Data Science

Midterm Review

1. Our mod of the day.

- 1. Our mod of the day.
- 2. Project 2

- 1. Our mod of the day.
- 2. Project 2
- 3. Equation(s) you were promised

- 1. Our mod of the day.
- 2. Project 2
- 3. Equation(s) you were promised
- 4. How this lecture is going to work.

Our moderator

Our moderator

1. Laura!

Project 2

Project 2

1. Well done, submission rate was much higher than Project 1.

This was written wrong:

$$(1+\frac{\lambda}{N})^{-2}$$

$$\sqrt{1+\frac{\lambda}{N}}$$

Let's say you had a variable where half the data was missing $(\lambda = 0.5)$ and you used N = 5 for the number of generated data sets:

$$\sqrt{1 + \frac{0.5}{5}} = 1.049$$

How much better would it be if you used an 'infinite' number of generated data sets?

$$\sqrt{1 + \frac{0.5}{5}} = 1.0$$

To a notebook!

Pooled Slope Estimate: One way to 'average' a pooled analysis (multiple imputation)

1. Calculate the linear regression for each imputed data-set

- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}

- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}
- 3. Just average them!

- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}
- 3. Just average them!

$$\beta_{1p} = \frac{\beta_{11} + \dots + \beta_{1n}}{n}$$

That was pretty naïve, let's try something more sophisticated:

1. Calculate the linear regression for each imputed data-set

- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}

- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}
- 3. Each linear regression has its own standard error: Z_i

- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}
- 3. Each linear regression has its own standard error: Z_i
- 4. Do the following fancy weighted average:

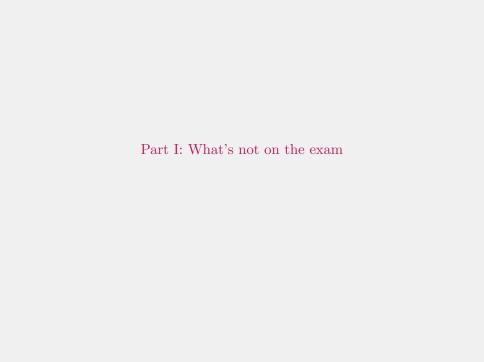
- 1. Calculate the linear regression for each imputed data-set
- 2. Each linear regression has its own slope: β_{1i}
- 3. Each linear regression has its own standard error: Z_i
- 4. Do the following fancy weighted average:

$$V_s = \frac{\sum Z_i}{n} + (1 + n^{-1}) * \frac{1}{n-1} * \sum (\beta_{1i} - \beta_{1p})^2$$

1. I will be clear about what will not be on the exam

- 1. I will be clear about what will not be on the exam
- 2. I will explain how the midterm will be administered.

- 1. I will be clear about what will not be on the exam
- 2. I will explain how the midterm will be administered.
- 3. We will discuss expectations



1. We didn't cover NLTK in the lectures, so don't worry about it.

- 1. We didn't cover NLTK in the lectures, so don't worry about it.
- 2. Everything else (lectures and readings and projects) is fair game.

- 1. We didn't cover NLTK in the lectures, so don't worry about it.
- 2. Everything else (lectures and readings and projects) is fair game.
- 3. Aside: It has been brought to my attention that NLTK is not on the course docker images: I will fix this.

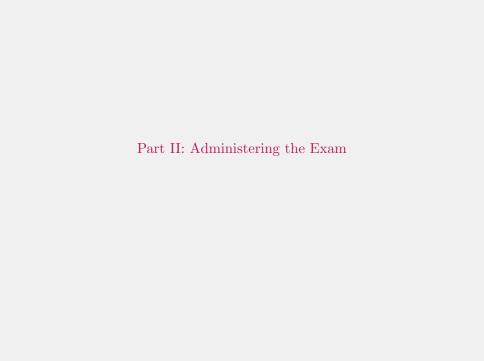
git

1. You should still learn it.

docker

docker

1. I'm not evil.



1. At 00:00 EDT on March $31^{\rm st}$ the exam will go live on ELMS

- 1. At 00:00 EDT on March $31^{\rm st}$ the exam will go live on ELMS
- 2. At 23:59 EDT on March 31st the exam will close on ELMS

- 1. At 00:00 EDT on March $31^{\rm st}$ the exam will go live on ELMS
- 2. At 23:59 EDT on March 31st the exam will close on ELMS
- 3. We will only answer clarifying questions that are asked in private

- 1. At 00:00 EDT on March 31^{st} the exam will go live on ELMS
- 2. At 23:59 EDT on March 31st the exam will close on ELMS
- 3. We will only answer clarifying questions that are asked in private
- 4. This is the only thing this semester where public discussion is forbidden.

- 1. At 00:00 EDT on March 31^{st} the exam will go live on ELMS
- 2. At 23:59 EDT on March 31st the exam will close on ELMS
- 3. We will only answer clarifying questions that are asked in private
- 4. This is the only thing this semester where public discussion is forbidden.
- 5. Please respect these rules.



1. Exam is designed for approx. 90 minutes

- 1. Exam is designed for approx. 90 minutes
- 2. Feel free to typeset, our eyes will thank you

- 1. Exam is designed for approx. 90 minutes
- 2. Feel free to typeset, our eyes will thank you
- 3. Explain. your. process.

- 1. Exam is designed for approx. 90 minutes
- 2. Feel free to typeset, our eyes will thank you
- 3. Explain. your. process.
- 4. Nothing will be accepted after 23:59 EDT on March 31st, 2021 CE (AD).

Thanks for your time!

:)