# My amazing title

*Your R. Name*
April DD, 20YY

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

Advisors:
*Advisor F. Name*
*Your Other Advisor*

# Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

# Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1 Introduction

The field of preventive epidemiology involves the identification of potentially modifiable risk factors that contribute to the burden of disease within human populations. Environmental epidemiology, in particular, considers the effect of environmental exposures — chemical or otherwise — which have been increasingly recognized as crucial determinants of human health (Vineis, 2018). Understanding the health effects of exposure to chemical pollutants is especially timely. It has been estimated that human activity releases chemicals at a rate of 220 billion tons per annum (Cribb, 2016). As a result, exposure to low levels of pollutants has become an inevitable peril of daily life. Scholars warn that such conditions of cumulative chronic toxicity pose an acute risk to the wellbeing of humans and our living environment (Naidu et al., 2021). To this, the quantification of such risks through environmental epidemiology studies can prompt critical regulatory action.

Studies concerning chemical pollutants in environmental epidemiology have historically focused on elucidating the effect and mechanisms of single exposures. However, humans are invariably exposed to numerous complex chemical mixtures which together contribute to the progression of adverse health outcomes — risk assessments of single pollutants likely fail to capture the true consequences of these complex exposures (Heys et al., 2016). Assessing mixtures of chemicals can also have more direct implications for public health interventions. The United States Environmental Protection Agency (U.S. EPA) currently passes regulations for individual pollutants.

In practice, though, regulation occurs by controlling the source of pollution, which is responsible for the production of a whole mixture of chemicals with specific joint effects on human health. As a result, the National Academies of Science has advocated for a multipollutant regulatory approach, which is likely to be more protective of human health (Committee on Incorporating 21st Century Science into Risk-Based Evaluations et al., 2017).

Hence, there are clear practical motivations for the development of studies and methodologies that examine the health effects of exposure to co-occurring chemical mixtures, hereafter referred to as exposure mixtures. However, expanding the focus of analysis from one exposure to multiple exposures introduces unique statistical challenges. In addition to the issue of small effect sizes and small sample sizes present in all exposure analyses, multiple exposure analyses must also contend with high-dimensionality, collinearity, non-linear effects, and non-additive interactions (Yu et al., 2022). The classic multiple linear regression framework fails to capture the true effects in this setting. In the past few years, a wide variety of statistical methods have been developed to overcome these challenges (Gibson et al., 2019; Yu et al., 2022), which have been accompanied by a host of comparative simulation studies for general mixture scenarios (e.g., Hoskovec et al., 2021; Lazarevic et al., 2020; Pesenti et al., 2023). However, there is not yet conclusive guidance about the ability of these methods to conduct inference on non-additive interactions between exposures.

The goal of this thesis is to explore the theory of emerging Bayesian regression techniques for quantifying complex interactions between environmental exposures. [**clarify goals**]

In an age where anthropogenic actions have radically reshaped the earth, humanistic inquiry can offer critical insights into our place within a rapidly evolving environment. I begin in Chapter 2 by contextualizing this thesis with a brief overview of cultural and social understandings of the topic of environmental exposures. Chapter 3 provides background on X Bayesian methods for analyzing exposure mixtures. Chapter 4 assesses the performance of these methods for conducting inference on non-additive interactions using a simulation study based on **X**. Chapter 5 explores an application on **X** data [**TBD**]. I conclude with a discussion of the implications of this work for the future study of complex interactions in exposure mixture studies.

# Chapter 2  R Markdown Basics

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).

## 2.1 Lists

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

## 2.2 Line breaks

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.

## 2.3 R chunks

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):

```r
summary(cars)
```

```
     speed           dist    
 Min.   : 4.0   Min.   :  2  
 1st Qu.:12.0   1st Qu.: 26  
 Median :15.0   Median : 36  
```

```
Mean   :15.4   Mean   : 43
3rd Qu.:19.0   3rd Qu.: 56
Max.   :25.0   Max.   :120
```

## 2.4 Inline code

If you'd like to put the results of your analysis directly into your discussion, add inline code like this:

> The `cos` of $2\pi$ is 1.

Another example would be the direct calculation of the standard deviation:

> The standard deviation of `speed` in `cars` is 5.288.

One last neat feature is the use of the `ifelse` conditional statement which can be used to output text depending on the result of an **R** calculation:

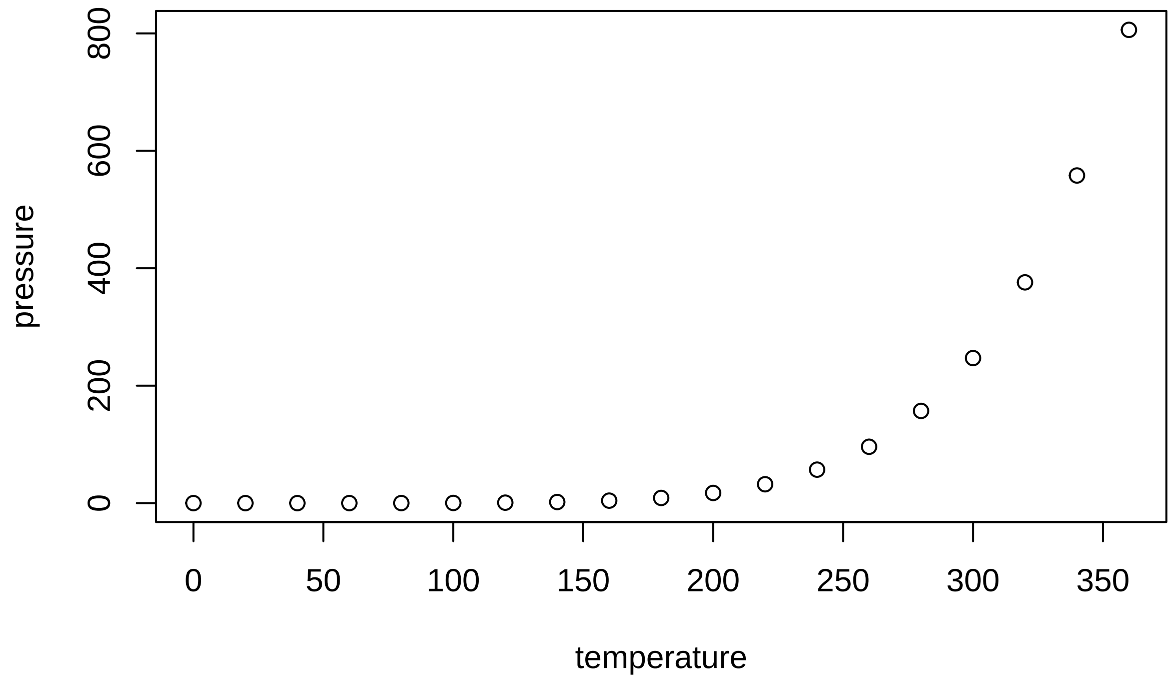
> The standard deviation is less than 6.

Note the use of `>` here, which signifies a quotation environment that will be indented.

As you see with `$2 \pi$` above, mathematics can be added by surrounding the mathematical text with dollar signs. More examples of this are in Mathematics and Science if you uncomment the code in Math.

## 2.5 Including plots

You can also embed plots. For example, here is a way to use the base **R** graphics package to produce a plot using the built-in `pressure` dataset:



Note that the `echo=FALSE` parameter was added to the code chunk to prevent printing of the **R** code that generated the plot. There are plenty of other ways to add chunk options. More information is available at `http://yihui.name/knitr/options/`.

Another useful chunk option is the setting of `cache=TRUE` as you see here. If document rendering becomes time consuming due to long computations or plots that are expensive to generate you can use knitr caching to improve performance. Later in this file, you'll see a way to reference plots created in **R** or external figures.

## 2.6 Loading and exploring data

Included in this template is a file called `flights.csv`. This file includes a subset of the larger dataset of information about all flights that departed from Seattle and Portland in 2014. More information about this dataset and its **R** package is available at http://github.com/ismayc/pnwflights14. This subset includes only Portland flights and only rows that were complete with no missing values. Merges were also done with the `airports` and `airlines` data sets in the `pnwflights14` package to get more descriptive airport and airline names.

We can load in this data set using the following command:

```r
flights <- read.file("data/flights.csv")
```

```
Reading data with read.csv()
```

The data is now stored in the data frame called `flights` in **R**. To get a better feel for the variables included in this dataset we can use a variety of functions. Here we can see the dimensions (rows by columns) and also the names of the columns.

```r
dim(flights)
```

```
[1] 52808    16
```

```r
names(flights)
```

```
 [1] "month"        "day"          "dep_time"     "dep_delay"   
 [5] "arr_time"     "arr_delay"    "carrier"      "tailnum"     
 [9] "flight"       "dest"         "air_time"     "distance"    
[13] "hour"         "minute"       "carrier_name" "dest_name"   
```

```
# read long paragraph file
longtext <- readLines("data/paragraphs.txt")
```

```
Warning in readLines("data/paragraphs.txt"): incomplete final
line found on 'data/paragraphs.txt'
```

```
# display text as vector
longtext
```

```
[1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit,
sed do eiusmod tempor incididunt ut labore et dolore magna
aliqua. Ut enim ad minim veniam, quis nostrud exercitation
ullamco laboris nisi ut aliquip ex ea commodo consequat. "
[2] ""
[3] "Duis aute irure dolor in reprehenderit in voluptate
velit esse cillum dolore eu fugiat nulla pariatur. Excepteur
sint occaecat cupidatat non proident, sunt in culpa qui
officia deserunt mollit anim id est laborum."
```

```
# display text as paragraphs
cat(longtext)
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed
do eiusmod tempor incididunt ut labore et dolore magna
aliqua. Ut enim ad minim veniam, quis nostrud exercitation
ullamco laboris nisi ut aliquip ex ea commodo consequat.
Duis aute irure dolor in reprehenderit in voluptate velit
esse cillum dolore eu fugiat nulla pariatur. Excepteur sint
occaecat cupidatat non proident, sunt in culpa qui officia
deserunt mollit anim id est laborum.
```

```
# display text without linewidth option specified
longtext
```

```
[1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor
[2] ""
[3] "Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu
```

Another good idea is to take a look at the dataset in table form. With this dataset having more than 50,000 rows, we won't explicitly show the results of the command here. I recommend you enter the command into the Console ***after*** you have run the **R** chunks above to load the data into **R**.

```
View(flights)
```

While not required, it is highly recommended you use the `dplyr` package to manipulate and summarize your data set as needed. It uses a syntax that is easy to understand using chaining operations. Below I've created a few examples of using `dplyr` to get information about the Portland flights in 2014. You will also see the use of the `ggplot2` package, which produces beautiful, high-quality academic visuals.

The example we show here does the following:

- Selects only the `carrier_name` and `arr_delay` from the `flights` dataset and then assigns this subset to a new variable called `flights2`.
- Using `flights2`, we determine the largest arrival delay for each of the carriers.

```
flights2 <- flights %>%
  select(carrier_name, arr_delay)
max_delays <- flights2 %>%
  group_by(carrier_name) %>%
  summarize(max_arr_delay = max(arr_delay, na.rm = TRUE))
```

A useful function in the `knitr` package for making nice tables in *R Markdown* is called `kable`. It is much easier to use than manually entering values into a table by copying and pasting values into Excel or LaTeX. This again goes to show how nice reproducible documents can be! (Note the use of `results="asis"`, which will produce the table instead of the code to create the table.) The `caption.short` argument is used to include a shorter title to appear in the List of Tables.

```r
kable(max_delays,
      col.names = c("Airline", "Max Arrival Delay"),
      caption = "Maximum Delays by Airline",
      caption.short = "Max Delays by Airline",
      longtable = TRUE,
      booktabs = TRUE)
```

Table 2.1: Maximum Delays by Airline

| Airline | Max Arrival Delay |
|---|---:|
| Alaska Airlines Inc. | 338 |
| American Airlines Inc. | 1539 |
| Delta Air Lines Inc. | 651 |
| Frontier Airlines Inc. | 575 |
| Hawaiian Airlines Inc. | 407 |
| JetBlue Airways | 273 |
| SkyWest Airlines Inc. | 421 |
| Southwest Airlines Co. | 694 |
| US Airways Inc. | 347 |
| United Air Lines Inc. | 472 |
| Virgin America | 366 |

The last two options make the table a little easier-to-read.

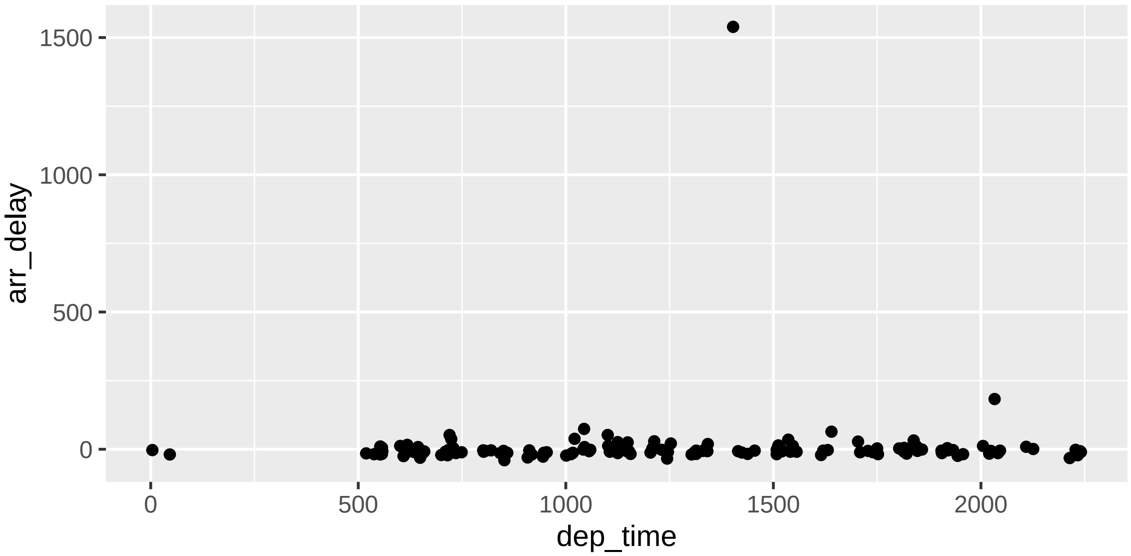
We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

```
flights %>% filter(arr_delay == 1539,
                   carrier_name == "American Airlines Inc.") %>%
  select(-c(month, day, carrier, dest_name, hour,
            minute, carrier_name, arr_delay))
```

```
    dep_time dep_delay arr_time tailnum flight dest air_time
  1     1403      1553     1934  N595AA   1568  DFW      182
    distance
  1     1616
```

We see that the flight occurred on March 3rd and departed a little after 2 PM on its way to Dallas/Fort Worth. Lastly, we show how we can visualize the arrival delay of all departing flights from Portland on March 3rd against time of departure.

```
flights %>% filter(month == 3, day == 3) %>%
  ggplot(aes(x = dep_time, y = arr_delay)) + geom_point()
```



*This is a proof.* There is a proof environment in which you can create equations

$$\hat{\beta}_0 + \hat{\beta}_1 x$$

□

## 2.7 Additional resources

- *Markdown* Cheatsheet - https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet

- *R Markdown* Reference Guide - https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf

- Introduction to `dplyr` - https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

- `ggplot2` Documentation - http://docs.ggplot2.org/current/

# Chapter 3 Mathematics and Science

## 3.1 Math

TEX is the best way to typeset mathematics. Donald Knuth designed TEX when he got frustrated at how long it was taking the typesetters to finish his book, which contained a lot of mathematics. One nice feature of *R Markdown* is its ability to read LaTeX code directly.

If you are doing a thesis that will involve lots of math, you will want to read the following section.

$$\sum_{j=1}^n (\delta\theta_j)^2 \leq \frac{\beta_i^2}{\delta_i^2 + \rho_i^2} \left[ 2\rho_i^2 + \frac{\delta_i^2\beta_i^2}{\delta_i^2 + \rho_i^2} \right] \equiv \omega_i^2$$

From Informational Dynamics, we have the following (Dave Braden):

After $n$ such encounters the posterior density for $\theta$ is

$$\pi(\theta|X_1 < y_1, \dots, X_n < y_n) \propto \pi(\theta) \prod_{i=1}^n \int_{-\infty}^{y_i} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)\ dx$$

Another equation:

$$\det\begin{vmatrix} c_0 & c_1 & c_2 & \dots & c_n \\ c_1 & c_2 & c_3 & \dots & c_{n+1} \\ c_2 & c_3 & c_4 & \dots & c_{n+2} \\ \vdots & \vdots & \vdots & & \vdots \\ c_n & c_{n+1} & c_{n+2} & \dots & c_{2n} \end{vmatrix} > 0$$

## 3.2 Statistics Symbols and Expressions

Exponent or Superscript: $x^2$

Subscript: $x_1, x_2, \dots, x_n$

Both combined: $x_1^{k+1}$.

Our favorite Greeks: $\sigma$, $\epsilon$, $\mu$

Defining a normally distributed random variable: $X \sim N(\mu, \sigma)$

How do we compute sample variance again?

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sometimes you'll need to consider asymptotics, that is, what happens as $n \to \infty$.

## 3.3 Additional information

Many of the symbols you will need can be found on Reed College's math page `http://web.reed.edu/cis/help/latex/math.html` and the Comprehensive LaTeX Symbol Guide (`http://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf`).

# Chapter 4 Tables, Graphics, References, and Labels

# Conclusion

If we don't want the 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

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

## A.1 In the main file 4:

## A.2 In Chapter 4:

# Appendix B The Second Appendix

R code

# Corrections

A list of corrections after submission to department.

Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail. The files samplethesis.tex and samplethesis.pdf show what the "Corrections" section should look like. Questions about what should appear in the "Corrections" should be directed to the Chair.

# References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with OpenGL.* Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with QuickTime.* Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel.* Boston, MA: Wesley Addison Longman.