STA303: Methods of Data Analysis II

Course guide

Prof. Liza Bolton

Winter 2022

Contents

THIS SITE IS STILL IN PROGRESS! The information is not yet official.

1

How to use this course guide

This course guide has been created using `bookdown`. You can download a PDF version of the whole guide with the download button above (down arrow into a tray). You can also put the website into dark mode and changed the font style, if you find s different display preferable.

If you would like a PDF copy of the slides, you can 'Print to PDF' in your browser. Shortcut: Cmd+P or Ctrl+P, and select 'Save as PDF' (or similar).

1.0.1 Communication policy reminder

All content and logistics questions must be asked on Piazza. Personal or private course matters should be emailed to sta303@utoronto.ca. Quercus mail or emails sent directly to teaching team members will not be answered. If you've missed an assessment due to illness or emergency, please fill out the appropriate form.

2

STA303/1002 syllabus

2.1 Key information

Winter 2022, half year, half credit course

Instructor: Prof. Liza Bolton (she/her)

Head TA: Amin Banihashemi (he/him)

✉

sta303@utoronto.ca

✉

Synchronous class meetings Wednesdays L0101 12:10–1:00 p.m. L0201 3:10–4:00 p.m.

Prof. office hours will take place in the hour after class.

TA office hours will be announced on Quercus.

2.1.1 Delivery

- This course will be run completely online until January 31, 2022. After that, *optional* in person components will be offered, subject to public health guidance.
- In person attendance is NOT required.
- Synchronous attendance is not required but some knowledge basket (see assessments) components are only available synchronously.
- All assessments will be completed online.

2.1.2 Pre-requisities

STA302H1/STAC67H3/STA302H5 (or for STA1002, STA1001H1)

It is Department policy to strictly enforce pre-requisites This is not up to the teaching team. You will be removed from the course if you do not meet the prerequisite course requirements. Please **reach out to the UG stats team (ug.statistics@utoronto.ca)** with any questions abour enrolment orbook an appointment during their office hours.

3

Start here!

3.1 Introductions

Hi folks,

Welcome to STA303! We're excited you're joining us on this statistical voyage. I look forward to introducing myself to you in our first class on Wednesday, but for now, there is are basic introductions below for me and our Head TA Amin. Feel free to skip to How this course works, I know there is a lot to read in the module!

Looking forward to a great semester! See you in class on Wednesday.

3.1.1 Professor Liza Bolton, Instructor

Email: sta303@utoronto.ca (Put "[Prof. Bolton]" in the subject line to the email me directly)

Pronouns: she/her

Before moving (back) to Canada in 2019, I had lived more than half my life in New Zealand. (I still mention New Zealand a lot in class…) My current research areas are in statistics education and online learning, as well as health disparities across ethnic groups. I used to run a small consulting company and called myself a Data Ambassador. Why? Well, lots of people are consultants. I even did an internship in management consulting once upon a time. But it wasn't a satisfying title for what I wanted my work with people to look like. I wanted something that focused on the communication and interpersonal side, not just high quality and appropriate analysis. People who aren't confident in their ability to analyse their own data need a go-between, someone who can be an ambassador for their data! While I don't do consulting any more, I love helping students build their technical and professional skills so they can go out into the world and be excellent ambassadors for data themselves.

Last movie I cried in: Kiki's Delivery Service

Favourite food: Corn. Popped, on the cob, in a chip, Mmmmm.

Book most often given as a gift: A Matter of Fact: Talking Truth in a Post-Truth World by Jess Berentson-Shaw

3.1.2 Amin Banihashemi, Head TA

Email: sta303@utoronto.ca (Amin will often be the one responding to your emails)

Pronouns: he/him

I'm a fourth-year PhD student at the Institute of Medical Science. I have been a TA for STA130 in DoSS for the past 3 years and this is my second semester as Head TA of STA303.

My area of research is clinical Neuroscience, something I am passionate about. I analyze images of brain and eye structures in neurodegenerative diseases. I investigate possible associations of these structures with each other and with the ability to remember well and carry out goal-oriented tasks successfully. I love creating reproducible statistical analysis workflows in R. I also like audiobooks, candlelight, and apple pie (which I make myself!)

\includegraphics[width=10d0%]{images/headers/map}

3.2 How this course works

This course is organized into five two-week modules of learning + two one week assessment-focus weeks.

All course material will be made available through this course site in Quercus (any links to outside sites will be found here). Take a moment to familiarize yourself with some of the tools and content areas found in the left navigation bar. You can move through content in a module by selecting the "Next" button at the bottom right of the page. You can also visit the Module section in the left navigation menu so see all the available modules and their contents.

All times listed are 'Toronto time', i.e. Eastern Time. Note that Daylight Savings Time begins Sunday, March 13, 2022. You may find this time converter helpful: https://www.timeanddate.com/worldclock/meeting.html

3.2.0.1 In most modules there be will be:

- A weekly module released on Monday morning.
- A quiz based on the content released in the module due on Tuesday at 6:00 p.m ET.
    - Special note for the Week 1 quiz: this quiz is available until January 26 at 6:00 p.m. ET. (with no penalty) because I know there is a lot to get used to in the first week.
- A synchronous class on Wednesday at 12:00 p.m. ET (L0101) and 3:00 p.m. ET (L0201).
    - Both sessions will be the same, you only need to attend one.
    - Please review the steps for attending a class.
    - Synchronous classes will be recorded. You're expected to watch the recording if you cannot attend live. They will be posted on the course overviewpage.
- Weekly writing Create phase due Thursday at 6:00 p.m. ET.
- Weekly writing Assess phase due Friday at 6:00 p.m. ET.
- Weekly writing Reflect phase due (next) Monday at 6:00 p.m. ET.
- Office hours.
    - Prof office hours will occur after the Tuesday synchronous classes, i.e. 11:10–12:00 p.m ET and 3:00–4:00 p.m., in the same Zoom call.
    - TA office hours: will begin in Week 2 (TBC) and the schedule will be updated here.

3.2.0.2 Students joining off the waitlist

You don't have to submit missed quizzes or writing activities or alert me that you joined the course late, these will be covered by the associated 'best of' policies. See the Syllabus for more information.

If you have a *friend* on the waitlist, they can sign up to receive materials here.

3.3 Hours expectations

While everyone has different work styles and learning needs, I want to provide some guidance around how I expect this course to look for students.

Plan to be doing 6–8 hours of work on STA303 each week. This may be comprised of:

- 1–2 hours on videos and readings
- 1 hour of attending synchronous class or reviewing the recording and activities
- 1–2 hours on knowledge basket assessments
- 1–3 hours on other assessments
- Remaining time attending office hours

3.3.1 Module flow

3.3.2 Communication

- Our course discussion board on Piazza is to be used for all content and administration questions. *Only* sensitive or personal issues/questions should be sent to sta303@utoronto.ca. We reserve the right not to respond to emails that should be Piazza posts.
  - Please ensure all course-related emails include your **UTORID**.
- There are several important forms that you may need if you miss an assessment due to **illness or emergency** or wish to request a **regrade** of an assessment.
- I will use Quercus announcements to share course information and updates. **Please make sure you read these**. I may also occasionally email or Querucs message you about things that relate specifically to you.

3.3.3 To do now

Press the next button below to continue through this module. In the following pages you will:

- Read the Syllabus.
- Join the Piazza discussion board.
- Understand the tools we will be using in this course.
- [Optional] Introduce yourself in theIntroductions discussion board.
- Learn about some of the services and supportsavailable to you as a U of T student.
- Make sure you have a U of T Zoom account https://utoronto.zoom.us/.

Header photo by Andrew Stutesman.

Assessments

4

Assessment overview

There are two special elements of assessment/grade calculation in this course that are important to be aware of in planning your approach to it.

- *Two roads diverged in a yellow wood*[1]*:* You can opt for Path A or Path B to get your final mark. I will calculate your marks under both paths and then assign you the higher of the two as your final mark.
- A-tisket, a-tasket[2], fill up your **knowledge basket**[3]:
  - Personalize which of these assessments you do based on your interests and skills you want to develop.
  - You can 'max out' your basket: just keep putting grades in until you get to 10% (Path A) or 5% (Path B).

Assessment	Path A	Path B
Mini-portfolio	5	0
Portfolio	20	25
Mini-mixed assessment	5	0
Mixed assessment	20	25
Final project	40	45
Learning basket	10	5
Total	100	100

4.1 Graduate student modification (1002H)

There is no difference in the grading scheme or assessment for graduate students enrolled in STA1002, other than an additional 'path' to your final grade where you may opt out of the 'basket' assessments, if you wish. This only applies to graduate students enrolled in STA1002, not to any students enrolled in STA303.

You don't need to advise me of your choice, I will calculate you mark all three ways below and give you the highest mark.

[1] The Road Not Taken, by Robert Frost. But, dear traveller, you can in fact take both roads and then get whichever gives you the better mark. https://www.poetryfoundation.org/poems/44272/the-road-not-taken

[2] A-tisket, a-tasket by Ella Fitzgerald, https://www.youtube.com/watch?v=1bgFkeDLpSI

[3] "Whaowhia te kete mātauranga" is a Māori proverb meaning "Fill up the basket of knowledge". Mātauranga is specifically traditional Māori knowledge. You can listen to the pronunciation here https://www.massey.ac.nz/student-life/m%C4%81ori-at-massey/te-reo-m%C4%81ori-and-tikanga-resources/te-reo-m%C4%81ori-pronunciation-and-translations/whakatauk%C4%AB-m%C4%81ori-proverbs/

Assessment	Path A	Path B	STA1002 ONLY
Mini-portfolio	5	0	0
Portfolio	20	25	25
Mini-mixed assessment	5	0	0
Mixed assessment	20	25	25
Final project	40	45	50
Learning basket	10	5	0

5

FAQs and Errata

5.1 Frequently asked questions

While Piazza is our main class question and answer board, this page will have some static Questions and Answers for frequently asked questions. Use Cmd + F (Mac) or Ctrl + F (PC) to search for keywords on this page.

Additionally, make sure you're familiar with the Syllabus. I've tried to explain as much as I could there.

5.1.1 Course admin

5.1.1.1 STA303 pre-requisites

I didn't take STA302 or an equivalent course, can I still take STA303?
No, sorry. We enforce pre-reqs strictly. This isn't up to me. You will be removed from the course. Please **reach out to the UG stats team (ug.statistics@utoronto.ca)** about questions of this nature.

- Book an appointment during their office hours.
- For grad courses, please email grad.statistics@utoronto.ca.

5.1.1.2 Recorded lectures

Where can I find recorded lectures and how soon can I expect them to be available?

Recorded lectures will be linked on the Course overview page. I aim to have the links up within 24 hours of class. They take some time to process.

5.1.1.3 Sections

Do L0101 and L0201 cover the same materials?

Yes. The only difference is when your synchronous class is. See the below question in "Attending synchronous class".

5.1.1.4 Attending synchronous class

Can I attend the synchronous class for the other section?

Yes!*

*If the number of attendees is getting too close to the call cap of 300 (this is a Zoom license thing), preference will be given to those enrolled in the session and others will be asked to leave. Any Team Up! activity bonuses will also be applied-section/session does not matter.

Why can't I access the class/office hour Zoom meeting?

You must have and be signed in with your University of Toronto Zoom account. Use the 'SSO' (single sign-on) option.

"Will STA303/STA1002 be able to be completed online-only?" Yes.

- **STA303** will be a flipped course, with content delivered online and opportunities for activities both in-person (subject to health advice) and online. All assessments will be completed and submitted online.
- If you are enrolled in an **in-person tutorial**, you will have an option to attend online instead.
- Synchronous attendance is NOT required to pass this course, but being able to attend (online) synchronously at the times in the timetable may make things easier for you.

*"Where is your **office**?"/"Can I come see you in-person?"* At this stage, I cannot offer in-person office hours nor student meetings in my office. Drop-ins are not currently allowed for any instructors or TAs in Statistical Sciences, unless they have told you they have organized another meeting space for this purpose.

5.1.2 Team Up!

5.1.2.1 Troubleshooting & FAQ advice from U of T CTSI

My screen seems to be lagging compared to those of my group members (eg. I'm not on the same question as the driver/members, the answer I chose does not appear on my group members' screens, etc.)

As long as you are logged in, you will receive your grade, so just continue to participate by communicating with your team. Almost always, your device will re-sync with the rest of your team's devices within a minute or two. If not, refresh your screen once. Repeatedly refreshing is not usually helpful!

I've been disconnected. Will my quiz progress be saved? Can I re-join my group?

Enter the Team Up! session again and you will automatically be put back into your group. Your progress will be saved, and you will return to the question you or your group was working on. This is true for the Driver as well as any team member.

How do I send my completed quiz results?

Click the large red "Submit to Quercus" button at the end of the Team Up! quiz.

Can we change our group driver?

You can request a driver change by pressing the red exclamation mark in the top right of your Team Up! quiz.

How is the driver for our group chosen, and how can the driver pass the group ID to others in remote classes?

Group members can decide who will be the Driver (ideally someone with a good internet connection). The Driver can pass the group ID to other members verbally through chat or microphone in breakout rooms.

The Driver of my group has to leave unexpectedly or their device has stopped functioning. How do we proceed with the quiz?

Request for a Driver Change using the red button at the top right of your Team Up! Session. This button only provides help for Driver Changes. You may also need to speak with your instructor or TA.

5.2 Other

5.2.1 References

Can you (Prof. Bolton) write me a reference? I'm desperate!

Please read my personal policy here to get a sense of under what circumstances I could write for you, but basically, a good mark in one class is not sufficient and you need to have at least two 'activities' with me. If you believe you meet my basic criteria, you can request a reference from me here. I will then accept or decline based on the the information provided.

Do you (Prof. Bolton) have any research opportunities available?

Research/work study/teaching assistant opportunities:

- See the Department website. The main round of TA recruitment occurs during the summer but there are occasionally emergency postings.
- Some information about opportunities with me on my website.
- **Reading courses:** STA496/497: Readings in Statistics must be registered for as part of special enrolment during July. I will not be taking on any further students.
  - If you're thinking about the future, I usually consider taking on a small number of STA497 students for a **half-credit, year long version.**
  - There is much more information about my past students, research interests and what a course with me might be like on my website: <https://www.lizabolton.com/reading_courses.html>.

5.3 Errata

ID	Location	Note

Appendix

6

Bits and pieces

6.1 Code to generate course art

```
# install.packages('readr')
# install.packages('tidyverse')
# install.packages("devtools")
# devtools::install_github("BlakeRMills/MetBrewer")

library(readr)
library(MetBrewer)
library(tidyverse)

course_code <- "STA303"

my_colours <- c(met.brewer("Cross", n = 8), met.brewer("Cross", n = 9))

set.seed(parse_number(course_code))

ngroup=17
names=paste("G_",seq(1,ngroup),sep="")
DAT=data.frame()

for(i in seq(1:30)){
  data=data.frame( matrix(0, ngroup , 3))
  data[,1]=i
  data[,2]=sample(names, nrow(data))
  data[,3]=prop.table(sample( c(rep(0,100),c(1:ngroup)) ,nrow(data)))
  DAT=rbind(DAT,data)
}
colnames(DAT)=c("Year","Group","Value")
DAT=DAT[order( DAT$Year, DAT$Group) , ]

ggplot(DAT, aes(x=Year, y=rev(Value), fill=Group )) +
  geom_area(alpha=1  )+
  theme_bw() +
  scale_fill_manual(values = my_colours)+
  theme(
    text = element_blank(),
    line = element_blank(),
    title = element_blank(),
    legend.position="none",
    panel.border = element_blank(),
    panel.background = element_blank(),
    plot.margin = margin(0, -2.7, 0, -2.7, "cm"))

ggsave(paste0(course_code, "-base.png"), width = 24, height = 2)
```

6.2 M1 supporting information on matrices (not assessed)

6.2.1 Background

6.2.1.1 Some true things about matrices

- The **rank** of a matrix is the number of linearly independent columns your matrix has.
- If the number of columns = the rank of the matrix all the columns are **linearly independent**. If the umber of columns is > the rank of the matrix, all the columns are *not* linearly independent.
- You can only **invert** a square matrix if all its columns are linearly independent. (Determinant non-zero).

Why do we care? In linear regression, to estimate β, our vector of coefficients, we calculate $(X^TX)^{-1}X^TY$. The elements of β can't be estimated if X^TX (a square matrix) isn't invertible.

Clarification to what I said in the tests activity: We usually perform regression with an intercept because we don't want to assume out line passes through the origin, **0**. So, if there is an intercept, (column of 1s in the model matrix) we must convert every categorical variable with k levels into $k-1$ dummy variables to have the intercept and still satisfies linear independence. IF we ditch the intercept, we can have k dummies, but this is only usually useful in the specific case of ANOVA.

6.2.2 Example

```
library(tidyverse)
library(palmerpenguins)
```

```
# function to replace NAs with 0 and text with 1
dummify <- function(x){
  if_else(is.na(x), 0, 1)
}

# create a smaller toy version of the penguin dataset (just for diplay purposes)
set.seed(24601)
pengwings <- penguins %>%
  group_by(species) %>%
  sample_n(4) %>%
  select(body_mass_g, species)

pengwings
```

```
## # A tibble: 12 x 2
## # Groups:   species [3]
##    body_mass_g species
##          <int> <fct>
##  1        3900 Adelie
##  2        3700 Adelie
##  3        3450 Adelie
##  4        3175 Adelie
##  5        4500 Chinstrap
##  6        3850 Chinstrap
##  7        3650 Chinstrap
##  8        4550 Chinstrap
##  9        5150 Gentoo
## 10        5000 Gentoo
## 11        5700 Gentoo
## 12        4600 Gentoo
```

```
# creating a version of the data where there is a dummy column for each level of species instead of one sp
wider <- pengwings %>%
  pivot_wider(id_cols = everything(), names_from = species, values_from = species, names_prefix = "species
  mutate_at(vars(starts_with("species.")), dummify)

wider
```

```
## # A tibble: 12 x 4
##    body_mass_g species.Adelie species.Chinstrap species.Gentoo
##          <int>          <dbl>             <dbl>          <dbl>
##  1        3900              1                 0              0
##  2        3700              1                 0              0
##  3        3450              1                 0              0
##  4        3175              1                 0              0
##  5        4500              0                 1              0
##  6        3850              0                 1              0
##  7        3650              0                 1              0
##  8        4550              0                 1              0
##  9        5150              0                 0              1
## 10        5000              0                 0              1
## 11        5700              0                 0              1
## 12        4600              0                 0              1
```

6.2.2.1 Classic regression, k-1 dummies (intercept)

```
mod1 <- lm(body_mass_g ~ species, data = pengwings)
summary(mod1)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species, data = pengwings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -512.50 -310.94  -34.37  348.44  587.50
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3556.3      206.8  17.199 3.42e-08 ***
## speciesChinstrap    581.2      292.4   1.988  0.07809 .
## speciesGentoo      1556.2      292.4   5.322  0.00048 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 413.6 on 9 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:   0.71
## F-statistic: 14.46 on 2 and 9 DF,  p-value: 0.001545
```

```
head(model.matrix(mod1))
```

```
##   (Intercept) speciesChinstrap speciesGentoo
## 1           1                0             0
## 2           1                0             0
## 3           1                0             0
## 4           1                0             0
## 5           1                1             0
## 6           1                1             0
```

Does the rank of the model matrix equal the number of columns?

```
qr(model.matrix(mod1))$rank == ncol(model.matrix(mod1))
```

```
## [1] TRUE
```

Okay, linearly independent, we're good to go!

6.2.2.2 Classic regression but trying to force k dummies (with an intercept)

```
mod2 <- lm(body_mass_g ~ species.Adelie + species.Gentoo + species.Chinstrap, data=wider)
summary(mod2)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species.Adelie + species.Gentoo +
##     species.Chinstrap, data = wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -512.50 -310.94  -34.38  348.44  587.50
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4137.5      206.8  20.010 9.04e-09 ***
## species.Adelie      -581.2      292.4  -1.988  0.07809 .
## species.Gentoo       975.0      292.4   3.334  0.00874 **
## species.Chinstrap       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.6 on 9 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:   0.71
## F-statistic: 14.46 on 2 and 9 DF,  p-value: 0.001545
```

```
model.matrix(mod2)
```

```
##    (Intercept) species.Adelie species.Gentoo species.Chinstrap
## 1            1              1              0                 0
## 2            1              1              0                 0
## 3            1              1              0                 0
## 4            1              1              0                 0
## 5            1              0              0                 1
## 6            1              0              0                 1
## 7            1              0              0                 1
## 8            1              0              0                 1
## 9            1              0              1                 0
## 10           1              0              1                 0
## 11           1              0              1                 0
## 12           1              0              1                 0
## attr(,"assign")
## [1] 0 1 2 3
```

```
qr(model.matrix(mod2))$rank
```

```
## [1] 3
```

Does the rank of the model matrix equal the number of columns?

```
qr(model.matrix(mod2))$rank == ncol(model.matrix(mod2))
```

```
## [1] FALSE
```

Not linearly independent. Why?

Well, this intercept column is a linear combination of the three species columns!

```
m <- model.matrix(mod2)

m[,1] == m[,2] + m[,3] + m[,4]
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

6.2.3 Regression NOT classic, actually ANOVA! (no intercept)

```
mod3 <- lm(body_mass_g ~ 0 + species.Adelie + species.Gentoo + species.Chinstrap, data=wider)
summary(mod3)
```

```
##
## Call:
## lm(formula = body_mass_g ~ 0 + species.Adelie + species.Gentoo +
##     species.Chinstrap, data = wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -512.50 -310.94  -34.38  348.44  587.50
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## species.Adelie      3556.2      206.8   17.20 3.42e-08 ***
## species.Gentoo      5112.5      206.8   24.73 1.39e-09 ***
## species.Chinstrap   4137.5      206.8   20.01 9.04e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.6 on 9 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9909
## F-statistic: 435.8 on 3 and 9 DF,  p-value: 4.659e-10
```

```
model.matrix(mod3)
```

```
##    species.Adelie species.Gentoo species.Chinstrap
## 1               1              0                 0
## 2               1              0                 0
## 3               1              0                 0
## 4               1              0                 0
## 5               0              0                 1
## 6               0              0                 1
## 7               0              0                 1
## 8               0              0                 1
## 9               0              1                 0
## 10              0              1                 0
## 11              0              1                 0
## 12              0              1                 0
## attr(,"assign")
## [1] 1 2 3
```

```
qr(model.matrix(mod3))$rank
```

```
## [1] 3
```

Challenge question for +100 stats respect points: How do you interpret these coefficients?

Does the rank of the model matrix equal the number of columns?

```
qr(model.matrix(mod3))$rank == ncol(model.matrix(mod3))
```

```
## [1] TRUE
```

Great, back to being linearly independent.

6.2.4 Further reading (if you want it)

As I said at the beginning, I'm not planning to assess you on any of this. If you're interested in knowing more, or just think matrix algebra is delicious, I think this is a delightfully approachable walk through. https://online.stat.psu.edu/stat462/node/132/

7

Module 1

7.1 Instructor information

Prof. Liza Bolton

Course email: sta303@utoronto.ca

Help hours: During the second half of Wednesday classes: 11:10—12:00 p.m. ET and 4:10—5:00 p.m. ET

Please do not email my individual email with STA303 questions or use Quercus mail. It makes it harder to keep track of and address your questions. Sending me messages on three different platforms will NOT speed up my response. I will not respond messages that don't follow the course communication policy.

You can see the latest version of my email autoresponder here.

7.2 Upward management tips

'Upward management' is basically managing your manager. [1] If you make their life easier and help them be more effective, this should also make *your* life easier and is a good investment in your own career and skills building.

While our course isn't a business, some of the basic parts of this concept apply well to your time at univeristy AND can set you up for success in graduate studies and your future career.

But why should you put effort into upward managing me? Well, while I will always seek to treat all students fairly and to listen to your feedback, YOU can make this easier or harder for me, and thus make this course better or worse for yourselves.

For example, if I have to use all my time and energy following up on unclear emails for more information and dealing with students who haven't followed instructions etc. etc….I won't have that energy to put into writing TeamUp! activities to give you extra practice and opportunities to earn bonus points.

7.2.1 Communicate using the tools your manager prefers.

- **In business**, this means knowing who likes face-to-face vs email, or whether your manager would rather receive an instant message than an email for a quick question.
- **In STA303**, this means:
  - Using Piazza for all course admin and content questions.
  - Using the appropriate forms for accommodations and regrade requests.

[1] I used to haaaate this concept because I learned it while doing a consulting internship at a large international professional services firm and I really struggled with my manager.

  - Emailing sta303@utoronto.ca for private issues not otherwise covered by the other tools, e.g. emailing me your Accommodation Serivces letter, requesting an extension to an assessment that conflicts with essential travel.
  - Asking questions in office hours.

7.2.2 Write good emails (when emails are appropriate)

- **In business**, this might mean:
  - Choosing who should be the main recipients vs CCed/BCCed.
  - Ensuring your contact details are clear in your signature.
  - Make sure the subject line is informative and short
  - Making the text of the email as clear and concise as possible.
  - Use proper grammar and punctuation.
  - Don't use emoji in formal emails. If in doubt, leave 'em out.
- **In STA303**, this looks like:
  - Everything in the right-hand column, plus the following.
  - Starting an email with "Hi Prof. Bolton," or "Hi Liza,".[2]
  - Sign off the email with your **preferred name** (i.e. what should I call you when I reply) and if your have different official name, include that and your UTORid below your name.
  - Subject line including [Prof. Bolton] or [TA name] if your email is for a specific person.

7.2.3 Understand your manager's goals.

- **In business**, this might look like understanding their KPIs and how you can help make sure these are met.
- **In STA303**, most of *my* goals are *for you*, like that you learn useful statistical skills, improve your writing skills etc. I also personally want to improve as an instructor and have fun talking about somethign I love.
  - You can help me with these goals by working on this course every week, trying your best, asking for help early and often, engaging with feedback gathering mechanisms and providing constructive feedback if something is not working.

7.2.4 Demonstrate self-management and resilience while also asking for help and flagging problems early.

- **In business**, this might look like:
  - Being proactive about addressing possible problems before they occur. Managers like a 'no surprises' policy.
  - Preparing a list of questions to cover in a meeting or to compile in an organised fashion into one email (instead of ten).
  - Searching for answers yourself before asking your manager and improving your strategies for finding available information.
- **In STA303**, this looks like:
- Putting in a little effort into find answers before posting on Piazza/asking in class. (*Have you searched Piazza? have you re-read/watched the assigned materials? have your checked the syllabus and recent announcements?*)
- Come to office hours often, lists of questions very welcome!
- Getting in touch (or asking your registrar to) **early** if you might hate to miss a lot of class. It is usually easier to find a solution if I know things ahead of time, or as soon as possible after.

[2] "Dear Madam" and "To my esteemed professor" make me uncomfortable.

7.3 Recap of linear models

7.4 Why model?

- The goal of a model is to provide a (relatively) simple summary of a dataset.
- We can describe data AND make predictions.

7.5 Linear models

In a linear model,

$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i$

The response is predicted by a linear function of explanatory (or predictor) variables plus an error term.

a.k.a.

DATA = MODEL + ERROR

7.6 Linear regression assumptions

L: your model is **L**inear.

I: Errors are **I**ndependent (usually satisfied if observations are independent).

N: Errors are **N**ormally distributed with expected value zero, $E[\epsilon_i] = 0$

E: **E**qual/constant variance (homoscedasticity), var$[\epsilon_i] = \sigma^2$.

We can express “I N E” above as assuming the errors are i.i.d Normal with mean of zero and variance σ^2,

$$\epsilon_i \sim N(0, \sigma^2)$$

7.7 What makes it a *linear* model?

A model is **linear** if it is *linear in the parameters.* That is, all the β enter the model in a linear way. It is totally fine if the predictor varaibles enter the model in a non-linear way.

7.7.1 Linear

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$
- $y_i = \beta_0 + \gamma_1 \delta_1 x_{1i} + \beta_2 \exp(x_{2i}) + \epsilon_i$

7.7.2 NOT linear

- $y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \epsilon_i$
- $y_i = \beta_0 exp(\beta_1 x_{1i}) + \epsilon_i$

Internally screaming “DON’T LET THE BETAS TOUCH” often helps me remember what is not linear. Additionally, don’t let them do anything *weird*, like get exponentiated.

7.8 Optional refresher reading

7.8.1 Brief discussion of assumptions with examples

Section 1.3 of *Broaden your Statistical Horizons* on the assumptions of OLS
https://bookdown.org/roback/bookdown-bysh/ch-MLRreview.html#ordinary-least-squares-ols-assumptions
[freely accessible]

7.8.2 Fitting linear models in R

Section 1.6 of *Broaden your Statistical Horizons* on Mulitple linear regression (bootstrapping not assessed)
https://bookdown.org/roback/bookdown-bysh/ch-MLRreview.html#multreg [freely accessible]

7.8.3 Delicious mathematics

Chapter 1 of *Wood, S. N. (2017). Generalized additive models : An introduction with r, second edition*
http://go.utlib.ca/cat/13435628 [you will need to log in to the U of T library for access]

7.9 Common statistical tests as linear regression

7.9.1 Introduction

Technical note: I have had to remove code answer checking, but if you get stuck the final 'Hint' is the solution. This is ungraded, so don't worry, just slightly inconvenient.

You have probably encountered several statistical tests in your studies so far.

7.9.1.1 Parametric

E.g. one-sample t-tests, paired t-tests, two-sample t-tests, one-way ANOVA, two-way ANOVA
Parametric tests make assumptions about the distribution of the population from which our sample data have been drawn.

7.9.1.2 Non-parametric

E.g. Wilcoxon signed rank, Mann Whitney-U, Kruskal-Wallace
Non-parametric tests do not assume that our outcome is Normally distributed. They are sometimes called 'distribution-free', but note that this is because they have fewer assumptions than parametric tests, not because they have no assumptions at all.

7.9.1.3 Aside: But why are there two types of tests?

Parametric tests are more **powerful**, i.e., they have a better chance of detecting an effect if there is one there to find. So why would you ever use a less powerful test? Well, with great power comes ~~great responsibility~~ more assumptions that must be valid to proceed.

There is a GIF in the web version. Not required content. https://tenor.com/view/spider-man-uncle-ben-with-great-power-comes-great-responsibility-its-true-just-saying-gif-24193883

Non-parametric tests are a great choice when your outcome is an ordinal variable, is ranks, or there are problematic outliers.

For the purposes of this lesson, we're going to focus more on parametric tests, but also take a look at the corresponding non-parametric tests with the slight white lie that they are just ranked versions of their parametric companions. This approach is pretty good as long as you have a reasonable sample size.

Imagine this:
You're on a ship trying to spot land. ***Parametric*** *tests are the crew member with the best eyesight, but they can be fussy and the conditions have to be right for them to work in or they will breakdown.*

Non-parametric *tests are the crew member with not quite as good eyesight, but they're more laid back about the conditions you make them work in.*

In the following sections we'll explore several of these tests.

7.9.2 One-sample t-test

I am assuming you've seen this in a 200-level statistics course or equivalent. Brief recap below.

7.9.2.1 Use case

You want to know if it is believable that the population mean is a certain value (our 'hypothesized value' below).

7.9.2.2 Assumptions

1. The data are continuous.
2. The data are normally distributed.
3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample

(Do these sound familiar from linear regression?)

7.9.2.3 Hypotheses

$$H_0 : \mu = \text{hypothesized val}$$
$$H_1 : \mu \neq \text{hypothesized val}$$

What are we doing? Finding the strength of evidence against the claim that the population mean is some hypothesized value.

The test statistic, t, is calculated as follows:

$$t = \frac{\bar{x} - \text{hypothesized val}}{s/\sqrt{n}}$$

We then compare this t value to the t-distribution with degrees of freedom df = n - 1 and find the area under the curve that represents the probability of values likes ours or more extreme.

7.9.2.4 Example

Suppose existing research suggests that the average weight of penguins is 4000 grams. You want to see if this makes sense for your new penguins data.

$$H_0 : \mu = 4000$$
$$H_1 : \mu \neq 4000$$

The `penguins` dataset is already loaded, you you don't have to run any libraries. Use the `t.test()` function run a one-sample t-test.

```
##
##  One Sample t-test
##
## data:  penguins$body_mass_g
## t = 4.6525, df = 341, p-value = 4.7e-06
## alternative hypothesis: true mean is not equal to 4000
## 95 percent confidence interval:
##  4116.458 4287.050
## sample estimates:
## mean of x
##  4201.754
```

7.9.2.5 Now as a linear model

First, consider the following, what would a linear regression with no predictor variables and just an intercept tell you?

Create a linear regression model called `mod1`(replace the blank below) that is an 'intercept only model' with `body_mass_g` as the response.

```
##
## Call:
## lm(formula = body_mass_g ~ 1, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1501.8  -651.8  -151.8   548.2  2098.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4201.75      43.36   96.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 802 on 341 degrees of freedom
##   (2 observations deleted due to missingness)
```

It turns out the estimate from this linear regression is the same as the sample mean.

```
mean(penguins$body_mass_g, na.rm = TRUE) #na.rm = TRUE removes missing values
```

```
## [1] 4201.754
```

Now, recall that with the t-test, we calculate our test statistic by subtracting the hypothesized value from the mean. Let's run the linear model again, but on the left-hand side of the formula, subtract the hypothesized value.

```
##
## Call:
## lm(formula = body_mass_g - 4000 ~ 1, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1501.8  -651.8  -151.8   548.2  2098.2
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   201.75      43.36   4.652  4.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 802 on 341 degrees of freedom
##   (2 observations deleted due to missingness)
```

Compare the results of this `summary(mod2)` and your earlier t-test. You should see that the t value, degrees of freedom and p-value are the same for both analyses.

Thus, our one sample t-test hypotheses,

$$H_0 : \mu = \text{hypothesized val}$$

$$H_1 : \mu \neq \text{hypothesized val}$$

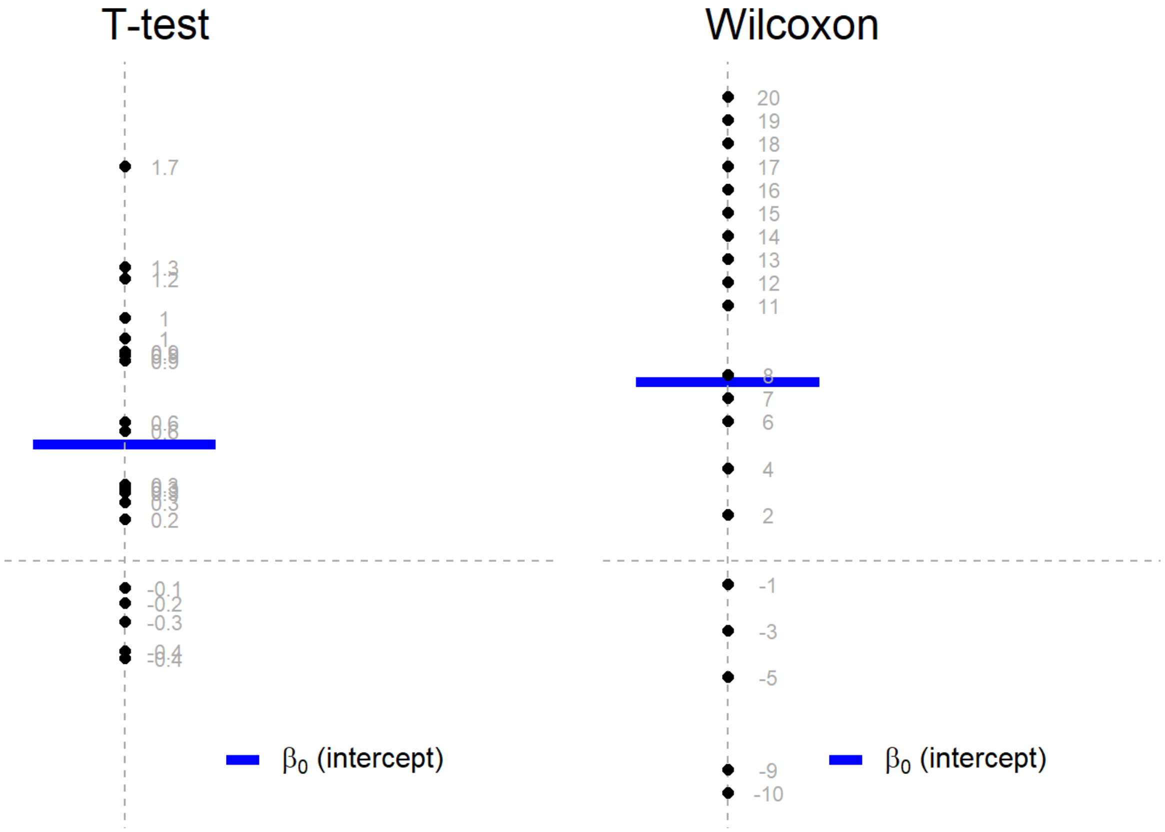
are equivalent to our linear regression hypotheses about the intercept,

$$H_0 : \beta_0 = \text{hypothesized val}$$

$$H_1 : \beta_0 \neq \text{hypothesized val.}$$

7.9.2.6 Wilcoxon signed-rank test

While the linear regression approach to the one-sample t-test is exact, we can also approximate the Wilcoxon rank-sign test with linear regression. See below.



Note: The above is just example from some toy data, but aims to illustrate how a t-test is treating the data and how the Wilcoxon test is treating the data.

```r
# Function to get signed rank of each observation
signed_rank = function(x) sign(x) * rank(abs(x))

# The wilcoxon test function
wilcox.test(penguins$body_mass_g, mu = 4000)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  penguins$body_mass_g
## V = 34723, p-value = 0.0004829
## alternative hypothesis: true location is not equal to 4000
```

```r
# Equivalent linear model
mod3 <- lm(signed_rank(penguins$body_mass_g-4000) ~ 1)
summary(mod3)
```

```
##
## Call:
## lm(formula = signed_rank(penguins$body_mass_g - 4000) ~ 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -334.63 -173.13  -22.13  187.87  305.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.63      10.53   3.479 0.000569 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 194.7 on 341 degrees of freedom
##   (2 observations deleted due to missingness)
```

[Optional] Check out the theory behind the rank transformation in section 3.0.2
https://lindeloev.github.io/tests-as-linear/#3_pearson_and_spearman_correlation

7.9.2.7 Paired sample t-test and Wilcoxon matched pair

A paired t-test is equivalent to a one sample t-test if you just consider $x_{\text{diff } i} = x_{1i} - x_{2i}$, i.e., $x_{\text{diff } i}$ is the difference of the paired values for each observation, and proceed with $x_{\text{diff } i}$ as you would in the one sample case. Likewise for the Wilcoxon

The R code (not evaluated here) would be as follows:

```r
# Built-in Wilcoxon matched pairs
wilcox.test(x1, x2, paired = TRUE)

# Equivalent linear model:
summary(lm(signed_rank(x1 - x2) ~ 1))
```

7.9.3 Dummy variables

Let's take a quick detour before we explore the next tests. We'll need to understand the concept of dummy variables and contrasts first.

7.9.3.1 The matrices we use for linear regression

Recall that we can express our linear regression in matrix form:

$$\mathbf{y} = X\beta + \varepsilon$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

We often talk about **X** as the **model matrix** (or design or regressor matrix) and it will be the focus of this section.

7.9.3.2 Getting our model matrix in R

Let's start by fitting a model with `body_mass_g` as the response and `flipper_length_mm` and `species` as the predictor variables.

(Note: Users of statistics use a lot of different words to refer to the same thing. Can you think of other terms people might use instead of *response* and *predictor*?)

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + species, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -927.70 -254.82  -23.92  241.16 1191.68
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4031.477    584.151  -6.901 2.55e-11 ***
## flipper_length_mm    40.705      3.071  13.255  < 2e-16 ***
## speciesChinstrap   -206.510     57.731  -3.577 0.000398 ***
## speciesGentoo       266.810     95.264   2.801 0.005392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 375.5 on 338 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7826, Adjusted R-squared:  0.7807
## F-statistic: 405.7 on 3 and 338 DF,  p-value: < 2.2e-16
```

Now we can use the `model.matrix()` function to extract the model matrix for `mod4`. I've applied `head()` to stop the entire thing being printed.

```
##    (Intercept) flipper_length_mm speciesChinstrap speciesGentoo
## 1            1               181                0             0
## 2            1               186                0             0
## 3            1               195                0             0
## 5            1               193                0             0
## 6            1               190                0             0
## 7            1               181                0             0
## 8            1               195                0             0
## 9            1               193                0             0
## 10           1               190                0             0
## 11           1               186                0             0
## 12           1               180                0             0
## 13           1               182                0             0
## 14           1               191                0             0
## 15           1               198                0             0
## 16           1               185                0             0
## 17           1               195                0             0
## 18           1               197                0             0
## 19           1               184                0             0
## 20           1               194                0             0
## 21           1               174                0             0
```

You'll notice that even though we only had an intercept and two variables, we have four columns in our model matrix. You should also notice that R has given the columns helpful names, and that we have a column for the Chinstrap species and the Gentoo species, but not the Adelie species.

Further, recall that when we are working with a categorical variables we call the different values the the variables can take **"levels"**. I may also refer to these as factor variables, and talk about the "levels of the factor".

What R is doing is dropping the first level (alphabetically) of the categorical variable and then creating **dummy variables** for each of the other levels.

The dropped level becomes our **reference level** and this should be familiar from interpreting summary output in previous courses where you have conducted multiple linear regressions with categorical variables.

A dummy variable is also called an indicator variable, and it *indicates* whether or not the given observation takes that level or not. I.e., if the 40th penguin in this dataset had a 1 in the speciesGentoo column, then I know it is a Gentoo penguin, and that it won't have a 1 in the speciesChinstrap column because each penguin can only have one species.

More generally, the sum across the row of the dummy variables for one categorical variable will either be 0 (if that observation has the reference level) or 1 (not the reference level) but you will never have more than one 'one' amongst the dummies for a given categorical variable.

7.9.3.2.0.1 [Unassessed aside] Why do we have to drop one of the levels? You may recall that for the matrix calculations required to get our vector of βs, we need to be able to invert X our matrix. We can only invert matrices for which all the columns are linearly independent and if we have the intercept AND dummies for all the levels of the categorical variable, our matrix will be linearly dependent.

Additional optional discussion here.

7.9.4 Two means

Back to tests!

Independent t-tests let you compare two means. I am assuming you've seen this in a 200-level statistics course or equivalent. Brief recap below.

7.9.4.1 Use case

You want to know if it is believable that two independent groups have the same population mean.

7.9.4.2 Assumptions

1. The data are continuous.
2. The data are normally distributed (in each group).
3. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample
4. The variances for the groups are equal.

Notice that these are the same assumptions as the one-sample t-test, but with the equality of variances assumption added.

7.9.4.3 Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

What are we doing? Finding the strength of evidence against the claim that the population means for both groups are the same. This differs from the one sample test because we have uncertainty about BOTH values here. Both are population parameters that we don't know.

The test statistic, t, is calculated as follows:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

We then compare this t value to the t-distribution with degrees of freedom $df = n_1 + n_2 - 2$ and find the area under the curve that represents the probability of values likes ours or more extreme.

7.9.4.4 Example

Conduct an independent t-test to test if the mean of `body_mass_g` is the same for male and female penguins (`sex`). Add your code below. Note: you must set `, var.equal = TRUE` as one of the arguments for it two be the independent t-test. If you don't set this we are conducting a *Welch's t-test*. I won't be covering this, but it is covered in the source credited at the end of this activity.

```
##
##  Two Sample t-test
##
## data:  body_mass_g by sex
## t = -8.5417, df = 331, p-value = 4.897e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -840.8014 -526.0222
## sample estimates:
## mean in group female   mean in group male
##             3862.273             4545.685
```

Now, based on what we've learned, write a linear model using the `lm()` function to do the same this as our independent t-test. Save the model as `mod5`.

```
##
## Call:
## lm(formula = body_mass_g ~ sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1295.7  -595.7  -237.3   737.7  1754.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3862.27      56.83  67.963  < 2e-16 ***
## sexmale       683.41      80.01   8.542  4.9e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 730 on 331 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.1806, Adjusted R-squared:  0.1781
## F-statistic: 72.96 on 1 and 331 DF,  p-value: 4.897e-16
```

Take a moment to match up parts of the outputs that are the same. There is a difference here in that the sign of the test statistics differs. That does not matter as out t-distribution is symmetrical and we're doing a two-tailed test.

7.9.4.5 Mann-Whitney U

Similar idea to before, except for this test it is just rank not signed rank.

```r
# Wilcoxon / Mann-Whitney U (multiple names)
wilcox.test(body_mass_g ~ sex, data = penguins)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  body_mass_g by sex
## W = 6874.5, p-value = 1.813e-15
## alternative hypothesis: true location shift is not equal to 0
```

```
# As linear model with our dummy-coded group_y2:
summary(lm(rank(body_mass_g) ~ sex, data = penguins))
```

```
##
## Call:
## lm(formula = rank(body_mass_g) ~ sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -183.68  -74.13  -16.63   91.32  162.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  128.630      6.948  18.512   <2e-16 ***
## sexmale       86.051      9.783   8.796   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.25 on 331 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.1895, Adjusted R-squared:  0.187
## F-statistic: 77.37 on 1 and 331 DF,  p-value: < 2.2e-16
```

7.9.5 ANOVA

7.9.5.1 Use case

You've probably seen 'ANOVA' in the context of model comparison, but it is also a popular test in psychology and other disciplines.

Let's look specifically at one-way ANOVA (or the F-test). It tests if all the means for several groups (more than 2) are the same or if at least one is different.

I hope this sounds a bit like the next evolution from the independent t-test…

(+ 1000 stats respect points to anyone who draws Pokemon-esque evolutions of these three tests…with regression as the mega-evolution…)

7.9.5.2 Assumptions

And it just so happens that the assumptions for the one-way ANOVA (also called the F-test) are EXACTLY the same as for the independent t-test

There is a GIF in the web version. Not required content.
https://gifst.blogspot.com/2019/03/snl-hi-saturday-night-live-hey-kate.html

1. The data are continuous.
2. The data are normally distributed (in each group).
3. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample
4. The variances for the groups are equal.

7.9.5.3 Hypotheses

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_1 : \text{at least one } \mu \text{ differs from the others}$$

7.9.5.4 Example

Let's now look at body mass across species. Suppose we wanted to know if was believable that the the means body mass in grams was the same across all three species. This is when we could fit a quick ANOVA to test this. The `aov()` allows us to do this.

```
summary(aov(body_mass_g ~ species, data = penguins))
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)    
## species       2 146864214 73432107   343.6 <2e-16 ***
## Residuals   339  72443483   213698                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

That looks a lot like the output from calling summary on `lm()`…in fact, aov is just a wrapper for lm! Which means it has been linear regression the whole time.

7.9.5.5 We'll stop here and talk further about this particular topic in class this week.

Your next topic is reproducible examples (reprexes).

7.9.6 Credits

Credit to **Jonas Kristoffer Lindeløv** for the excellent resource this resource is based on. There are more examples there than we will cover in this course.

8

Module 2

9

Module 3

10

Module 4

11

Module 5

12

Resources

12.1 Course tools overview

While we've tried to keep things as streamlined as possible, there are still several different tools we'll be using this semester. Your U of T login should work with all of them. The below PDF file provides an overview of how you'll be interacting with each one.

- At the bottom of the page is an embedded slideshow introducing you to the JupyterHub.
- You can always access Piazza from the Navigation Menu on the left.
- Instructions for setting up your U of T Zoom are on the Zoom page and links are in the Navigation menu and on the home page.

12.1.1 Admin

Logo	Description
QUERCUS UNIVERSITY OF TORONTO	Quercus will be used for timed assessments, some submissions and announcements.
zoom	Synchronous classes and office hours will be hosted via Zoom. You MUST join using your U of T Zoom account to be admitted. Get your account: utoronto.zoom.us
F	Microsoft Forms will be used for several important administrative forms. You will need to be signed in to your U of T account in the same browser to access these.

12.2 Using RStudio with the JupyterHub

We will be using R through RStudio to conduct analyses in this course. If you have a local installation of R you are welcome to continue using that, but, for this course, you do not need to have R and RStudio installed.

Instead, assessments and activities will be shared through the U of T JupyterHub. This gives you access to RStudio in your browser through your U of T login on any internet-connected device. It means you don't have to fight package installations and we can instead focus on the good stuff.

Please read through the following slides, experiment with the example sharing link, make sure you know how to knit an Rmd to pdf + export the pdf, and practice navigating and moving files.

Link: https://rstudio-with-jupyerhub-uoft.netlify.app.

12.3 Zoom, Zoom, Zoom, Zoom...

Access to STA303 synchronous meetings and office hours is restricted to our students.

Set up your U of T Zoom account

12.3.1 Make sure your Zoom is up to date

To participate fully, you will need Desktop client or mobile app: version 5.3.0 or higher. You can check your desktop client or mobile app version by following these instructions.

12.3.2 Customize!

Once you have logged in, please customize your profile:

- Update your name to your **preferred name** (what you would like us to call you in class) Note: this may not be allowed with your U of T settings, so don't worry if this doesn't work.
- Add a **profile picture** (please make it a photo of YOU or an avatar that looks like you...we don't want Snoopy or Joe Biden[1] in class)

12.3.3 VPN

There is a University of Toronto VPN (UTORvpn) that you have access to as a student. It may help with video quality and access to U of T resources.
If you are based in mainland China, the Alibaba Cloud Enterprise Network (CEN)Links to an external site. service should help with your Quercus access.

12.3.4 Notes:

1. Please always use your real name and face for this course, and be cautious about changing them and your virtual background for other meetings. A joke background for a call with family or friends may not be appropriate for class.
2. For class meetings, the settings will always be that your camera and microphone are off to begin with so you have the control to check these things first.
3. We do ask that, when possible, you use your microphone in office hours, breakout groups and any other small group meetings and strongly prefer that you use your camera AND microphone. We trust you to make the best choice for your environment, comfort and learning.
4. You may get a **"This meeting is for authorized participants only message".** Choose the "Sign in with SSO" option to sign in.

[1] Yes, these are real images students have used.

12.3.5 Changing your profile picture on Zoom and Quercus

Follow these instructions to add a profile picture (or bitmoji style avatar if you'd prefer) to Quercus and Zoom. I want this experience to be more social and less faceless. Please don't use photos of cartoon characters, etc. A good photo will be a close-up of your face so we can see who you are even when the photo is small.

12.3.6 What to do if you experience technical difficulties during class?

First, (if possible) send me a chat note that you're having technical difficulties and are working to resolve them.

Second, leave the meeting and re-enter. This often resets things and resolves the problem. Before entering the meeting, make sure all of your devices are properly plugged in and Bluetooth devices are connected.

If that doesn't fix things, exit the meeting again and update your Zoom Client. This is the Zoom software that should be on your computer. Here's a short video tutorial explaining how to update the software:
https://www.youtube.com/watch?v=E7zERcVLUBM.

After updating, enter the meeting again to see if this resolved your problems.

Our synchronous classes are recorded, so if your technology is just going catastrophically wrong, go get a cup of tea/coffee/water and relax, you can catch up with the recording when it is posted on Quercus.

12.3.7 What to do if your instructor or TA is experiencing technical difficulties on Zoom

First, check the **chat** to see if the instructor or the project mentor have said what is going on and what they are doing to fix things and follow any instructions they give.

Second, if they have disappeared completely, wait 10 minutes (or until the end of the meeting time, whichever comes first) before closing the call. (You can do other things in the meantime, but be ready to jump back in).

Third, expect to see an announcement on Quercus afterwards telling you what to do (e.g. it might be to watch a video I'll record later, to review some slides or perhaps there is nothing to do and i'll see you next time).

12.4 Student support services and resources

12.4.1 Mental health support

You may find yourself feeling overwhelmed, depressed, or anxious. Lots of people feel the same way. There is help available from mental health professionals 24 hours a day via online and phone-based services. Here are some that are available to U of T students:

- MySSP - My Student Support Program 1-844-451-9700, or outside of Canada call 001-416-380-6578
- Good2Talk Student Helpline 1-866-925-5454, or text GOOD2TALK to 686868
- Distress Centres of Greater Toronto 416-408-4357, or text 45645

There is also the new Navi tool for U of T students, it is a chatbot and your questions are totally anonymous.
http://uoft.me/navi

The student union are also curating a list here: https://www.utsu.ca/mental-health/

12.4.2 General University resources

The following are some important links to help you with academic and/or technical service and support:

- **Health & Wellness** can help with appointments with a range of clinicians, nutrition, immunizations, sexual and reproductive health and much more. Many of their services continue to be available online.
- **Arts & Sciences** student resources through Sidney Smith Commons Online
- **General** student services and resources at Student Life
  - Tips for dealing with multi-choice questions (MCQs)
  - Book an appointment with a learning strategist (they can help you with strategies for MCQs also)
- Full **library** service through the University of Toronto Libraries
- Resources on **academic support** from the Academic Success Centre
- Learner support at the **Writing** Centre
- Information about **Accessibility** Services
- Quercus Information in the Canvas Student Guide
- Logistical and social support for **international students** at the Center for International Experience

Visit the A&S online resources for students page for resources available to support you through your online studies. If you have further questions, please email ask.artsci@utoronto.ca.

12.4.3 Financial support

A list of University financial supports, work-study opportunities, as well as provincial and federal government programs is available on the University's Financial Support & Funding Opportunity directory.

12.4.4 Arts & Science COVID19 FAQ

The **Arts & Science Undergraduate FAQ page** addresses frequently asked questions that are specific to undergraduate students taking courses with the Faculty of Arts & Science. On this page you will find information for:

Messages from Dean Woodin can be found on the A&S latest updates page.