MATH 254 - Syllabus - Fall 2022

Statistical Modeling and Applications

Tural Sadigov

8/23/22

Table of contents

0.1 Course Description

MATH 254 is the second course in statistics with an emphasis on how statistics are used in the real world. Students will learn to form and test hypotheses and models using a wide class of statistical techniques, and special emphasis will be placed on situations where those techniques fail or are misunderstood. Students will learn to implement their ideas in the statistical programming language R. Topics include bootstrap methods, randomization, permutation tests, linear, multilinear, polynomial regression, logistic regression, and selected topics in machine learning. Emphasis will be placed on applications of statistics in real-world data sets and also interpretation & presentation of results.

0.2 Class info

- **Instructor:** Dr. Tural Sadigov
- **Office:** CJ 107
- **Contact:** tsadigov@hamilton.edu
- **Teaching Assistants:** Courtney Shay and Emily Weinstein
- **Class times:** MWF 11:00 – 11:50 am in Ben 201
- **Office hours:** MWF 1:30 pm – 3:30 pm in CJ 107

0.3 Textbooks

- **Textbook - 1:** OpenIntro Statistics

  4th Edition, by D. Diez, M. Cetinkaya-Rundel, C.D. Barr

  https://www.openintro.org/book/os/
- **Textbook - 2:** Introductory Statistics with Randomization and Simulation

  1st Edition, by D. Diez, C.D. Barr, M .Cetinkaya-Rundel

  https://www.openintro.org/book/isrs/
- **Textbook - 3:** Introduction to Modern Statistics

  1st Edition, by Mine Cetinkaya-Rundel and Johanna Hardin

  https://openintro-ims.netlify.app/, https://www.openintro.org/book/ims/
- **Textbook - 4:** Statistics: A Critical Look

  1st Edition, by Chinthaka Kuruwita

  https://stat2-critical-look.netlify.app/
- **Textbook - 5:** An Introduction to Statistical Learning with Applications in R,

  2nd Edition, by G. James, D. Witten, T. Hastie, R. Tibshirani

  https://www.statlearning.com/

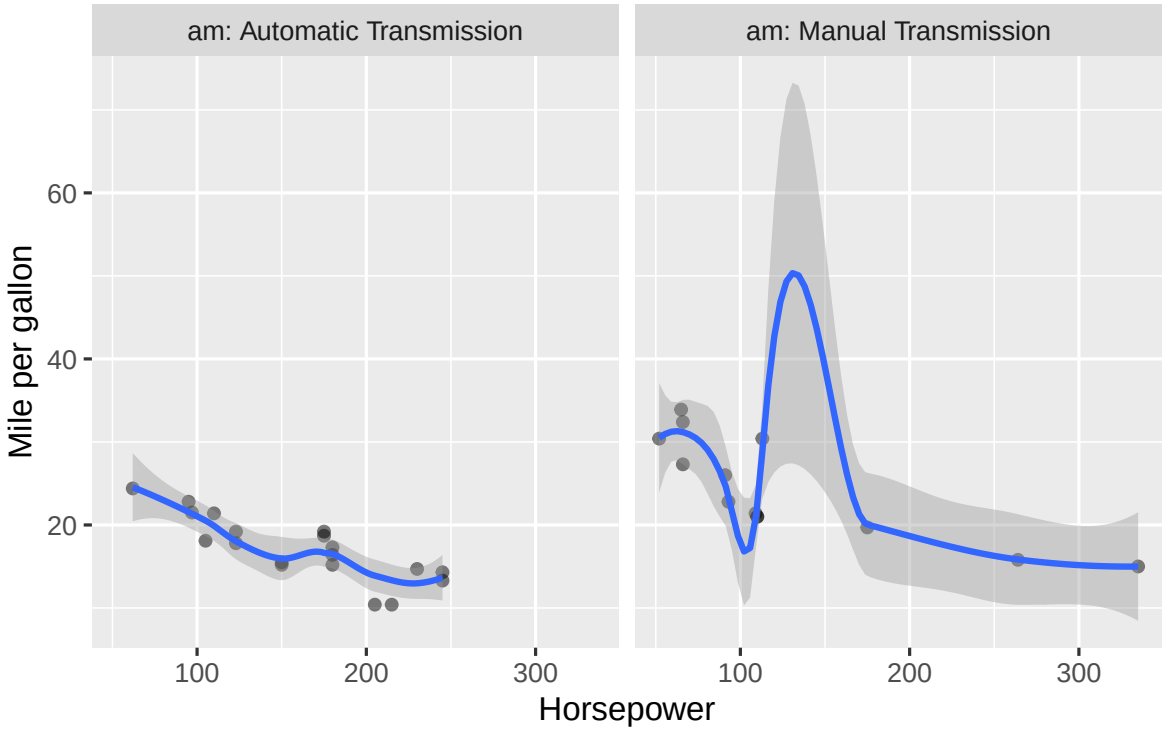
0.4 Software and webpages

- **Software:**
  - R, https://www.r-project.org/
  - R Studio, https://www.rstudio.com/
  - R Studio Cloud, https://rstudio.cloud/
- Blackboard page of the course
- Gradescope page of the course
- GitHub page of the course

0.5 Demo on R/R Studio

You will be writing your lab reports, project, homework solutions and possibly exam papers using Quarto (formerly known as R Markdown) within R Studio. Quarto is an open-source scientific and technical publishing system. It enables you to create reports, journal articles, books, blogs, websites and more, and has the ability to combine text with code. Here is an R code chunk that creates a scatterplot withing this very document.

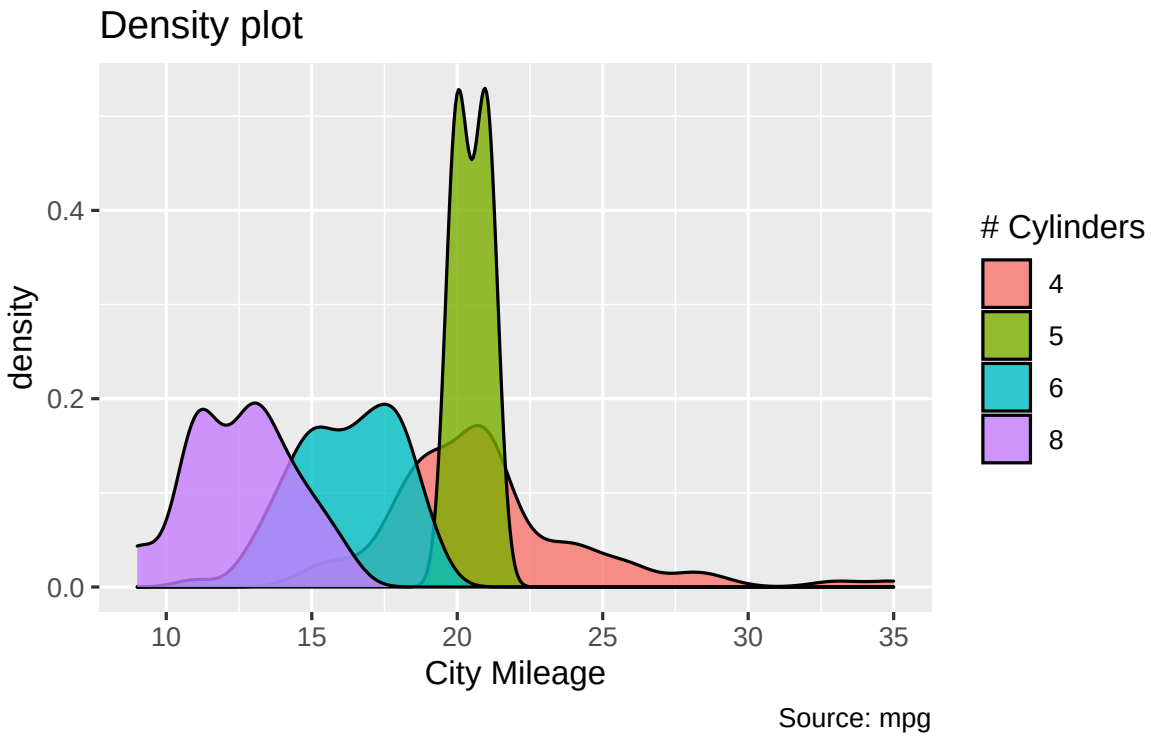
```r
# this is a R example: scatterplot
# if you do not have tidyverse package, install it by uncommenting
# the line below
# install.packages("tidyverse")
library(tidyverse)
mtcars$am <- factor(mtcars$am,
                    labels = c("Automatic Transmission",
                               "Manual Transmission"))
mtcars %>%
  ggplot(aes(x = hp, y = mpg)) +
  geom_point(alpha = 0.5) +
  geom_smooth() +
  facet_wrap(.~am, labeller = label_both) +
  xlab("Horsepower") +
  ylab("Mile per gallon")
```



Here is an example of smoothed histograms (densities).

```
library(ggplot2)

# Plot
mpg %>%
  ggplot(aes(cty)) +
  geom_density(aes(fill=factor(cyl)), alpha=0.8) +
  labs(title="Density plot",
       caption="Source: mpg",
       x="City Mileage",
       fill="# Cylinders")
```



We can take the previous plot and make it interactive within the document. One can play with it by clicking on the number of cylinders bars (colors) on the right.

```
library(ggplot2)
library(plotly)

# Plot
my_plot <- mpg %>%
  ggplot(aes(cty)) +
  geom_density(aes(fill=factor(cyl)), alpha=0.8) +
  labs(title="Density plot",
       caption="Source: mpg",
       x="City Mileage",
       fill="# Cylinders")
ggplotly(my_plot)
```

We can also write Python chunks in Quarto.

```python
# python code
import pandas as pd
r.iris.head()
```

```
   Sepal.Length  Sepal.Width  Petal.Length  Petal.Width Species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa
```

Note that above, we used data from R within Python! The same code can be written in R like below:

```r
# R code
iris %>%
  head()
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

One can also write SQL queries within Quarto and even create web apps/dashboards with R (R Shiny) and embed them into a Quarto document. See an example of a web app created with R: https://turalsadigov.shinyapps.io/my_new_ames_app/.

0.6 Topics

We have two modules in this class.

0.6.1 Module - 1 - Classical Statistics

This module covers fundamentals of classical statistics.

1. R/R Studio/Quarto

2. Statistics: cases, variables, data-matrices, experiments, numerical and graphical summaries
3. Sampling distributions, Central Limit Theorem (review)
4. Hypothesis testing, decision errors, Common statistical tests (z, t) (review)
5. Bootstrap and randomization
6. Permutation tests
7. p-values, dicing and slicing
8. χ^2 tests

0.6.2 Module - 2 - Machine Learning

In this module, we start working with both inference and prediction.

9. Linear, multiple, polynomial regression
10. Model assessments, Diagnostics
11. Variable selection (forward, backward, mixed selections)
12. Bootstrap methods for regression
13. Classification, Logistic regression
14. K nearest neighbor classification
15. K nearest neighbor regression

0.7 Schedule

Table 1: Tentative Schedule

Dates	Topics/Events
	`MODULE 1 - CLASSICAL STATISTICS`
August 26	Introduction/Syllabus
August 29 - Sept. 2	Statistics-1 Review, R, Sampling distributions
September 5 - 9	Central Limit Theorem, Confidence Intervals
September 12 - 16	Hypothesis testing, Common statistical tests (z, t)
September 19 - 23	Bootstrap and randomization
September 26 - 30	Permutation tests
October 3 - 7	Midterm Review, p-values, dicing and slicing,
October 6	Midterm Test

Dates	Topics/Events
October 10 - 14	χ^2 tests
	`MODULE 2 - MACHINE LEARNING`
October 17 - 21	Linear and multiple linear regression
October 24 - 28	Linear, multiple, polynomial regression
October 31 - Nov. 4	Variable selection, forward/backward/mixed selections
November 7 - 11	Classification, Logistic regression
November 10	Project
November 14 - 18	Multiple Logistic regression, K nearest neighbor classification
November 19 - 27	Thanksgiving recess
November 28 - Dec. 2	K nearest neighbor regression
December 5 - 9	Labs all week, k-means clustering, Final Review

0.8 Grading Categories

There are six main grading categories.

0.8.0.1 Homework (10%)

Weekly homework of around 5-7 problems will be assigned on each Monday, and it will be due on next Monday, 5 PM. You will have access to the homework through Blackboard, and you will submit your homework on Gradescope. We have two graders who will grade your work. Lowest HW grade will be dropped at the end of the semester.

0.8.0.2 Labs/Mini Projects (10%)

Weekly labs (a.k.a mini projects) will be given at the end of each Wednesday class. These mini projects will have questions based on the new material we cover in that week. You will be able to work on some of them right away using what you have learned on Monday and Wednesday, and you will continue working on it during Friday's class (the last 20 minutes). Note that you will be working in a group of 2 people on each lab, and you are allowed to use the book, your notes and R. Deadline for submitting the labs is Friday, 5 pm. No make-up lab will be given. Missed labs are counted as zeros. The lowest lab score will be dropped at the end of the semester.

0.8.0.3 Midterm Test (25%)

There will be single midterm test during the semester. When grading tests, partial credits will be given for answers that are mostly, but not completely correct. These tests will test your

understanding of the concepts covered in the class and your ability to solve problems. You will need to use R during the midterm test.

No make-up tests will be given unless there is a convincing reason which will need to be supported by necessary documentations and which I would be informed about at least one week in advance.

0.8.0.4 Project (20%)

At the beginning Week of October 10th, you will be assigned to a group of 2 students to work on a project that will last 4.5 weeks. You will choose a data set you would like investigate from many available resources that will be provided to you, and start analyzing the data using descriptive statistics first and then applying the techniques you have been learning and that are appropriate your research question(s). It is essential that your project have a story and flow. The Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) and the American Statistical Association conduct yearly competition called Undergraduate Statistics Project Competition (USPROC). Guidelines on how to write a project report will be the same ones from USPROC. USPROC has two separate competitions: The Undergraduate Class Project Competition (USCLAP) and The Undergraduate Research Project Competition (USRESP). You will focus on the guidelines for USCLAP. A report template is also provided to you at https://www.causeweb.org/usproc/report-template. Here is the list of Fall 2019 project winners and their projects: https://www.causeweb.org/usproc/usclap/2019/fall/winners. Run the data set with me before you start working on the project. Note that you will be working on the project with your team, and I will serve as a mentor.

0.8.0.5 Project Checkpoints (5%)

Starting with the week of October 10th, each week that project is NOT due on, you will submit a draft of your work as a checkpoint for your project during that week, and get some feedback. These are drafts of your project, and show your progress. In other words, the project has 4 checkpoints and final subsmission:

- **October 12 - Wed - 5 pm:** Checkpoint 1 would have the data set, a draft with a research question and evidence of data read into R Studio.
- **October 19 - Wed - 5 pm:** Checkpoint 2 would have a draft with the research question, some descriptive analysis and references section started.
- **October 26 - Wed - 5 pm:** Checkpoint 3 would have a draft with detailed introduction/background, methods that would be applied and some results.
- **November 2 - Wed - 5pm:** Checkpoint 4 would have some main results, updated methods that applied and discussion, and detailed references and appendix.

- **November 9 - Wed - 5 pm:** Final submission would have everything outlined in ttps://www.causeweb.org/usproc/report-template.

0.8.0.6 Final Test (30%)

There will be a cumulative final test at the end of the semester. Like midterm tests, when grading the final test, partial credits will be given for answers that are mostly, but not completely correct. These tests will test your understanding of the concepts covered in the class and your ability to solve problems. You will need to use R during the final test.

0.9 Grade Scale

Plus and minus grades will be given according to the following table in the inserted image.

A+	=	98 – 100	C+	=	77 – 79
A	=	93 – 97	C	=	73 – 76
A-	=	90 – 92	C-	=	70 – 72
	=			=	
B+	=	87 – 89	D+	=	65 – 69
B	=	83 – 86	D	=	60 – 64
B-	=	80 – 82	F	=	0 – 59

0.10 QSR Center

Assistance to students taking mathematics and statistics courses at Hamilton College is provided by the Quantitative and Symbolic Reasoning Center, located in CJ 303. QSR Center has generous drop-in hours and well-trained tutors. You can get a Peer Tutor to work with you one-on-one at the QSR Center. For more information see https://www.hamilton.edu/academics/centers/qsr.

0.11 Honor code

All of your work (Homework, Labs, Tests, Project) have to be done according to the honor code. The work you submit for homework, labs, and tests must be your own. For homework you will probably find it beneficial to consult with other students about the material and this kind of conversation and collaboration is encouraged. An examination, however, must be solely the student's own work. See the website:https://www.hamilton.edu/student-handbook/studentconduct/honor-code.

0.12 Accommodations for Students with Disabilities

Hamilton College will make reasonable accommodations for students with properly documented disabilities. If you are eligible to receive an accommodation(s) and would like to make a formal request for this course, please discuss it with me as soon as possible. You will need to provide Allen Harrison, Assistant Dean for Accessibility Resources (aharriso@hamilton.edu) with appropriate documentation of your disability.

0.13 Mental Health Support

There are times that each of us may feel overwhelmed, anxious, or depressed. There are many resources available on campus to help and support you:

- Counseling Center (www.hamilton.edu/offices/counselingcenter, 315-859-4340) located at 100 College Hill Road offers individual and group therapy, peer counselors, psychiatric treatment, and a 24-hour hotline. If you need immediate assistance, phoning the Counseling Center and selecting option 2 will connect you with a counselor, 24 hours a day, 7 days a week.
- Associate Dean of Students for Student Support, Sarah Solomon (315-859-4463; ssolomon@hamilton.edu)
- Associate Dean of Students for Academics, Adam Van Wynsberghe (315-859-4600; avanwyns@hamilton.edu)
- Your faculty adviser
- Your RA and Area Director in your residence hall

If at any time you feel suicidal or in danger of harming yourself or others, please reach out for support! The Hamilton community cares and is available to help. Campus Safety is available 24/7 for urgent concerns at 315-859-4000.