Introduction to Regression

Stat 230

March 28 2022

Something about me

- First year at Carleton
- Originally from Nepal
- PhD in Applied Statistics from UC-Riverside
- Diverse education background
- Avid learner and traveler

My webpage

COVID-19 related policies

- Stay home when sick. (Even if you don’t have COVID-19, you should stay home if you aren’t feeling well.)
- Follow CDC on testing, quarantine, and isolation.
- Follow the College mask-wearing policy

What will you learn?

- Expressing and interpreting basic models
- Fit and visualize regression models in R
- Assess the model fit with the data
- Understand various modeling goals

What will a typical day/week look like?

Before Class:

1. Some reading/video to introduce some topics
2. Will be updated in the weekly planner

During Class:

1. Mini-lectures
2. Hands-on group activities

How to do well in class?

Be up to date with the class materials

- read the assigned chapters
- watch posted videos

Follow the in-class activity diligently

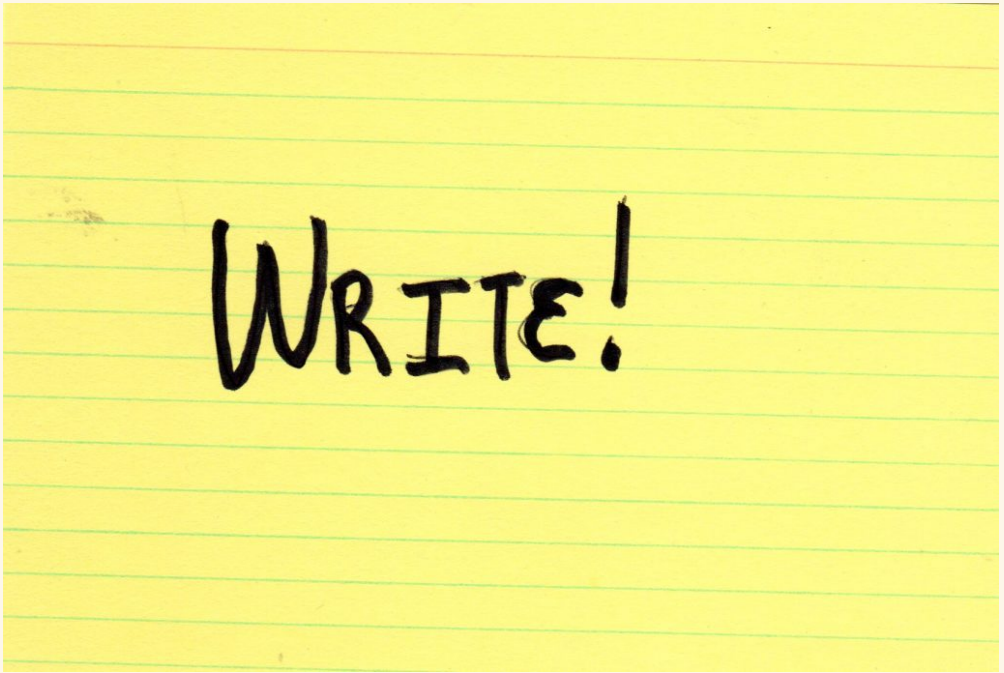
- Try to learn the most here in class
- Do not leave the class with more confusion

Ask questions

- Do not hesitate to ask questions
- I may assume that you understood, if no one ask questions!

Your introduction

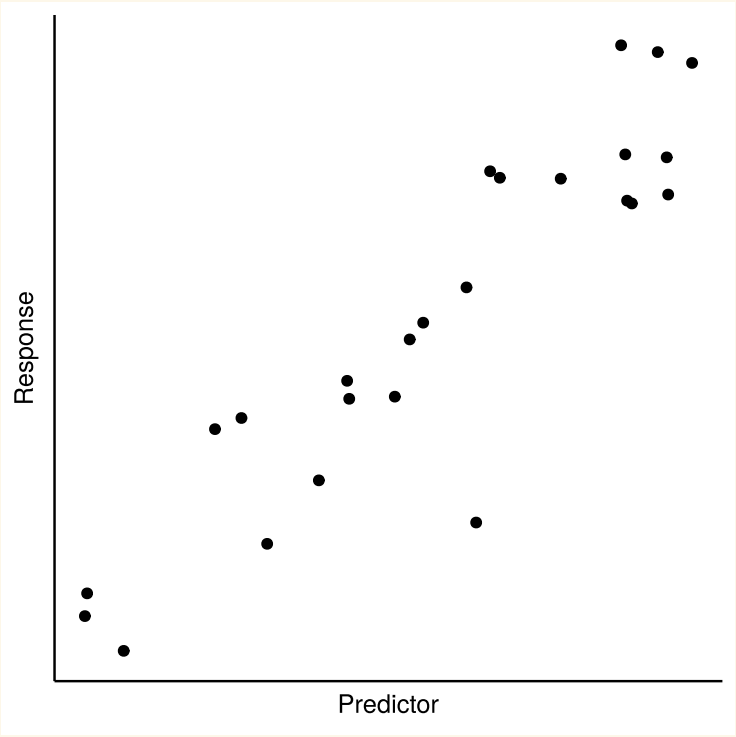
- Your name?
- What gender pronouns do you use?
- Favourite Mathematician/Scientist/Person?
- Recent fun memories?



Please fill in!

Linear Regression

To describe (model) the relationship between a response variable and a predictor variable.



Many ways to refer

- **x-variable:** Predictors, covariates, explanatory variable
- **y-variable:** Response, outcome variable

Code Output Regression

```
ggplot(sim_data) +
  aes(x = predictor, y = response) +
  geom_point() +
  labs(
    x = "Predictor Variable",
    y = "Response",
    title = "A Scatter Plot"
  ) +
  theme(plot.title = element_text(hjust=0.5, size=20, face='bold'))
```

Real-life example: `cars` dataset

- Contains information on speed of cars and the distances taken to stop
- Recorded in the 1920s and contains 50 observations on 2 variables

Show 5 entries Search:

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16

Showing 1 to 5 of 50 entries

Previous 1 2 3 4 5 … 10 Next

Relationship between stopping distance and speed

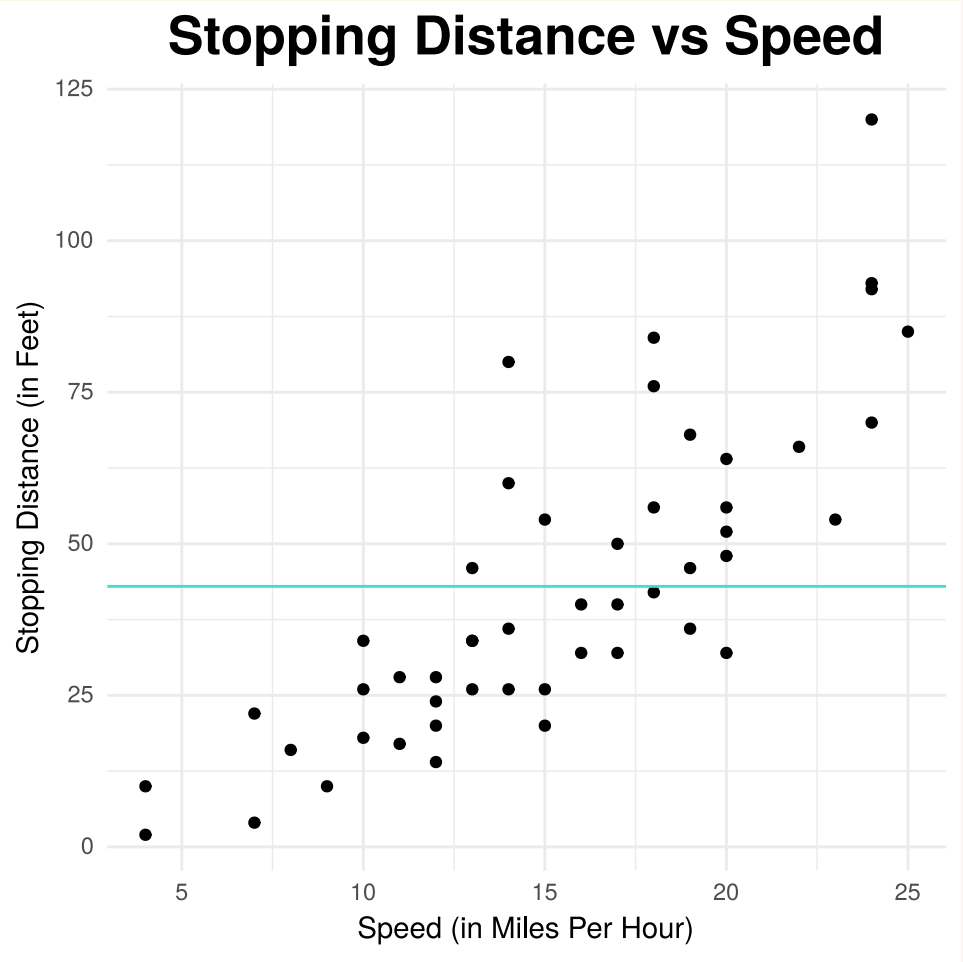
- We have pairs of data (x_i, y_i) for $i = 1, 2, \ldots, n$
- Here i is an index for each of the n observations
- We use x_i as the predictor (explanatory) variable
- The predictor variable is used to help predict or explain the response (outcome)

In our current example, the predictor is the speed and the response is the stopping distance

Why not just model with a horizontal line?

Plot

Explanation



Statistical Modeling

Build a model (Theory)

- model a population's mean response to explanatory variable(s)
- model the error (variation) of the response around the mean at each X

Fit a model to data (Computation)

- Using statistical software like R
- Interactive platform like RStudio

R and RStudio

R is a free statistical software

Widespread use among research statisticians, consultants, and others who use lots of stats in their research

RStudio is an interface with R

- Offers same functionality as R but in a nicer user interface
- Plus has some useful additional features

R is like a car engine and RStudio is like a car dashboard

Your Turn 1

05:00

- Your turn features are regularly used in this class
- This is your time to gauge your learning
- Think of these as lab time under my supervision
- Feel free to sit with your group members next time

Please download the class activity RMarkdown (.Rmd) file from moodle and go over to maize R server. Today, we will briefly go over some basic R codes!!

Basic commands from R console

```
log(2^2) # basic arithmetic (log is ln)
```

```
1:5 # vectors of numbers 1 thru 5
```

```
c(3,1,2) # c = concatenate
```

```
x <- 10:1 # assignment
```

```
x^2 # arithmetic on vectors
```

```
x[c(2,4)] # subsetting
```

Command outputs

```
log(2^2) # basic arithmetic (log is ln)
[1] 1.386294
```

```
1:5 # vectors of numbers 1 thru 5
[1] 1 2 3 4 5
```

```
c(3,1,2) # c = concatenate
[1] 3 1 2
```

Command outputs

```
x <- 10:1 # assignment
```

```
x^2 # arithmetic on vectors
 [1] 100  81  64  49  36  25  16   9   4   1
```

```
x[c(2,4)] # subsetting
[1] 9 7
```

R Objects

Objects are created by many actions in R (assignment, loading data, running certain commands)

Type `ls()` or look in the Environment tab to see objects

R has a number of basic data types

```
str(x) # reveals the structure
 int [1:10] 10 9 8 7 6 5 4 3 2 1
```

Common types: `numeric`, `int`, `factor`, `character`, `data.frame`, `lm`, `logical`

R Workspace (WS) and Working Directory (WD)

Workspace

All objects created in R are stored in the workspace (Environment tab).

Working Directory

- Folder where R looks for a WS, R files, and saves files
- Check your location with getwd()

```
getwd()  # my working directory
[1] "/home/deepak/Desktop/Research/STAT230-Spring/lecture_notes"
```

Make your work Reproducible

It is desirable to make our work reproducible so that anyone with access to our data can easily reproduce the results

Making our work reproducible also means that we can conveniently reuse or modify previous works!

R Markdown helps us make our work reproducible

R Markdown

A R Markdown (.Rmd) file contains the commands and written analysis used to complete a data analysis for an assignment or report.

- The .Rmd file is compiled ("knit") to generate an HTML, PDF, or Word document with **written analysis, R commands and output.**

You can write-up homework and reports as Markdown docs

- Work with simple .Rmd templates, add more complexity to it and knit often!

.Rmd file

A typical .Rmd file needs 3 elements:

- **Header:** info like the title, author, and output type preferences
- **Written analysis:** You can create formatted test elements like bold words, lists, section labels, latex equations, etc. on your final pdf/word/html document.
- **R chunks:** R commands that you want evaluated.

After knitting the file, a pdf/html/Word file will be created in the same folder as the .Rmd file

Practical Tips

When you knit a .Rmd file, Rstudio basically starts a completely new R session behind the scenes with a working directory in the same location as the .Rmd file.

You must have all commands in your .Rmd for your file to knit correctly. Make sure your .Rmd has:

- a data loading command
- any package loading library commands

Knit your .Rmd doc often to catch small errors as they occur