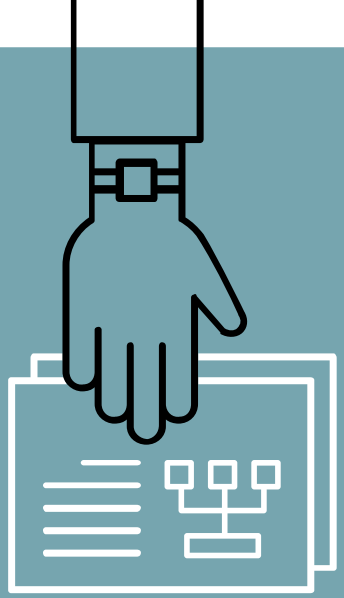
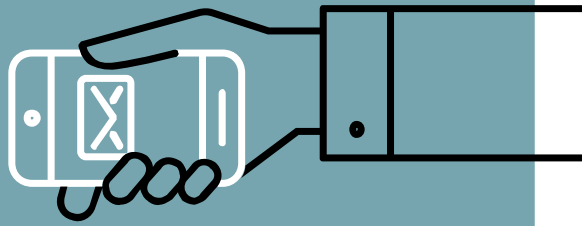
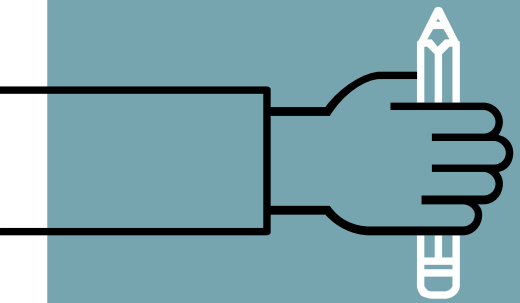
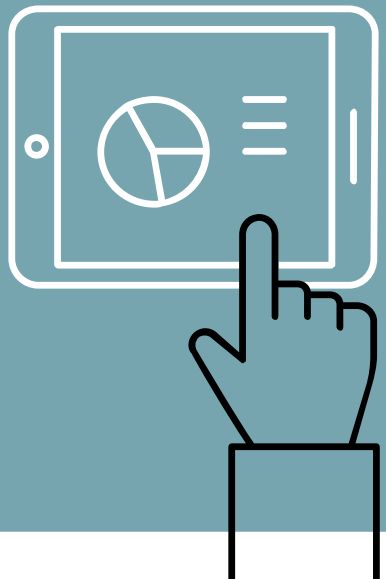




BST 270
Reproducible Data Science

Winter 2021
Session 5





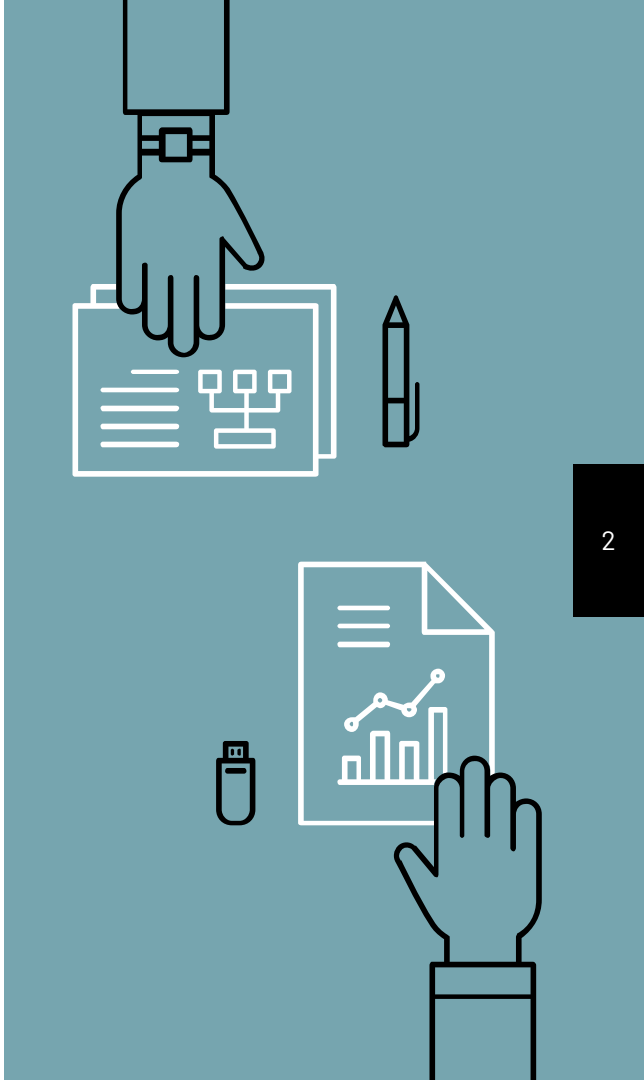
Module 5 Discussion

Do people still do cross validation if they have a replication dataset?

- If I understand the question correctly, yes. Cross validation is like having multiple validation sets to test the performance and generalizability of a model. A replication dataset is for replicating the results of the model on a completely different dataset.

In the examples they split the dataset into the number of rows contained in the dataset and have the training dataset made up of every observation except one, and the testing dataset be made up of one observation. What is the advantage to doing this over something like 5-fold cross validation (which will have more than one observation in the test dataset)?

- Leave one out cross validation (LOOCV) is a special case of k-fold cross-validation where k is equal to the size of data (n). LOOCV is ideal for small datasets (not many rows) because it allows for the largest amount of training data. However, it is prone to higher variance since the validation set is only 1 observation. If you have a larger number of rows, I would suggest using k-fold cross-validation to avoid this variance issue.



Module 5 Discussion

Although it similar to residual errors, how popular is the brier score in statistics? I haven't heard about it before today.

- In my experience it's more common in machine learning when checking the calibration of a model. I like this post and I think it explains it really well.

The Brier Score and AUC are used for binary outcome variables, but what validation metrics are used for categorical outcome variables that have 3 or more levels?

- You can still use the Brier score for categorical outcome variables. The only difference is using 1 for the correct class (category) and 0 if it's anything else (all other categories) - so one category versus all of the others rather than one versus another.

Module 5 Discussion

If you're testing/developing some new method, how should you balance between simulated data resembling a real dataset you're working with vs. satisfying the assumptions you're making about method (even if the dataset might not satisfy them)?

- I've seen researchers create multiple simulated datasets. One to show how their method works when all assumptions are met, and one or more that resemble empirical data or data that doesn't meet all of the assumptions. That way you can also see how your method breaks down when the assumptions aren't met.

Can we go over the intuition behind the estimate of concordance measure?

- Concordance, or the C-statistic, is interpreted as the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome occur than a randomly selected subject who did not experience the outcome.
- I like this resource

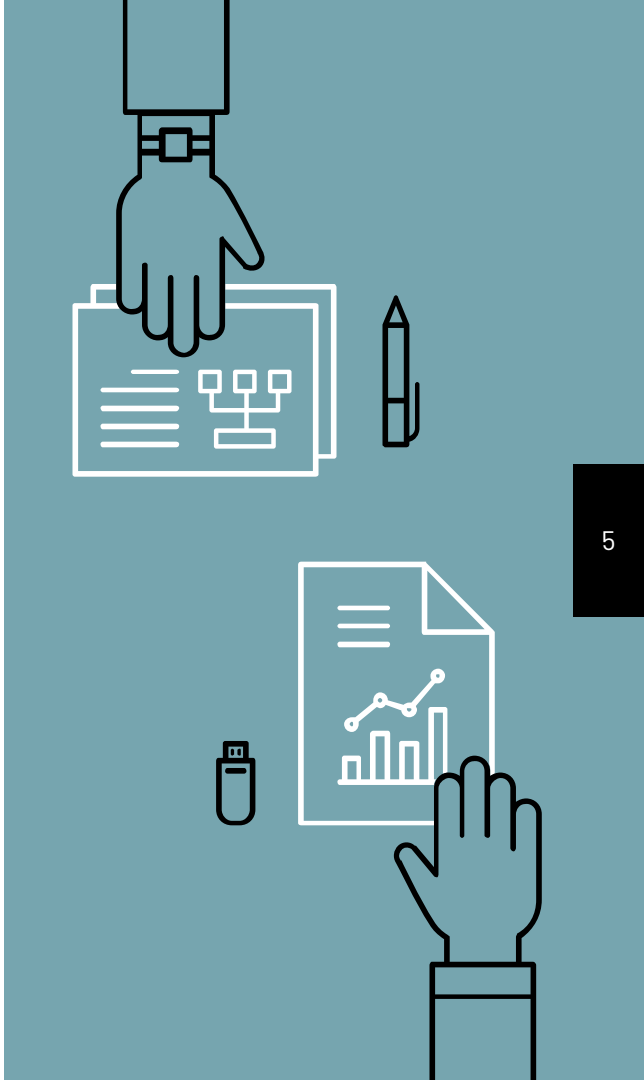
Module 5 Discussion

Is the only difference between cross validation and cross-study validation the number of datasets used?

- No. Cross validation uses one dataset and splits it into some number *k* groups for validation. Cross-study validation uses multiple independent datasets for validation (similar to replication).

Do you have strategies for communicating stats measures (AUC, p-values, regression outputs) to non-stats people who are probably more interested in the scientific result rather than the math behind it? How much of the math do you explain?

- This is such an important skill! And it takes years to master. If I am explaining this to clinicians or a very non-technical audience I don't mention/explain any of the math and just say what the outcome/result means in the context of the problem. For example: "The AUC for our model is .9, meaning our model is better at predicting cardiovascular disease than the standard model." (In this example the other AUC was .6)
- If the audience is a bit more technical, I might go into more detail but still summarize results without using a ton of jargon.



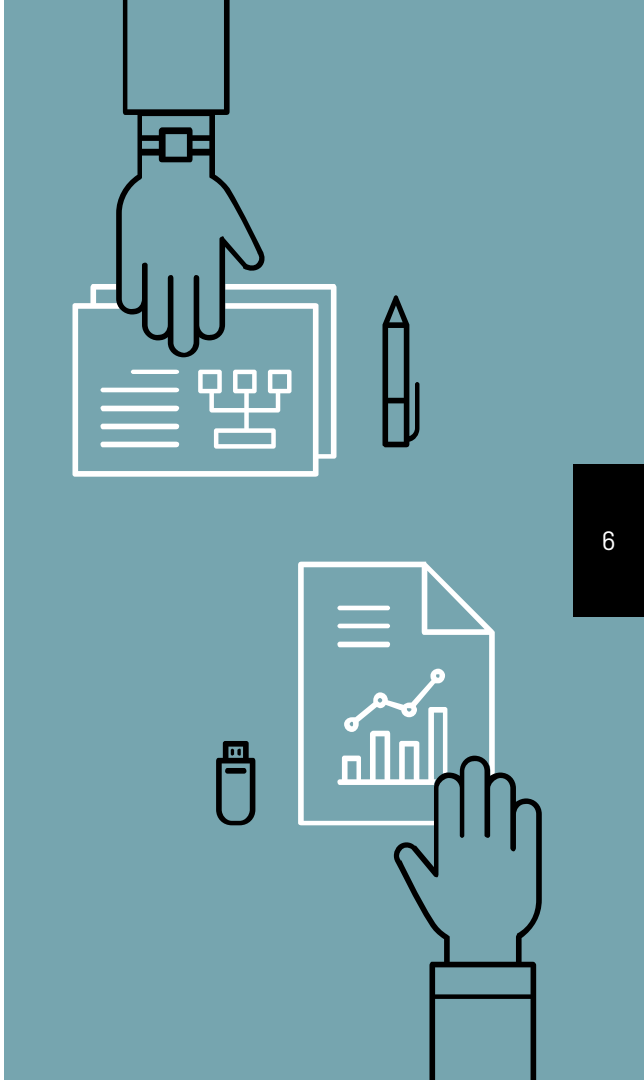
Module 5 Discussion

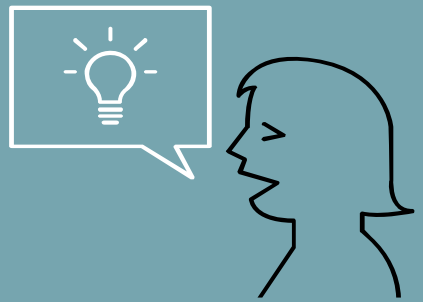
I understand the importance of all of the topics covered in this module, but how do they tie to reproducibility? For instance, using validation metrics to assess if a prediction model is reliable and using cross validation to select an algorithm are both clearly important in ensuring your research is valid, but it seemed like this section was a bit disconnected from the previous module in terms of staying on the topic of reproducibility.

- I agree and to be honest I'm not sure this module really fits in. The main purpose of the module is exactly what you pointed out and it is really about making sure your methods/results are correct/valid before sharing with other researchers.

What exactly are the pros and cons of using the AUC vs. Brier score for binary outcomes?

- AUC is more common so people will usually know what you're talking about, and the Brier score isn't as well known. AUC is also a little easier to interpret/understand, but the Brier score gives more information about model performance. I personally like to report both when applicable.





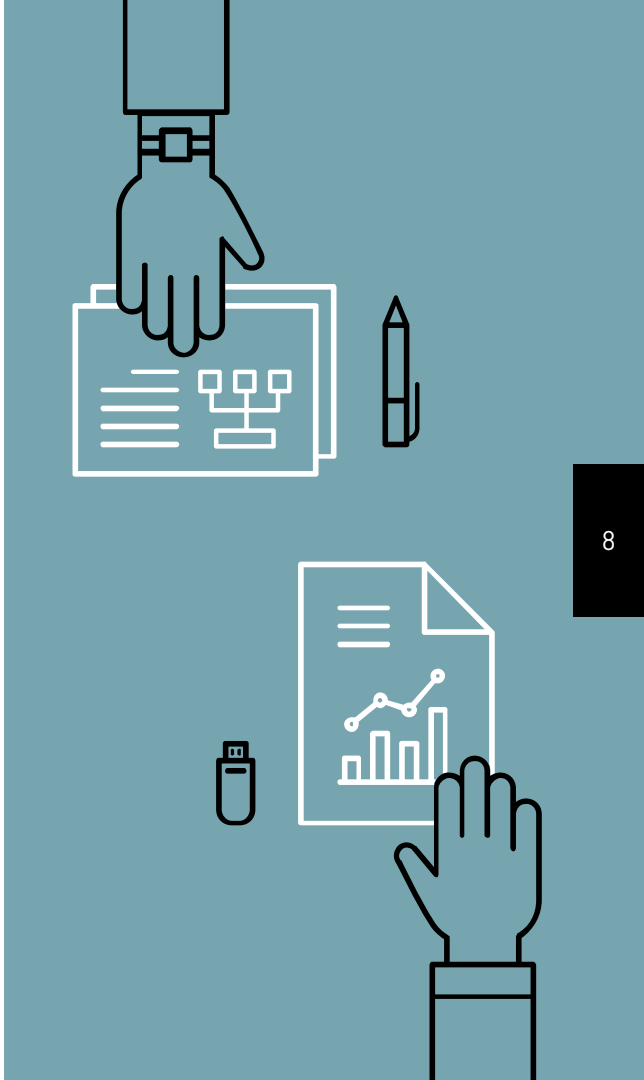
In-Class Project

We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility

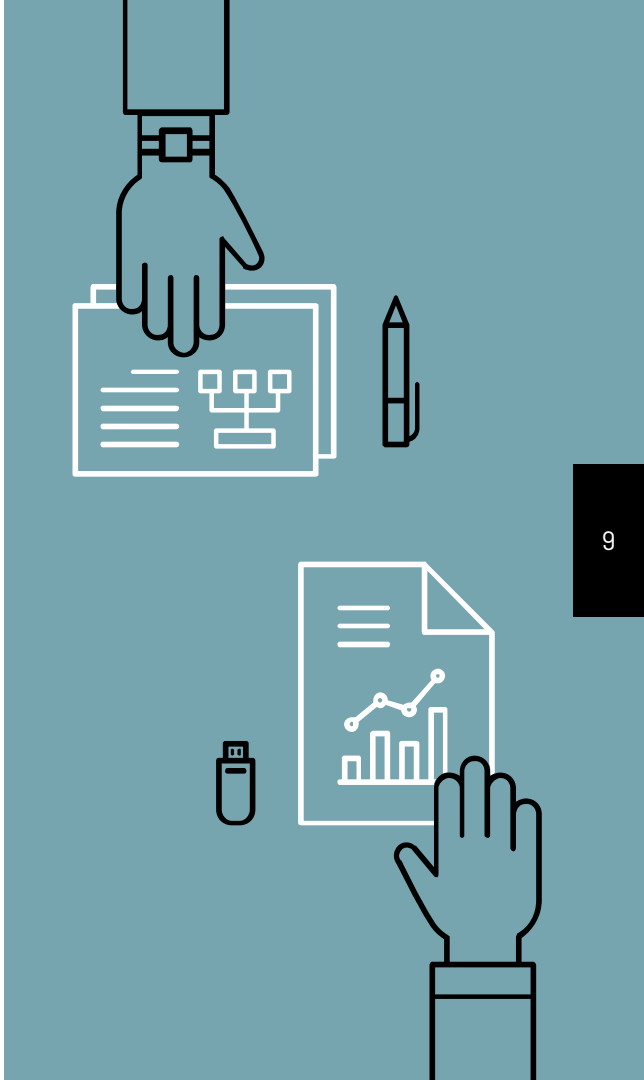


We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility



Homework

- Watch Module 6 part 1 videos
  - 6.1.1 - 6.3.10

- Submit Module 6 part 1 discussion points