Analysis of CSV File Reading in R

Quang Nguyen* and Robert Tedesco*

* Loyola University Chicago

STAT 407 Fall 2021

Outline

- Background
- Experiment
- ANOVA
- Additional Analysis
- Discussion

Background

- Statisticians struggle with reading in large dataset (Matthews, 2021)
- `read.csv()`, part of `base R` (R Core Team, 2021), is old and slow.
- There are better and more efficient data I/O algorithms in `R` (Gillespie & Lovelace, 2021)
  - `rio::import()` (Chan et al., 2021)
  - `readr::read_csv()` (Wickham & Hester, 2021) - part of `tidyverse` (Wickham et al., 2019)
  - `data.table::fread()` (Dowle & Srinivasan, 2021)

Experiment

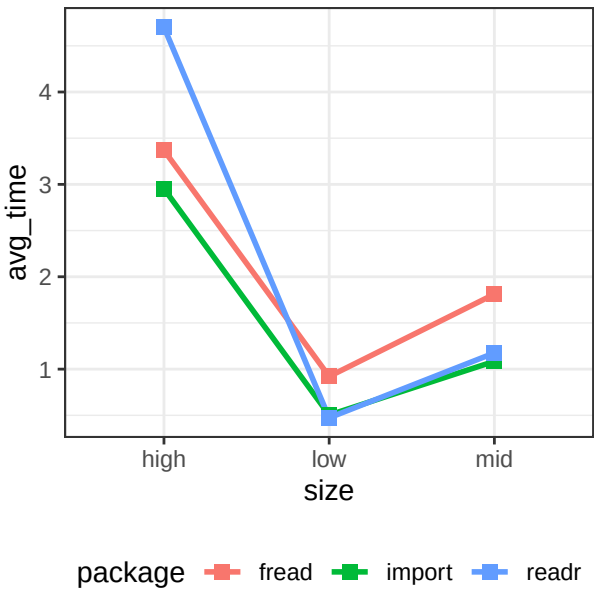
- 3^3 factorial design (Montgomery, 2012)
  - File size: low (~0.5GB), medium (~1.2GB), high (~2GB)
  - Computer: Lenovo, Dell, MacBook
  - Function: `rio::import()`, `readr::read_csv()`, `data.table::fread()`
- $n = 20$ replicates
- $y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$
  - $i, j, k = 1, 2, 3$
  - $l = 1, 2, ..., 20$

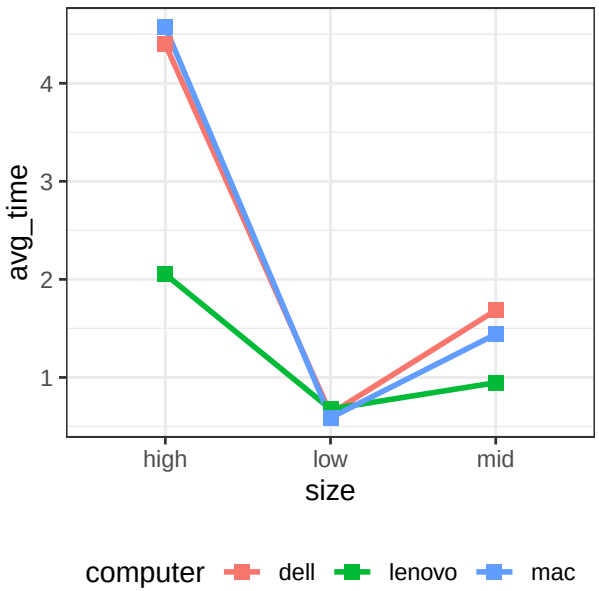
Data

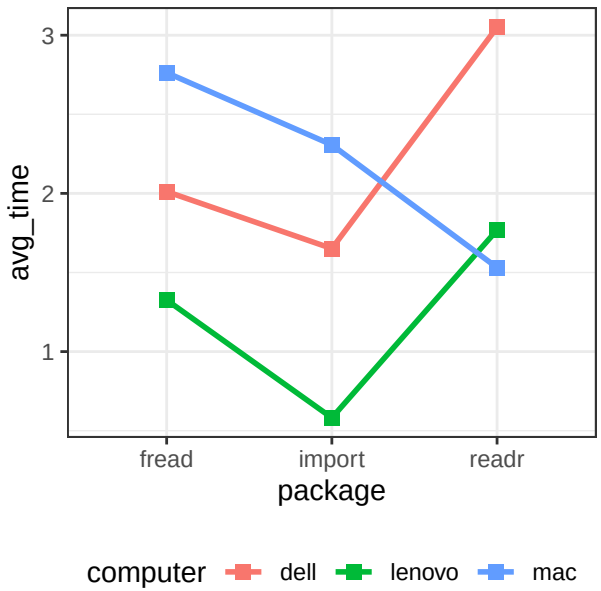
- Generate 3 files of different sizes consisting of columns of randomly-draw samples from $N(0, 1)$, then export to CSV.
- 3^3 possible combinations of `R` package, file size, and computer.
- Use the 3 reading functions to load the CSV files, then `Sys.time()` to measure time elapsed.
- Obtain 20 replicates for each combination according to the order of a random sample of 1 through 3^3.

Data Understanding

- Interaction plots







ANOVA

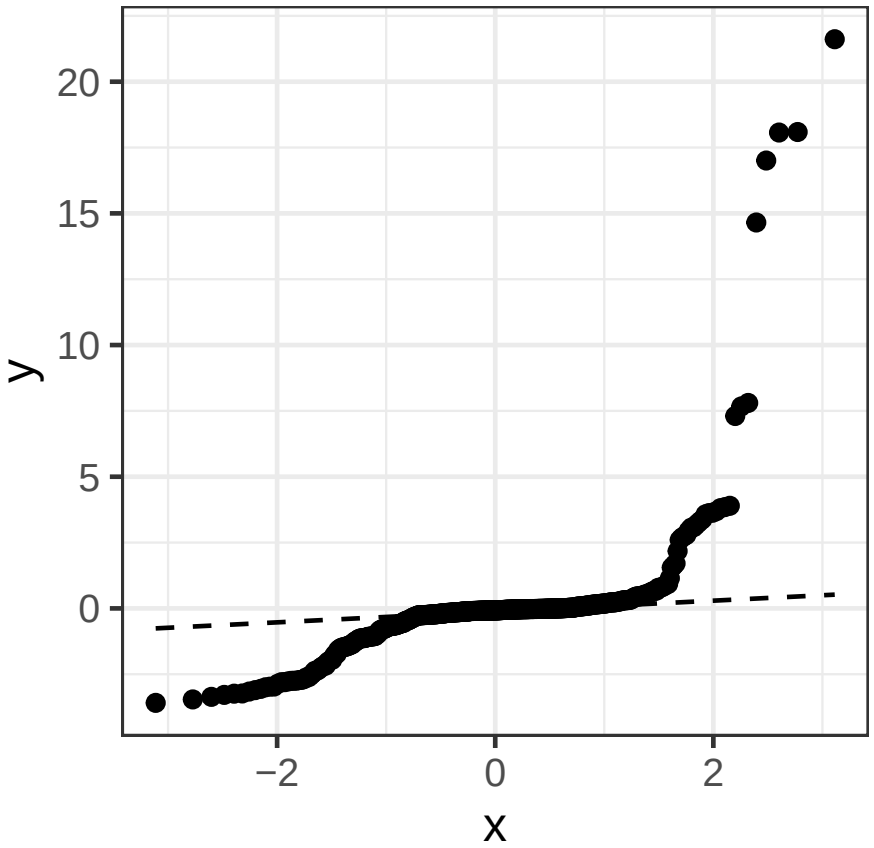
- All model terms are significant. . .

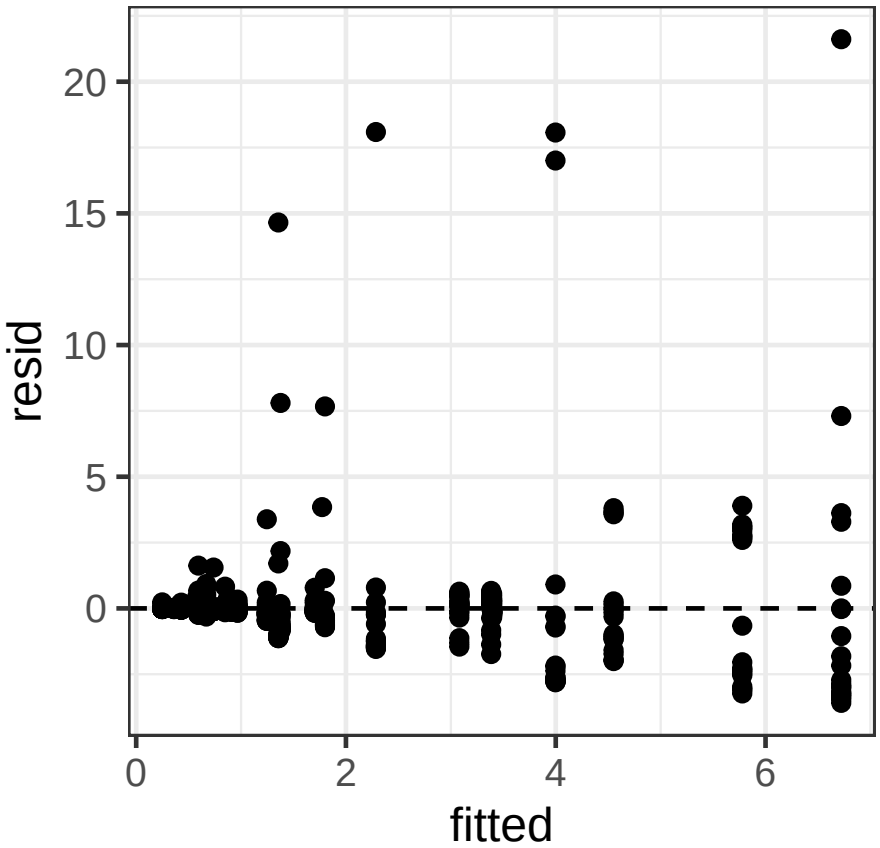
Term	df	SS	MS	F	p-value
size	2	908.9852	454.4926	98.1034	0.0000
computer	2	118.6155	59.3078	12.8017	0.0000
package	2	38.4974	19.2487	4.1549	0.0162
size:computer	4	136.3553	34.0888	7.3582	0.0000
size:package	4	87.9878	21.9970	4.7481	0.0009
computer:package	4	114.5762	28.6441	6.1829	0.0001
size:computer:package	8	144.3655	18.0457	3.8952	0.0002
Residuals	513	2376.6214	4.6328	NA	NA

- However . . .

ANOVA

- There are issues with both **normality** and **homoscedasticity**





- Confirmed by Shapiro-Wilk and Levene tests (both with p-value ≈ 0)

ANOVA

- Transforming the data (log, Box-Cox) did not fix issues with model assumptions
- This leads us to consider a nonparametric approach,
  - No distribution assumption about the data
  - Permutation test for a three way factorial designs
  - `asbio::perm.fact.test()` (Aho, 2021)

ANOVA

- A permutation test gives significant results

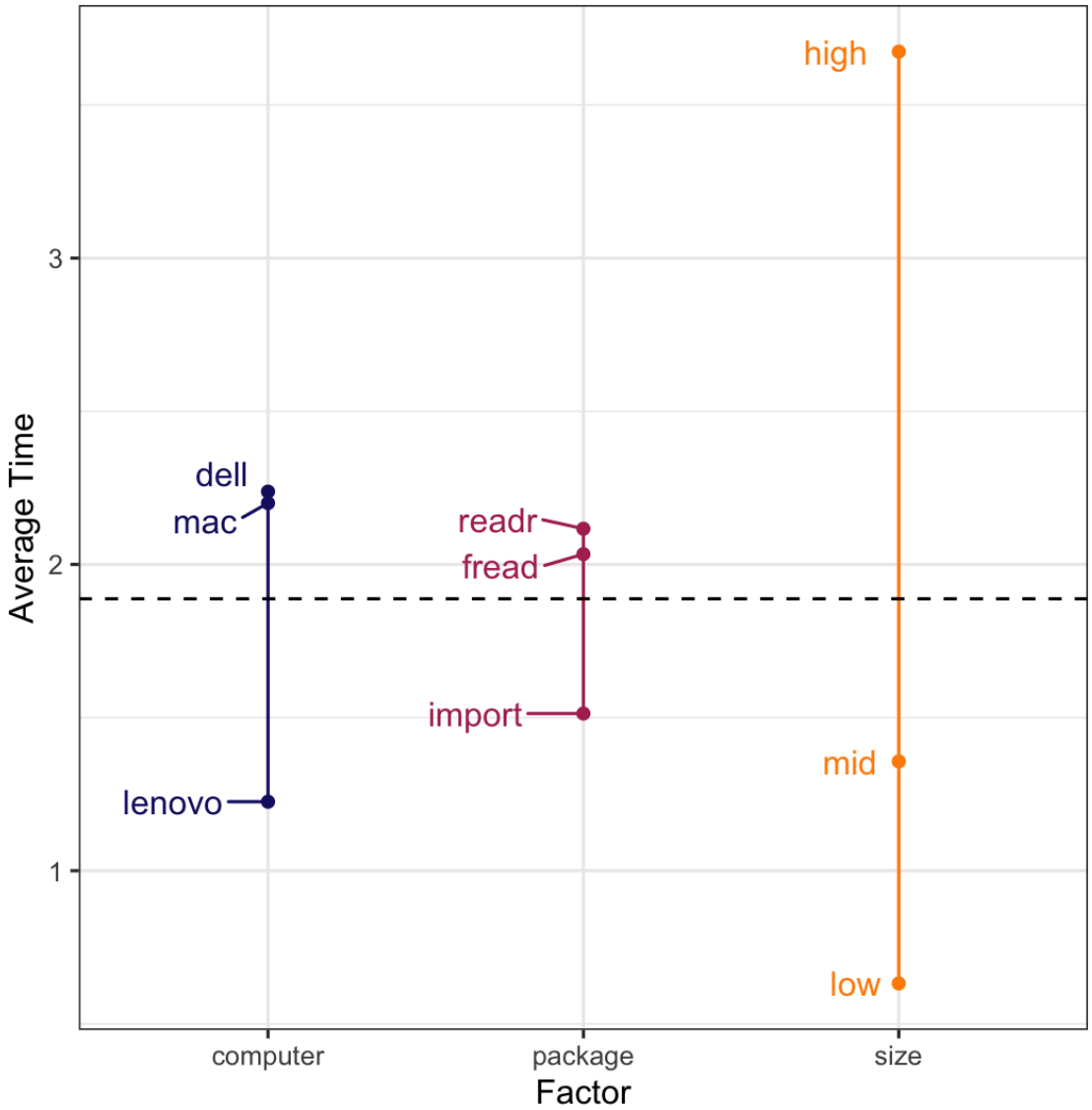
Term	Initial F	df	p-value
size	98.1034	2	0.0001
computer	12.8017	2	0.0001
package	4.1549	2	0.0123
size:computer	7.3582	4	0.0001
size:package	4.7481	4	0.0007
computer:package	6.1829	4	0.0002
size:computer:package	3.8952	8	0.0003
Residual	NA	513	NA

ANOVA

- Post-hoc analysis: pairwise permutation tests
- Use False Discovery Rate (Benjamini & Hochberg, 1995) as p-value adjustment method
- Which pairs differ in mean reading time?
  - Speed: all pairs
  - Computer: Lenovo-Dell, Lenovo-Mac
  - Package/Function: `data.table::fread()` - `rio::import()`, `rio::import()` - `readr::read_csv()`

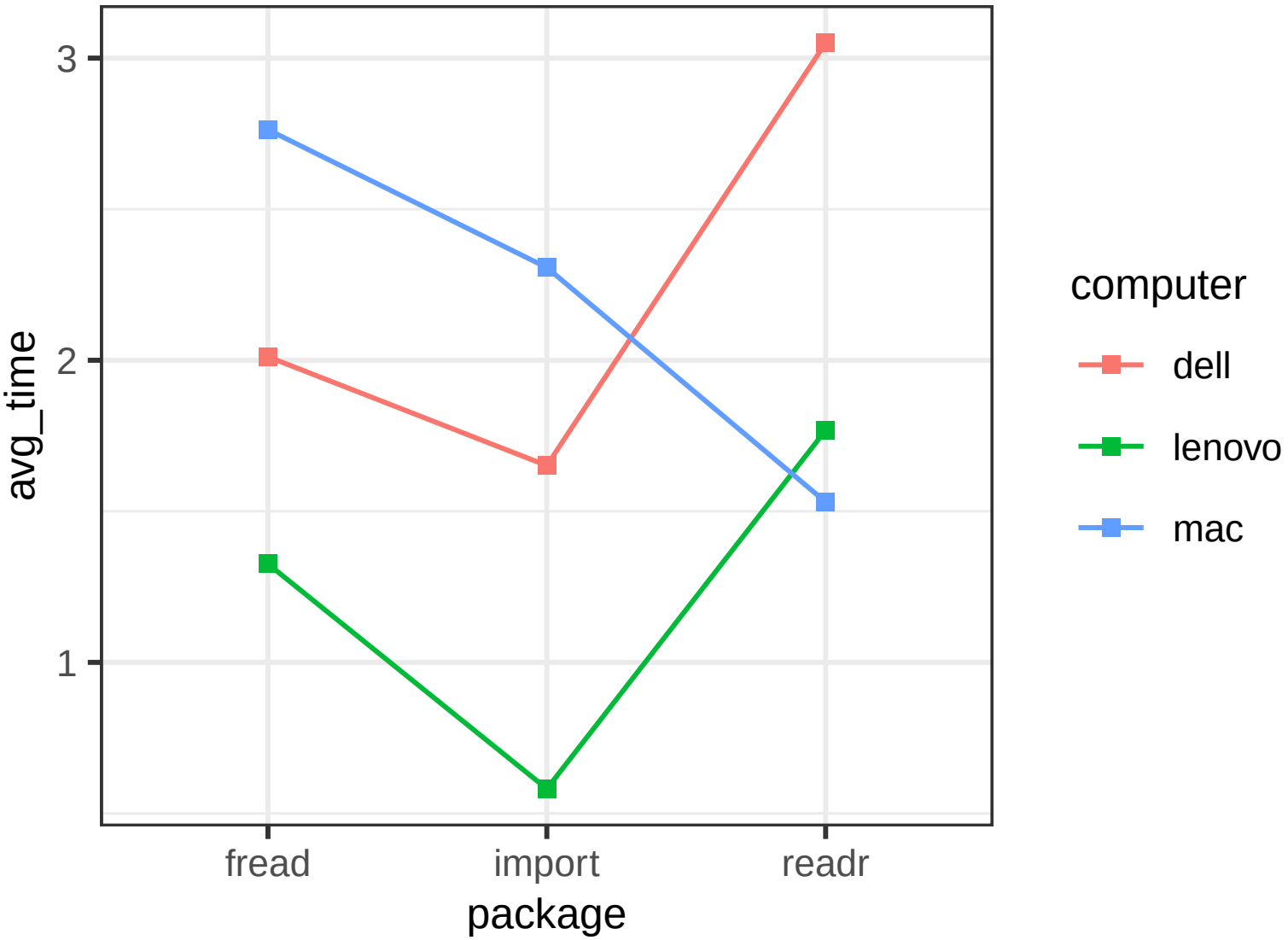
ANOVA

- Univariate effects plot of the factors



Additional Analysis

- Investigate reading method (package/function) for MacBook only

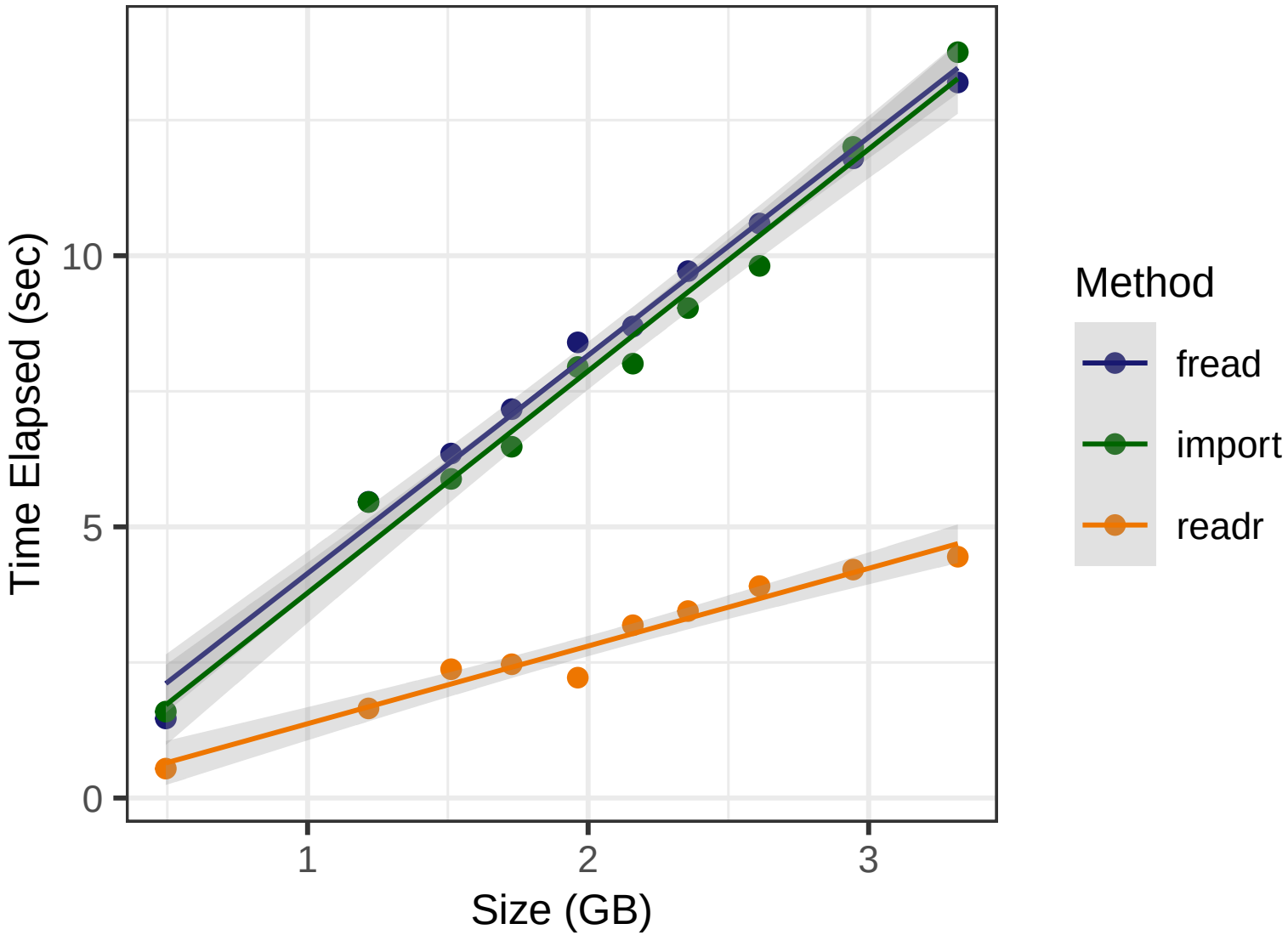


Additional Analysis

- A simulation study
- Simulate more data, in addition to the existing 3 data files
- Consider 10 file sizes (in GB)
  - **0.49**, **1.22**, 1.51, 1.73, **1.96**, 2.16, 2.36, 2.61, 2.94, 3.32
- Use MacBook to read in each file using the 3 functions, and record the time elapsed

Additional Analysis

- Clearly there's a difference in algorithm reading time when using Mac



Additional Analysis

- ANCOVA
  - Response: time elapsed (in seconds)
  - Treatment: function (3 levels)
  - Covariate: file size (in GB)

Term	SS	df	F	p-value
(Intercept)	10.5154	1	8.4657	0.0073
meth	187.3580	2	75.4188	0.0000
size	192.4662	1	154.9502	0.0000
Residuals	32.2950	26	NA	NA

Additional Analysis

- Model assumptions are met.

statistic	p.value	method
0.9732	0.6286	Shapiro-Wilk normality test

statistic	p.value	df	df.residual	method
2.3633	0.1133	2	27	Levene test

Additional Analysis

- Multiple comparisons - Tukey contrasts

contrast	estimate	se	statistic	adj p-value
import - fread	-0.2873	0.4984	-0.5764	0.8338
readr - fread	-5.4391	0.4984	-10.9127	0.0000
readr - import	-5.1518	0.4984	-10.3362	0.0000

- `readr` significantly differs from the other 2 methods when using Mac.

Discussion

- Take advantage of more efficient algorithms!!!
- Extend to other file formats (`.tsv`, `.json`...) and other forms of data (images, audio, ...)
- Other variables to consider
  - On battery vs charging
  - Amount of RAM stored in the `R` environment
  - Different environments: RStudio Desktop, RStudio Cloud, Colab Notebook, Terminal/Command Line...
  - Reading files in locally vs from a DropBox/Google Drive link
  - `R` (`readr::read_csv()`) vs `python` (`pandas.read_csv()`)

Cheers.

- Acknowledgments
  - Lance Davis
  - Mike Perry
  - Greg Matthews

- Greg's YouTube video: youtu.be/E5KJkooW4RY
- GitHub: github.com/qntkhvn/read_speed
- Question?

References I

Aho, K. (2021). *Asbio: A collection of statistical tools for biologists*. https://CRAN.R-project.org/package=asbio

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. http://www.jstor.org/stable/2346101

Chan, C., Chan, G. C., Leeper, T. J., & Becker, J. (2021). *Rio: A swiss-army knife for data file i/o*. https://CRAN.R-project.org/package=rio

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

Gillespie, C., & Lovelace, R. (2021). *Efficient r programming*. https://csgillespie.github.io/efficientR

Matthews, G. (2021). *Old dog (statistician) learns new trick (read_csv)*. https://youtu.be/E5KJkooW4RY.

References II

Montgomery, D. C. (2012). *Design and analysis of experiments, 8th edition*. John Wiley & Sons.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Hester, J. (2021). *Readr: Read rectangular text data*. https://CRAN.R-project.org/package=readr