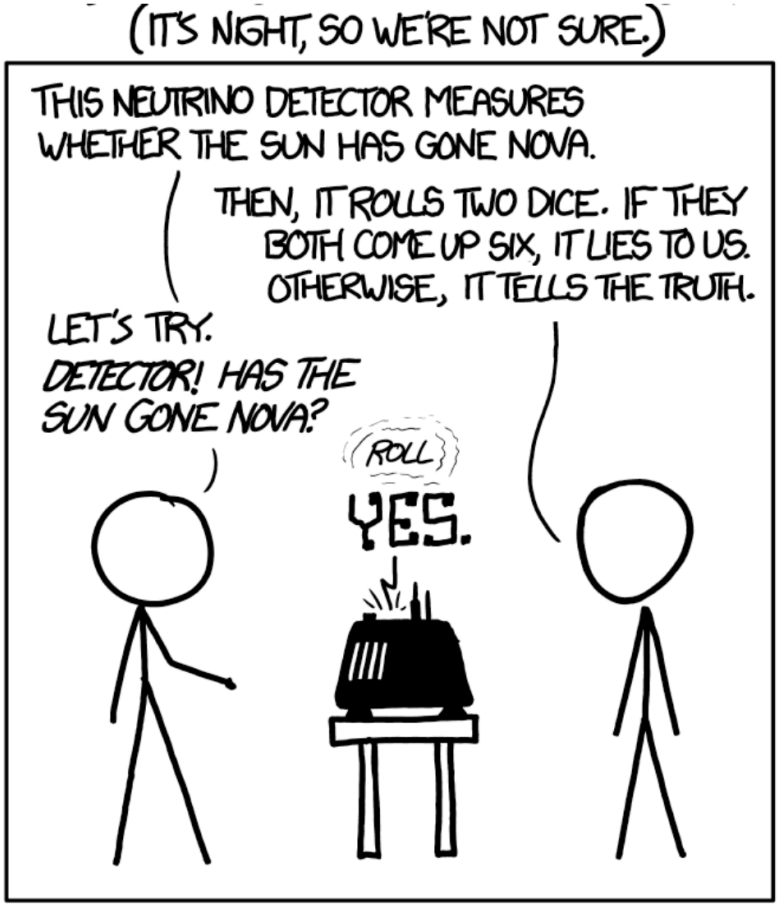
Late assignments carry a very severe penalty of **10% off per day late**. This is primarily to protect the time of the TAs.

Pleas of mercy for extenuating circumstances will be accepted only with written documentation. Taking extra time to do a better job is probably not a worthwhile strategy, because the late penalty is so costly.

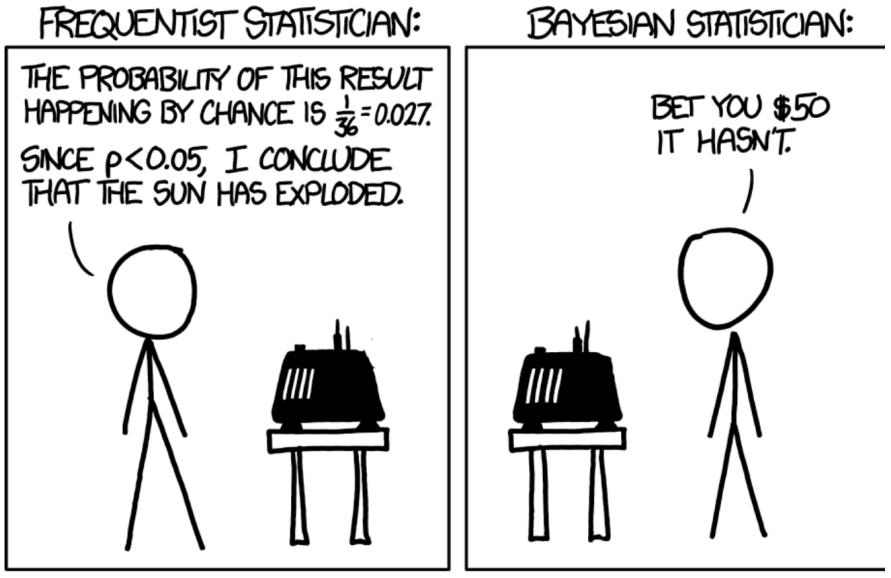
Some Admin & Logistics

- Important dates (**McGill**): Add/Drop class: Sep 10th, 2024
  - Key dates: Assignments and Rapid Reviews
- Communication channels: *mycourses* & emails : Be Mindful of time schedules, office hours
  - We all have other activities outside the classroom
- Course Evaluations : Feedback throughout the course is MORE THAN WELCOME!
- Academic Integrity : Plagiarism, misconducts, & any form of discrimination, aggressive behavior or harassment towards ANY member of the community is **UNACCEPTABLE**
- Other resources at Student Wellness Hub : FEEL FREE TO ASK!

A little bit of Bayes



A little bit of Bayes



Descriptive Epidemiology

- What is Descriptive Epidemiology?
- How much bias is *acceptable*? (👀)
- Descriptive vs. Causal (🧐)

Matthew P Fox, Eleanor J Murray, Catherine R Lesko, Shawnita Sealy-Jefferson, On the Need to Revitalize Descriptive Epidemiology, American Journal of Epidemiology, Volume 191, Issue 7, July 2022, Pages 1174–1179, https://doi.org/10.1093/aje/kwac056

Not every research needs to be causal,

but everything should be robust (❗)

Framework for descriptive Epidemiology

1. Research Question and Background

2. Population and Data Source

3. Outcome Ascertainment/Covariates

4. Analysis Plan: Targeted Measure of Occurrence

5. Results (Presentation of)

6. **Bias**, **Limitations** and **Interpretations**

Where are we on our Epi knowledge?

Baseline Assessment

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

Other resources

Reproducible research

Can you reproduce a project that you completed 3 weeks ago?
How about 3 months ago?
3 years ago?

What if your data changed, even minimally, could you easily redo your analyses?
How long would it take you to update your manuscript?

Getting setup and starting a project in R Studio

1. Put all the files you received into a folder and give it a name (e.g., `704_course`)
2. Choose `File > New Project`
3. Associate the project with the course folder (e.g., `704_course`)
4. From now on, you can open your project by clicking directly on the `.Rproj` file in that folder (or using the drop-down menu on the top right)
5. Now create an R Notebook file (`File > New File > R Notebook`)

A note on computer code and syntax

- There are two main "dialects" of R: base R and tidyverse
- I generally use tidyverse syntax
- There are three main differences:
  - The pipe `%>%`
  - `dplyr` verbs (e.g., `select()`, `filter()`, `mutate()`)
  - graphing with `ggplot2` (but this antedated the Tidyverse)
- However a case can be made that base R is kinder, gentler and more efficient
- This course is not about learning computer syntax so plenty of sample code will be provided
- It is about understanding what code to use and understanding its output

Tidyverse examples

- Instead of `tidy(lm(y ~ x))`, type `lm(y ~ x) %>% tidy()`
- For data management, you can easily filter and select columns:

```
data_subset <- data %>%
  filter(female == 0) %>%           # keep female respondents
  select(income, education) %>%     # keep two variables
  mutate(educ_2 = education^2) %>%  # create edu^2
  drop_na()                         # listwise delete missings
```

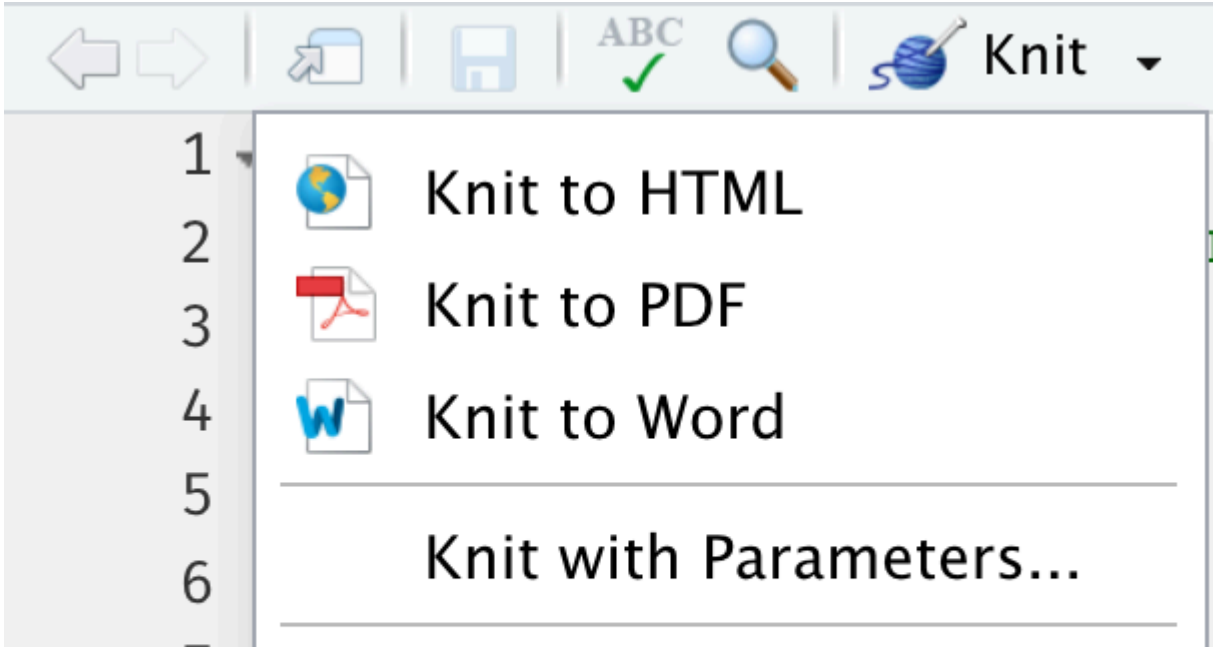
A few useful keyboard shortcuts

Object/Action	Windows	MacOS
<-	Alt + -	Option + -
%>%	Ctrl + Shift + M	Cmd + Shift + M
New code block	Ctrl + Alt + I	Cmd + Option + I
Run highlight	Ctrl + Enter	Cmd + Enter
Clear console	Ctrl + L	Ctrl + L

R can help reproducibility

Can integrate code and text.
Uses RMarkdown which is simple intuitive language which can be out put to different formats

- HTML
- pdf
- LaTex
- Word



What is R Markdown?

1. "An authoring framework for data science." (✔)
2. A document format (`.Rmd`). (✔)
3. An R package named `rmarkdown`. (✔)
4. "A file format for making dynamic documents with R." (✔)
5. "A tool for integrating text, code, and results." (✔)
6. "A computational document." (✔)
7. Wizardry. (🧙)

R Markdown

`R Markdown` depends on `knitr` and `Pandoc`

`knitr` executes the computer code embedded in Markdown, and converts R Markdown to Markdown

`Pandoc` renders Markdown to the output format you want (such as PDF, HTML, Word, etc)

— R Markdown: The Definitive Guide
— `RStudio` has a Markdown quick reference drop down help menu

Basic R-Markdown anatomy

1. The **metadata (YAML)**

2. The **text**

3. The **code**

4. The **output**

Gentle tutorial

R Markdown for writing reproducible scientific papers

R Chunk

```
```{r echo=TRUE, results='hide'}
glimpse(mockdata)
```
```

default options

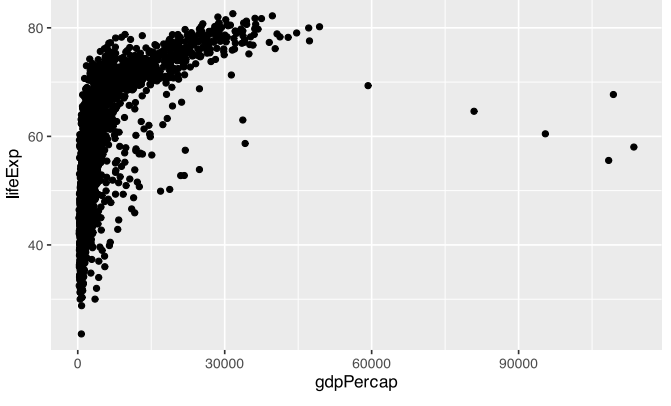
```
## List of 54
##  $ eval         : logi TRUE
##  $ echo         : logi FALSE
##  $ results      : chr "markup"
##  $ tidy         : logi FALSE
##  $ tidy.opts    : NULL
##  $ collapse     : logi FALSE
##  $ prompt       : logi FALSE
##  $ comment      : chr "##"
##  $ highlight    : logi TRUE
##  $ size         : chr "normalsize"
##  $ background   : chr "#F7F7F7"
##  $ strip.white  : 'AsIs' logi TRUE
##  $ cache        : logi FALSE
##  $ cache.path   : chr "L1_EPIB704_v0_cache/html/
##  $ cache.vars   : NULL
##  $ cache.lazy   : logi TRUE
##  $ dependson    : NULL
##  $ autodep      : logi FALSE
##  $ cache.rebuild: logi FALSE
##  $ fig.keep     : chr "high"
##  $ fig.show     : chr "asis"
##  $ fig.align    : chr "center"
##  $ fig.path     : chr "L1_EPIB704_v0_files/figu
```

Visualizations

Great way to understand and share insights into your data
R is great for visualizations - either base R or tidyverse with **ggplot2** package

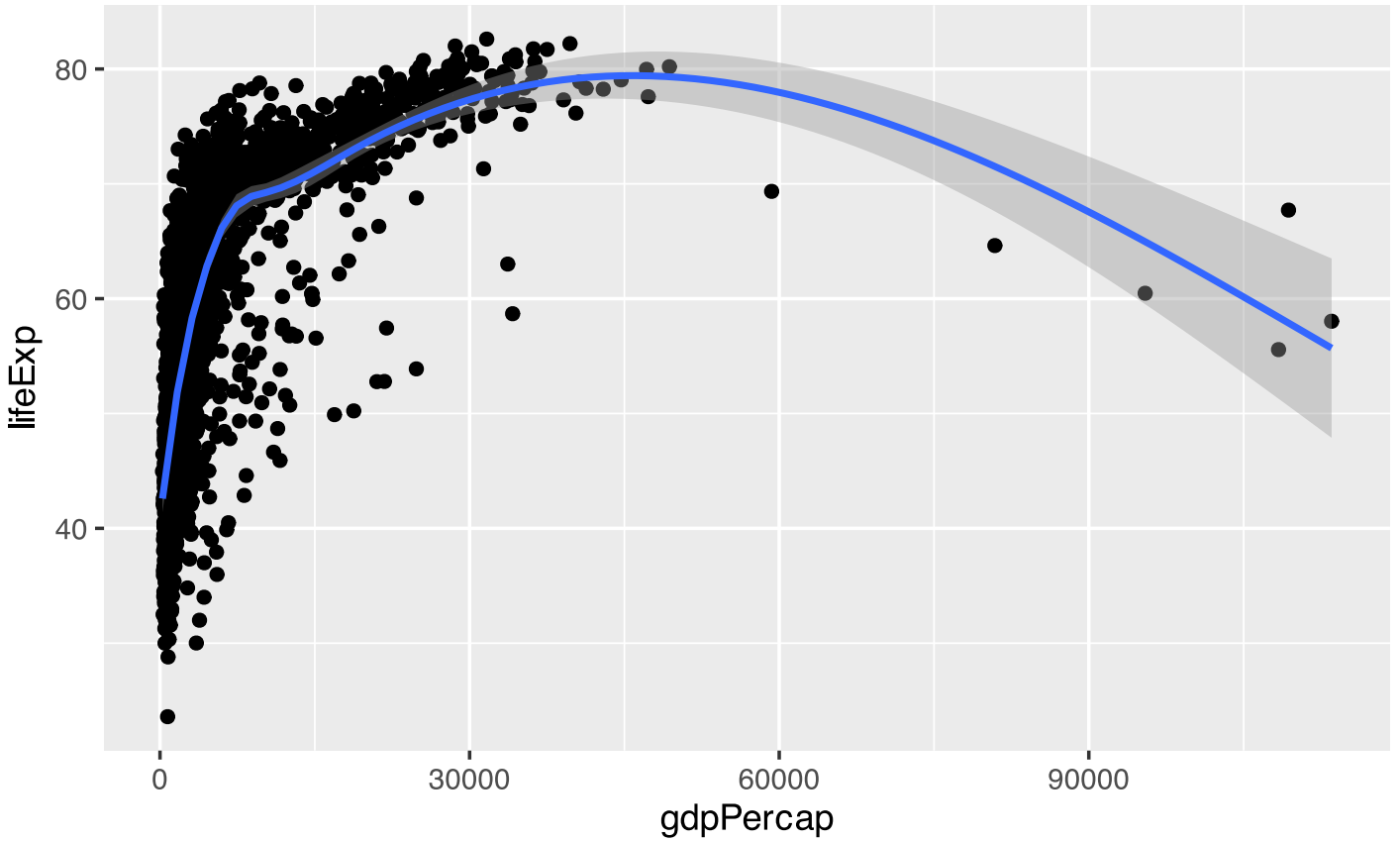
```
library(gapminder)
p <- ggplot(data = gapminder,
            mapping = aes(x = gdpPercap, y
  geom_point()
```

```
p
```



Build plots with layers

```
p + geom_smooth()
```



Increasing complexity

```
p <- ggplot(data = gapminder,
            mapping = aes(x = gdpPercap,
                          y = lifeExp,
                          color = continent
                          fill = continent)
    geom_point() +
    geom_smooth(method = "loess") +
    scale_x_log10()
```

```
p
```