UNIVERSITY OF SÃO PAULO
SCHOOL OF ARTS, SCIENCES AND HUMANITIES
GRADUATE PROGRAM IN MODELING COMPLEX SYSTEMS

Daniel Vartanian

Ecology of sleep and circadian phenotypes of the Brazilian population

São Paulo
2023

Daniel Vartanian

Ecology of sleep and circadian phenotypes of the Brazilian population

Preliminary version

Dissertation presented to the School of Arts, Sciences and Humanities at the University of São Paulo as part of the requirements for the degree of Master of Science by the Graduate Program in Modeling Complex Systems (PPG-SCX).

Area of Concentration: Fundamentals of Complex Systems.

Supervisor: Prof. Dr. Camilo Rodrigues Neto

São Paulo

2023

ERRATA

Qualifying exam text by Daniel Vartanian, titled **“Ecology of sleep and circadian phenotypes of the Brazilian population”**, presented to the School of Arts, Sciences and Humanities at the University of São Paulo, as part of the requirements for the degree of Master of Science by the Graduate Program in Modeling Complex Systems (PPG-SCX), in the concentration area of Fundamentals of Complex Systems.

Approved on _____ , _____ .

Examination committee

Committee chair:

Prof. Dr. _____

Institution _____

Examiners:

Prof. Dr. _____

Institution _____

Evaluation _____

Prof. Dr. _____

Institution _____

Evaluation _____

Prof. Dr. _____

Institution _____

Evaluation _____

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

ACKNOWLEDGEMENTS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nullius in verba

(The Royal Society, n.d.)

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

RESUMO

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

LIST OF FIGURES

Figure 1 – Distribution of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype. The categorical cuts follow a quantile approach going from extremely early $(0|-0.11)$ to the extremely late $(0.88-1)$. 44

Figure 2 – Distribution of mean aggregates of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, with latitude decimal degree intervals. Higher values of MSF_{sc} indicate a late chronotype tendency. 47

Figure 3 – **?(caption)** 74

Figure 4 – **?(caption)** 74

Figure 5 – **?(caption)** 75

Figure 6 – **?(caption)** 76

Figure 7 – **?(caption)** 77

Figure 8 – **?(caption)** 79

Figure 9 – **?(caption)** 80

Figure 10 – **?(caption)** 81

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

$_F$	Subscript indicating a relation with work-free days.
$_W$	Subscript indicating a relation with workdays.
BT	Local time of going to bed.
FD	Number of work-free days per week.
GU	Local time of getting out of bed.
HO	Horne & Ostberg's morningness-eveningness questionnaire (same as *MEQ*).
LE	Light exposure.
LE_{week}	Average weekly light exposure.
MCTQ	Munich ChronoType Questionnaire.
$MCTQ^{PT}$	Portuguese version of the MCTQ.
MEQ	Morningness-Eveningness Questionnaire.
MSF	Local time of mid-sleep on work-free days.
MSF_{sc}	Chronotype proxy. The midpoint between sleep onset and sleep end on work-free days. A sleep correction ($_{SC}$) is made when a possible sleep compensation related to a lack of sleep on workdays is identified.
MSW	Local time of mid-sleep on workdays.
PRC	Phase response curve.
SD	Sleep duration.
SD_{week}	Average weekly sleep duration.
SE	Local time of sleep end.

SI "Sleep inertia". Despite the name, this abbreviation represents the time the respondent takes to get up after sleep end. It is used this way by the MCTQ authors.

SJL Absolute social jetlag.

SJL_{rel} Relative social jetlag.

SJL_{sc} Jankowski's sleep-corrected social jetlag.

$SJL_{sc\text{-}rel}$ Jankowski's relative sleep-corrected social jetlag.

$Sloss_{week}$ Weekly sleep loss.

SO Local time of sleep onset.

Slat Sleep latency or time to fall asleep after preparing to sleep on work-days.

SPrep Local time of preparing to sleep.

TBT: Total time in bed.

WD Number of workdays per week.

Table of contents

Table of contents 13

Welcome 17

Citation 17

I Preliminary Pages **19**

Errata 20

Inscription 21

Acknowledgments 22

Epigraph 23

Abstract 24

Resumo (PT-BR abstract) 25

Thesis outline 26

Abbreviations 27

Terms and definitions 29

II Chapters **31**

1 Introduction 32

1.1 On chronobiology 32

1.2 On complex systems 32

1.3 On the entrainment phenomenon 32

1.4 Aims 32

1.5 Projects developed 32

1.6 Other related activities 32

1.6.1 Classes . . . 32
1.6.2 Teaching internship . . . 33
1.6.3 Publications . . . 33
1.6.4 Conferences . . . 33
1.7 Research compediums . . . 34
1.8 Data models & Data plans . . . 34
1.9 Softwares . . . 34
1.9.1 Projects . . . 35

2 Similarities between different versions of the MCTQ-PT . . . 36
2.1 Study framework . . . 36

3 The {mctq} R package . . . 37
3.1 Study framework . . . 37

4 Ecology of sleep and circadian phenotypes of the Brazilian population . . . 38
4.1 Study framework . . . 38

5 Rule-based model of the 24h light/dark cycle entrainment phenomenon . . . 39
5.1 Study framework . . . 39

6 A biological approach for the latitudinal cline of the chronotype 40
6.1 Study framework . . . 40
6.2 Abstract . . . 40
6.3 Main text . . . 41
6.3.1 Introduction . . . 41
6.3.2 Results . . . 43
6.3.3 Discussion . . . 45
6.4 Methods . . . 48
6.4.1 Ethics information . . . 48
6.4.2 Measurement instrument . . . 48
6.4.3 Sample . . . 49
6.4.4 Analysis . . . 50

6.4.5 Data availability . . . 51
6.4.6 Code availability . . . 52
6.5 Acknowledgments . . . 52
6.6 Ethics declarations . . . 52
6.6.1 Competing interests . . . 52
6.7 Additional information . . . 52
6.8 Rights and permissions . . . 53

7 Discussion and conclusions . . . 54
7.1 Discussion . . . 54
7.2 Limitations . . . 54
7.3 Conclusions . . . 54
7.4 Future perspectives . . . 54

REFERENCES . . . 55

Appendices 61

A Appendice: Chapter 2 supplemental material . . . 61
A.1 Load and embed texts . . . 61
A.2 Text similarity . . . 64
A.2.1 How similar is the *data questionnaire* when compared to the *EUCLOCK questionnaire*? . . . 66
A.2.2 How similar is the *data questionnaire* when compared to the *MCTQ^{PT} questionnaire*? . . . 68

B Appendice: Chapter 3 supplemental material . . . 71

C Appendice: Chapter 4 supplemental material . . . 72
C.1 Data wrangling . . . 72
C.2 Distribution of main variables . . . 73
C.3 Geographic distribution . . . 75
C.4 Age pyramid . . . 76
C.5 Correlation matrix . . . 77

C.6 Age series . . . 78
C.7 Chronotype . . . 79
C.8 Latitude series . . . 80
C.9 Statistics . . . 81
C.9.1 Numerical variables . . . 81
C.9.2 Sex . . . 82
C.9.3 Sex and Age . . . 85
C.9.4 Longitudinal range . . . 89
C.9.5 Latitudinal range . . . 91
C.9.6 Region . . . 93
C.9.7 State . . . 95

D Appendice: Chapter 5 supplemental material . . . 97

E Appendice: Chapter 6 supplemental material . . . 98
E.1 Hypothesis . . . 98
E.2 Assumptions . . . 99
E.3 Data preparation . . . 100
E.4 Restricted model . . . 100
E.4.1 Residual diagnostics . . . 103
E.4.2 Heteroskedasticity . . . 107
E.4.3 Collinearity diagnostics . . . 108
E.4.4 Measures of influence . . . 109
E.5 Full model . . . 110
E.5.1 Residual diagnostics . . . 113
E.5.2 Heteroskedasticity . . . 116
E.5.3 Collinearity diagnostics . . . 118
E.5.4 Measures of influence . . . 119
E.6 Nested regression models test . . . 119
E.7 Group test . . . 121

Welcome

> **!** Important
>
> You are seeing the printed preliminary version of this master's thesis. Please note that this a reproducible document that includes a lot of data and coding information, hence the thesis is best visualized in its web version: <danielvartan.github.io/mastersthesis>. The web version will have the most updated information.
>
> The final print version is going to be rendered accordingly with USP's guidelines to create theses and dissertations documents and will be available at <teses.usp.br>.

The analyses contained in this document are 100% reproducible. They were performed using the R programming language and the Quarto publishing system (an evolution of R Markdown). Click on the GitHub icon in the menu to access the thesis code repository.

Citation

For attribution, please cite this work as:

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Master's thesis]. School of Arts, Sciences and Humanities, University of Sao Paulo, São Paulo. https://danielvartan.github.io/mastersthesis/

BibTeX citation:

```
@mastersthesis{vartanian2023,
  title = {Ecology of sleep and circadian phenotypes of the Brazilian population},
  author = {Daniel Vartanian},
  year = {2023},
  address = {São Paulo},
  school  = {University of Sao Paulo},
  langid = {portuguese},
  url = {https://danielvartan.github.io/mastersthesis/}
```

```
  note = {Preliminary version}
}
```

Part I

Preliminary Pages

Errata

This is the development version of the thesis (version <1.0.0). Any necessary corrections will be listed here after its approval.

Inscription

I dedicate this work to the skeptics, the radicals, the ignorant, the uncivilized, the subversives, the wild dogs, the irreducibles, the irreconcilables. To the true engines of change. To the destabilizers, who possess equal or greater importance than the stabilizers. To those who act on principle, even knowing that there is no ultimate reward or any meaning in life.

Acknowledgments

To Salete Perroni (Sal), my partner in life and in the fight.

To my Mother, for her unconditional love.

To my sister and my brother, for their love and partnership in life.

To my friends in science, Alicia Rafaelly Vilefort Sales, Juliana Viana Mendes, and Maria Augusta Medeiros de Andrade.

To my friend and Professor Humberto Miguel Garay Malpartida, for his support; for his principles; and for his integrity, which was demonstrated when the need arose.

To Professor Camilo Rodrigues Neto, for introducing me to and teaching me about the science of complex systems since 2012; for guiding my dissertation; for the patience and the virtue in taking on and mediating the process of transitioning my guidance in my master's degree after the breakdown of relations with my former supervisor.

To Professor Carlos Molina Mendes, for the speed, impartiality, patience, and virtue in mediating the process of transitioning my guidance in my master's degree.

To my fellow friends: Alex Azevedo Martins; Amanda Moreira; Augusto Amado, Carina (Cacau) Prado; Cauê Teles; Ítalo Alves Bezerra do Nascimento; Júlia Mafra; Leonardo Kazuhiko Kawazoe; Letícia Nery de Figueiredo; and Reginaldo Noveli.

To President Lula (Yes!), who saved Brazil from fascism and approved the long-overdue adjustments to postgraduate scholarships.

To the local Student Movements, which truly support their category, **unlike** the National Association of Postgraduates (ANPG), the Association of Postgraduates (APG) Helenira 'Preta' Rezende) of the University of São Paulo (USP), and the Central Directory of Students (DCE) Alexandre Vannucchi Leme of USP.

To the Support Program for Student Permanence and Education (PAPFE) of USP, which enabled me to get this far.

To the Coordination for the Improvement of Higher Education Personnel (CAPES), for funding this work and enabling my presence in postgraduate studies.

Epigraph

Nullius in verba (The Royal Society n.d.)

Abstract

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Master's thesis]. School of Arts, Sciences and Humanities, University of São Paulo, São Paulo. https://danielvartan.github.io/mastersthesis/

Keywords: chronobiology. chronotype. circadian phenotypes. sleep. sleep-wake cycle. entrainment. latitude. ecology. mctq.

Resumo (PT-BR abstract)

Vartanian, D. (2023). *Ecologia do sono e de fenótipos circadianos da população brasileira* [Dissertação de Mestrado]. Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo. https://danielvartan.github.io/mastersthesis/

Palavras-chaves: cronobiologia. cronotipo. fenótipos circadianos. sono. ciclo sono-vigília. entrainment. latitude. ecologia. mctq.

Thesis outline

All articles shown in the next chapters are targeted to specific science journals and will be submitted for publication soon.

Abbreviations

> **Note**
>
> You are reading the work-in-progress of this thesis. This chapter should be readable but is currently undergoing final polishing.

- $_{F}$: Subscript indicating a relation with work-free days.
- $_{W}$: Subscript indicating a relation with workdays.
- **BT**: Local time of going to bed.
- **FD**: Number of work-free days per week.
- **GU**: Local time of getting out of bed.
- **HO**: Horne & Ostberg's morningness-eveningness questionnaire (same as *MEQ*).
- **LE**: Light exposure.
- **LE_{week}**: Average weekly light exposure.
- **MCTQ**: Munich ChronoType Questionnaire.
- **$MCTQ^{PT}$**: Portuguese version of the MCTQ.
- **MEQ**: Morningness-Eveningness Questionnaire.
- **MSF**: Local time of mid-sleep on work-free days.
- **MSF_{sc}**: Chronotype proxy. The midpoint between sleep onset and sleep end on work-free days. A sleep correction ($_{SC}$) is made when a possible sleep compensation related to a lack of sleep on workdays is identified.
- **MSW**: Local time of mid-sleep on workdays.
- **PRC**: Phase response curve.
- **SD**: Sleep duration.
- **SD_{week}**: Average weekly sleep duration.
- **SE**: Local time of sleep end.
- **SI**: "Sleep inertia". Despite the name, this abbreviation represents the time the respondent takes to get up after sleep end. It is used this way by the MCTQ authors.
- **SJL**: Absolute social jetlag.
- **SJL_{rel}**: Relative social jetlag.
- **SJL_{sc}**: Jankowski's sleep-corrected social jetlag.
- **$SJL_{sc\text{-}rel}$**: Jankowski's relative sleep-corrected social jetlag.

- **$\mathbf{Sloss_{week}}$**: Weekly sleep loss.
- **SO**: Local time of sleep onset.
- **Slat**: Sleep latency or time to fall asleep after preparing to sleep on workdays.
- **SPrep**: Local time of preparing to sleep.
- **TBT**: Total time in bed.
- **WD**: Number of workdays per week.

Terms and definitions

The definitions below don't pretend to be an exhaustive list of terms about chronobiology and complex systems. They are only here to be a quick support for the reader.

See Marques and Oda (2012) and Aschoff, Klotter, and Wever (1965) to get an extensive list of chronobiology terms and definitions.

- **Biological rhythms**: Any rhythms present in living beings. These can be classified as *exogenous* (not related to the biological unit) and *endogenous* (rhythms that originated within the biological unit) (Aschoff 1981). Examples: menstrual cycle (infradian rhythm); sleep-wake cycle (circadian rhythm); cardiac cycle (ultradian rhythm).
- **Chronotype**: Any kind of temporal phenotype (Ehret 1974; C. S. Pittendrigh 1993). Usually, it refers to circadian phenotypes in a spectrum that goes from morningness to eveningness (Horne and Ostberg 1976; Roenneberg, Wirz-Justice, and Merrow 2003). It can also be seen as an organism's phase of entrainment (Roenneberg et al. 2012).
- **Circadian rhythm**: A rhythm with a period close to a day/24h, an approximation to the period of the earth's rotation (C. S. Pittendrigh 1960). From the Latin *circā*, around, and *dĭes*, day (Latinitium n.d.). Example: the sleep-wake cycle.
- **Complex system**: There are several definitions. Here are some that I found to be of use:
  - "Systems that don't yield to compact forms of representation or description." (David Krakauer apud Mitchell 2013)
  - "A system of many interacting parts where the system is more than just the sum of its parts" (Mark Newman apud Mitchell 2013)
  - Systems with many connected agents that interact and exhibit self-organization and emergence behavior, all without the need for a central controller (adapted from Camilo Rodrigues Neto's definition, supervisor of this thesis).
  - Dialectics at its finest (my working definition).

- **Entrainment**: A shift and alignment of biological rhythms induced by a zeitgeber input (Kuhlman, Craig, and Duffy 2018). For example: a shift/alignment of an organism's circadian rhythm when exposed to light.
- **Infradian rhythm**: A rhythm with a period greater than a day/24h. From the Latin *infrā*, below, and *dĭes*, day (Latinitium n.d.). Example: the menstrual cycle.
- **Period**: Cycle duration of an oscillation or, in a more technical way, the duration between two identical and consecutive phases in an oscillation (Kuhlman, Craig, and Duffy 2018).
- System theory: Two definitions can be of use:
  - Science or discipline that investigates models, principles, and laws that are valid to systems in general (Bertalanffy 1968).
  - "The attempt of a reductionist scientific tradition to come to terms with complexity, nonlinearity, and change through sophisticated mathematical and computational techniques, *a groping toward a more dialectical understanding* that is held back by its philosophical biases and the institutional and economic contexts of its development." (Levins 1998)
- **Ultradian rhythm**: A rhythm with a period below a day/24h. From the Latin *ultrā*, beyond, and *dĭes*, day (Latinitium n.d.). Example: the cardiac cycle.
- **Zeitgeber**: Any periodic environmental signal/cue that can influence or regulate biological rhythms. From the German *zeit*, time, and *geber*, donor (Cambridge University Press n.d.). Two main well known zeitgebers are light exposure and environment temperature (C. S. Pittendrigh 1960).

Part II

Chapters

1 Introduction

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

1.1 On chronobiology

1.2 On complex systems

1.3 On the entrainment phenomenon

1.4 Aims

1.5 Projects developed

1.6 Other related activities

1.6.1 Classes

While this thesis was developed, the following Graduate classes from the University of São Paulo (USP) were completed.

- 2022/2: *SCX5000 - Mathematical and Computational Methods I* (10 credits) (Concept: **C**);
- 2022/2: *SCX5002 - Complex Systems I* (10 credits) (Concept: **A**);
- 2023/1: *SCX5001 - Mathematical and Computational Methods II* (10 credits) (Concept: **A**);
- 2023/1: *SCX5017 - Introduction to Data Science* (10 credits) (Concept: **A**);
- 2023/1: *EAH5001 - Pedagogic Preparation* (4 credits) (Concept: **A**).

Please note that the unfortunate **C** concept above happened in the same semester when the author broke relations with his former supervisor (*Mario Pedrazzoli*).

In total, 44 credits were completed until the thesis publication date. Another 12 special credits related to an article publication (see Viana-Mendes et al. (2023)) were requested, according to program regulations, to the Graduate Program Coordination Commission (CCP) and should be evaluated by the end of October 2023. A minimum of 50 credits is needed for the thesis defense.

1.6.2 Teaching internship

Scholarship students of the Coordination for the Improvement of Higher Education Personnel (CAPES) must participate in the Teaching Improvement Program (PAE). This internship is ongoing and will end in December 2023.

The internship activities comprise being an Assistant Professor for USP's undergraduate class *ACH0042 - Problem-Based Learning II*. A teaching plan was written while taking the *EAH5001* graduate class mentioned above and can be accessed at the link below.

Vartanian, D., Bernardes, M. E. M., & Rodrigues Neto, C. (2023). *Plano de ensino: ACH0042 - Resolução de Problemas II*. https://doi.org/10.13140/RG.2.2.33335.50086

1.6.3 Publications

The following article was published during the development of this thesis.

Viana-Mendes, J., Benedito-Silva, A. A., Andrade, M. A. M., Vartanian, D., Gonçalves, B. da S. B., Cipolla-Neto, J., & Pedrazzoli, M. (2023). Actigraphic characterization of sleep and circadian phenotypes of PER3 gene VNTR genotypes. *Chronobiology International*. https://doi.org/10.1080/07420528.2023.2256858

1.6.4 Conferences

An abstract about the main investigation was published and presented on a poster in the Sao Paulo School of Advanced Science on Ecology of Human Sleep and Biological Rhythms of the São Paulo Research Foundation (FAPESP). This was an in-

ternational school with 100 students and young researchers (50 from all states of Brazil and 50 international) that was held between 2022-11-16 to 2022-11-26.

Vartanian, D., & Pedrazzoli, M. (2022). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Poster]. São Paulo Research Foundation; São Paulo School of Advanced Science on Ecology of Human Sleep and Biological Rhythms. https://doi.org/10.13140/RG.2.2.25343.07840

1.7 Research compediums

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Research compendium]. https://danielvartan.github.io/mastersthesis/

1.8 Data models & Data plans

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Data Management Plan]. DMPHub. https://doi.org/10.48321/D1DW8P

1.9 Softwares

Vartanian, D. (2022). *{entrainment}: a rule-based model of the 24h light/dark cycle entrainment phenomenon* [Software, Python library]. https://github.com/danielvartan/entrainme

Vartanian, D. (2023). *{mctq}: tools to process the Munich ChronoType Questionnaire (MCTQ)* [Software, R Package v0.3.2]. https://docs.ropensci.org/mctq/

Vartanian, D. (2023). *{lockr}: easily encrypt/decrypt files* [Software, R package v0.3.0]. https://github.com/danielvartan/lockr

Vartanian, D. (2023). *{lubritime}: an extension for the lubridate package* [Software, R package]. https://github.com/danielvartan/lubritime

Vartanian, D. (2023). *{tesesusp}: a Quarto format for USP theses and dissertation* [Software, LaTeX/R format, v0.1.0]. https://github.com/danielvartan/tesesusp/

1.9.1 Projects

The author is also currently working on the development of the project below.

Sales, A. R. V., Vartanian, D., Andrade, M. A. M., Pedrazzoli, M. (2023). *Associations between the duration and quality of sleep in third-trimester pregnant women and the duration of labor* [Master's project]. School of Arts, Sciences and Humanities, University of Sao Paulo, São Paulo. https://bit.ly/3S6O0MB

2 Similarities between different versions of the MCTQ-PT

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

> **i** Target
>
> 1. Chronobiology International (IF 2022: 2.8/JCR | A1/2017-2020).
> 2. Journal of Biological Rhythms (IF 2022: 3.5/JCR | A2/2017-2020).

> **i** Note
>
> The following study was performed by **Daniel Vartanian (DV)** and Camilo Rodrigues Neto (CR).
>
> **DV** and **CR** contributed to the design, implementation and statistical analysis of the study and . **DV** wrote the manuscript. All authors discussed the results and revised the final manuscript.
>
> (*Future reference*): Vartanian, D., & Rodrigues Neto, C. (2024). Similarities between different versions of the MCTQ-PT. *Chronobiology International*.

2.1 Study framework

3 The {mctq} R package

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

> **i** Target
>
> 1. Journal of Statistical Software (IF 2022: 5.8/JCR | A1/2017-2020).
> 2. Journal of Open Source Software (B1/2017-2020).

> **i** Note
>
> The following study was performed by **Daniel Vartanian (DV)**, Ana Amélia Benedito-Silva (AA), Mario Pedrazzoli (MP) and Camilo Rodrigues Neto (CR).
> **DV** contributed to conception, design, coding and implementation of the software. **AA**, **MP** and **CR** contributed as science advisers and reviewers. **DV** wrote the manuscript.
> (*Future reference*): Vartanian, D, Benedito-Silva, A. A., Pedrazzoli, M., & Rodrigues Neto, C. (2024). {mctq}: tools to process the Munich ChronoType Questionnaire (MCTQ). *Journal of Statistical Software*.

3.1 Study framework

4 Ecology of sleep and circadian phenotypes of the Brazilian population

> **Important**
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

> **Target**
>
> 1. Chronobiology International (IF 2022: 2.8/JCR | A1/2017-2020).
> 2. Journal of Biological Rhythms (IF 2022: 3.5/JCR | A2/2017-2020).

> **Note**
>
> The following study was performed by **Daniel Vartanian (DV)**, Mario Pedrazzoli (MP) and Camilo Rodrigues Neto (CR).
> **DV** and **CR** contributed to the design, implementation and statistical analysis of the study and . **DV** and **MP** collected the data. **DV** wrote the manuscript.
> (*Future reference*): Vartanian, D., Pedrazzoli, M., & Rodrigues Neto, C. (2024). Ecology of sleep and circadian phenotypes of the Brazilian population. *Chronobiology International*.

4.1 Study framework

5 Rule-based model of the 24h light/dark cycle entrainment phenomenon

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

> **i** Target
>
> 1. Journal of Open Source Software (B1/2017-2020).

> **i** Note
>
> The following study was performed by **Daniel Vartanian (DV)** and Camilo Rodrigues Neto (CR).
>
> **DV** contributed to the design, implementation of the software. **CR** contributed as a science adviser and reviewer. **DV** wrote the manuscript. All authors discussed the results and revised the final manuscript.
>
> (*Future reference*): Vartanian, D, & Rodrigues Neto, C. (2024). {entrainment}: A rule-based model of the 24h light/dark cycle entrainment phenomenon. *Journal of Statistical Software*.

5.1 Study framework

6 A biological approach for the latitudinal cline of the chronotype

> **Note**
>
> You are reading the work-in-progress of this thesis. This chapter should be readable but is currently undergoing final polishing.

> **Target**
>
> - Scientific Reports (IF 2022: 4.6/JCR | A1/2017-2020).

> **Note**
>
> The following study was performed by **Daniel Vartanian (DV)**, Mario Pedrazzoli (MP) and Camilo Rodrigues Neto (CR).
> **DV** contributed to the design and implementation of the study. **DV** and **MP** collected the data. **DV** and **CR** performed the statistical analysis. **DV** wrote the manuscript.
> (*Future reference*): Vartanian, D., Pedrazzoli, M., & Rodrigues Neto, C. (2024). A biological approach for the latitudinal cline of the chronotype. *Scientific Reports*.

6.1 Study framework

6.2 Abstract

Chronotypes are temporal phenotypes (Ehret 1974; C. S. Pittendrigh 1993). Observable traits, like weight and eye color. Our current understanding of these traits is that they are linked to our environment and are the result of evolution pressures for creating an inner temporal organization (Aschoff 1989; Paranjpe and Sharma 2005). A way that organisms have found to anticipate events. Having such an important function in nature, these internal rhythms need to be closely aligned with environmental changes. The agents that shift these oscillations towards the environment are called zeitgebers and the shift phenomenon is called entrainment (Roenneberg, Daan, and Merrow 2003; Roenneberg et al. 2010). The main zeitgeber for humans is light exposure, particularly the light of the sun (Khalsa et al. 2003; Minors, Waterhouse, and Wirz-Justice 1991; Roenneberg et al. 2007). Considering the major role of light on entrainment, several

studies hypothesized that the latitude shift of the sun could influence or even define the chronotypes of different populations (Horzum et al. 2015; Hut et al. 2013; Leocadio-Miguel et al. 2017, 2014; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler and Rahafar 2017). For example, populations that live close to the equator would be, on average, more entrained to the light-dark cycle and have morning-leaning characteristics. Here we test this hypothesis using a biological measure, the chronotype state, provided by the Munich ChronoType Questionnaire (Roenneberg, Wirz-Justice, and Merrow 2003). We tested the latitude hypothesis on a sample with $73,825$ subjects living in different latitudes in Brazil. Our results show that, even with a wide, big, and aligned sample, the latitude is associated only with negligible effect sizes. The entrainment phenomenon appears to be much more complex than previously imagined, opening new questions and contradictions that need to be further investigated.

6.3 Main text

6.3.1 Introduction

Humans can differ from one another in many ways. These observable traits, like hair color or height, are called phenotypes and are also presented in the way that our body functions.

A chronotype is a temporal phenotype (Ehret 1974; C. S. Pittendrigh 1993). This word is usually used to refer to endogenous circadian rhythms, i.e., rhythms which periods that are close to a day or 24 hours (*circa diem*). The current body of knowledge of Chronobiology, the science that studies biological rhythms, indicates that the evolution of these internal oscillators is linked to our oscillatory environment, like the day and night cycle, which, along with our evolution, created environmental pressures for the development of a temporal organization (Aschoff 1989; Paranjpe and Sharma 2005). A way in which an organism could predict events and better manage its needs, like storing food for the winter.

But a temporal system wouldn't be of much use if it could not follow environmental changes. To those environmental signals that can regulate the biological rhythms are given the name zeitgeber (from the German Zeit, time, and Geber, giver). These zeitgebers produce inputs in our bodies that can shift and align those rhythms. This phe-

nomenon is called entrainment (Roenneberg, Daan, and Merrow 2003; Roenneberg et al. 2010).

The main zeitgeber known today is the light, particularly the sun's light (Khalsa et al. 2003; Minors, Waterhouse, and Wirz-Justice 1991; Roenneberg et al. 2007). Considering its influence in entraining the biological temporal system, several studies hypothesize that the latitudinal shift of the sun, related to the earth's axis, would produce, on average, different temporal traits in populations that live close to the equator line when compared to populations that live close to the planet's poles (Horzum et al. 2015; Hut et al. 2013; Leocadio-Miguel et al. 2017, 2014; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler and Rahafar 2017). That is because the latter ones would have greater oscillations in sun activity and an overall weak solar zeitgeber. This is the latitude hypothesis, that can also appear as an environmental hypothesis of circadian rhythm regulation.

Recently there have been attempts to test the latitude hypothesis in different settings, but, at least in humans, none of them have been successful in seeing a significant effect size related to the latitudinal cline. Some of these approaches worked with secondary data and with small samples. One of the most serious attempts of testing this hypothesis was made by Leocadio-Miguel et al. (2017) in 2017. They measured the chronotype of $12,884$ Brazillian subjects on a wide latitudinal spectrum using the Morningness–Eveningness Questionnaire (MEQ). Their results showed a negligible effect size. One possible reason for this is that the MEQ measures psychological traits and not biological states (Roenneberg, Winnebeck, and Klerman 2019), i.e., the circadian oscillation itself, therefore, it's not the best way to answer the question (Leocadio-Miguel et al. 2014).

This article brings a novel attempt to test the latitude hypothesis, using, this time, a biological approach provided by the Munich ChronoType Questionnaire (MCTQ) (Roenneberg, Wirz-Justice, and Merrow 2003). Furthermore, the test was carried out on the biggest chronotype sample ever collected in a same country. A sample made of $73,825$ subjects, all living in the same timezone in Brazil, with only one week of difference between questionnaire responses.

6.3.2 Results

The midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, had an overall mean of 04:26:51. The distribution curve is shown in Figure 1.

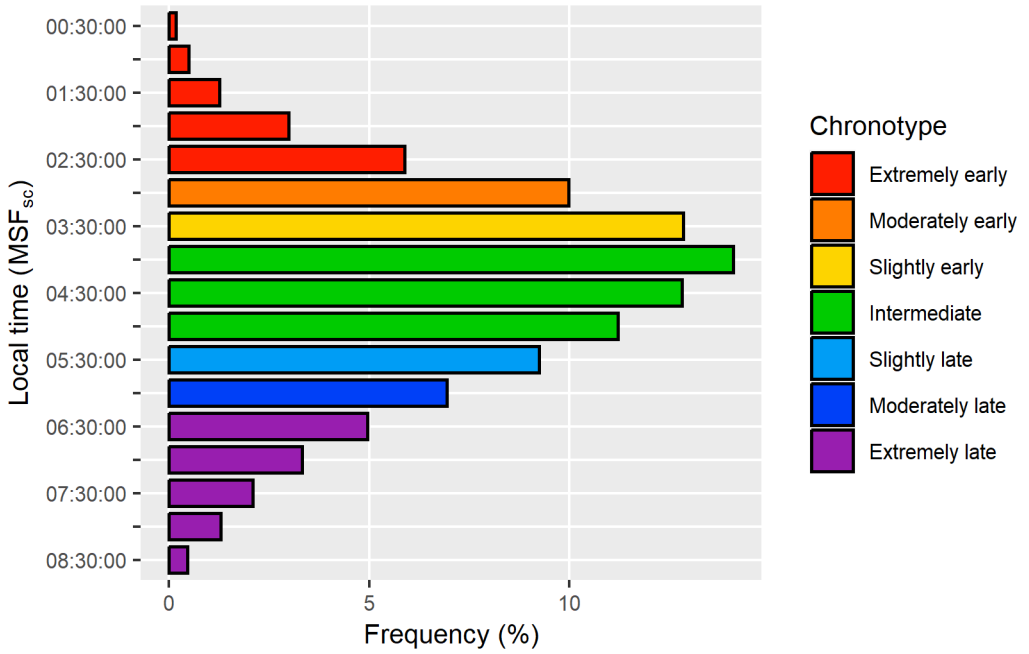
That's the midsleep point of Brazilian subjects with an intermediate/average chronotype. One can imagine, following the 7-9h sleep recommendation for healthy adults of the American Academy of Sleep Medicine (AASM) (Watson et al. 2015), that this average person would, if he/she had no social obligations, typically wake up at about 08:26:51.

```
library(dplyr)
library(here)
library(targets)
library(tidyr)

data <-
  targets::tar_read("geocoded_data", store = here::here("_targets")) |>
  dplyr::select(
    age, sex, state, region, latitude, longitude, height, weight, work, study,
    msf_sc, sjl, le_week,
    ) |>
  tidyr::drop_na(msf_sc, age, sex, latitude)
```

```
library(here)
library(latex2exp)

source(here::here("R/plot_chronotype.R"))

data |>
    plot_chronotype(
    col = "msf_sc",
    x_lab = "Frequency (%)",
```

```
      y_lab = latex2exp::TeX("Local time ($MSF_{sc}$)"),
      col_width = 0.8,
      col_border = 0.6,
      text_size = 10,
      chronotype_cuts = FALSE,
      legend_position = "right"
    )
```

Figure 1 – Distribution of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype. The categorical cuts follow a quantile approach going from extremely early $(0| - 0.11)$ to the extremely late $(0.88 - 1)$.



The MSF_{sc} curve had a skewness of 0.291 and a kurtosis of 2.766. However, the distribution was not normal accordingly to Kolmogorov-Smirnov test ($D = 0.0373$; p-value $< 2e - 16$) and D'Agostino Skewness test ($Z3 = 31.544$; p-value $< 2.2e - 16$).

A linear regression model was created with MSF_{sc} as the response variable and with age and sex as predictors ($R^2 = 0.054433$; $F(2, 73822) = 2130$, p-value $< 2e - 16$). A Box-Cox transformation of the response variable was needed to attend to the linear regression model assumptions ($\lambda = -1.1919$; $MSF_{sc}^{\lambda-1}/\lambda$). All coefficients were significantly different than 0 (p-value $< 2e - 16$) and, accordingly to D'Agostino

Skewness test, the residuals were normal (Z3 $= -1.2704$; p-value < 0.2039). Residual homoscedasticity was verified by a Score Test for Heteroskedasticity ($\chi^2 = 0.00$; p-value $= 1$). No collinearity was found between the predictor variables (variance inflation factor: age $= 1.0014$; sex $= 1.0014$).

Another model was created on top of the first one, adding the latitude as a predictor variable ($R^2 = 0.06204$; $F(3, 73821) = 1630$, p-value $< 2e-16$). All coefficients were significantly different than 0 (p-value $< 2e-16$) and the residuals were normally distributed accordingly to the D'Agostino Skewness test, (Z3 $= 0.0703$; p-value < 0.944). Residual homoscedasticity was verified by a Score Test for Heteroskedasticity ($\chi^2 = 0.00$; p-value $= 1$). No collinearity was found between the predictor variables (variance inflation factor: age $= 1.0067$; sex $= 1.0018$; latitude $= 1.0056$). The longitude was not used as a predictor because it presented colinearity with the latitude variable.

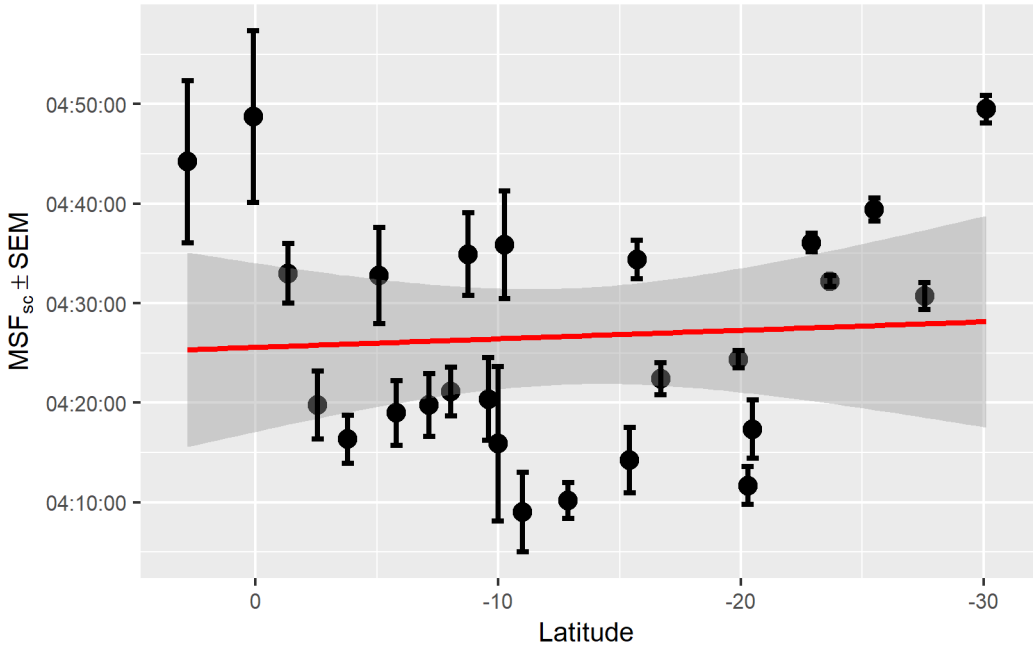
An F test for nested models showed a significant reduction of the residual sum of squares ($F(1, 73821) = 600$, p-value $< 2e-16$), meaning that the latitude seems to produce an effect on the chronotype. However, when estimating Cohen's f^2 effect size, the result was negligible (Cohen 1992) $((0.06204 - 0.054433)/(1 - 0.06204) = 0.0081102)$.

6.3.3 Discussion

The results show that even with a wide latitudinal spectrum and with a big and aligned sample of biological states the latitude effect does not reveal itself in a non-negligible size. Several studies indicate the existence of this effect on the chronotype (Hut et al. 2013; Leocadio-Miguel et al. 2017; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler 2008; Randler and Rahafar 2017; Roenneberg, Wirz-Justice, and Merrow 2003), but, at this time, at least in humans, no empirical evidence can support this claim. Our results are very similar to Leocadio-Miguel et al. (2017), which also found a negligible effect size (Cohen's $f^2 = 0.004143174$). The inconsistency of the latitude effect can be visualized in Figure 2.

```r
library(dplyr)
library(here)
library(latex2exp)

source(here::here("R/plot_latitude_series.R"))

data |>
  dplyr::filter(age <= 50) |>
  plot_latitude_series(
    col = "msf_sc",
    y_lab = latex2exp::TeX("$MSF_{sc} \\pm SEM$"),
    line_width = 2,
    point_size = 3,
    error_bar_width = 0.5,
    error_bar_linewidth = 1,
    error_bar = TRUE,
    text_size = 10
    )
```

Figure 2 – Distribution of mean aggregates of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, with latitude decimal degree intervals. Higher values of MSF_{sc} indicate a late chronotype tendency.



Despite the lack of evidence, is not uncommon to hear talks insisting that this effect is real and already proven. We suspect that this behavior may be derived from a lack of understanding of statistical models and techniques. Although it may be logical and aligned with the overall theory for the evolution of biological temporal systems, it's our role as scientists to eliminate contractions, not pursue them.

As Karl Popper said, science begins and ends with questions (Popper 1979). The absence of a strong entrainment with the solar zeitgeber shows that the entrainment phenomenon is more complex than we previously imagined. Other hypotheses for the human circadian entrainment, like the entrainment to self-selected light, proposed by Anna Skeldon and Derk-Jan Dijk (Skeldon and Dijk 2021), need to be tested and may produce significant results.

It's important to notice that the results shown here are preliminary. The data still needs some cleaning and to be balanced with Brazil's latest population census. The latitude coordinates used in the analysis are related to the subject's state capital and, hence, have low resolution. Even with these results, it may be that a significant latitude effect can still appear at the end of the research.

Despite the several strengths that the dataset used in this study has, it is also important to notice its weaknesses and limitations. The fact that all the subjects were measured in the Spring season is one of them. Since the objective is to catch individuals in different seasonal patterns, the ideal moment to collect this kind of data is in the wintertime, when there is a greater insolation gradient between the equator and the poles. Another one is that this dataset can be influenced by the presence of a Daylight Saving Time (DST) event. This latter issue is explored in more detail in the methods section.

6.4 Methods

6.4.1 Ethics information

Abiding by Brazilian law, all research involving human subjects must have the approval of a Research Ethics Committee (REC) affiliated with the Brazilian National Research Ethics Committee (CONEP). This approval request is ongoing.

6.4.2 Measurement instrument

Chronotypes were measured using the core version of the standard Munich ChronoType Questionnaire (MCTQ) (Roenneberg, Wirz-Justice, and Merrow 2003). MCTQ is a widely validated and widely used self-report questionnaire for measuring the sleep-wake cycle and chronotypes (Roenneberg, Winnebeck, and Klerman 2019). It quantifies the chronotype as a state, a biological circadian phenotype, using as a proxy the local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}). A sleep correction (SC) is made when a possible sleep compensation related to a lack of sleep on workdays is identified (Roenneberg 2012).

Subjects were asked to complete an online questionnaire based on the MCTQ Portuguese translation created by Till Roenneberg & Martha Merrow for the EUCLOCK project (Roenneberg and Merrow 2006) (statements mean cosine distance $= 0.921$). They were also asked to provide sociodemographic (e.g., age, gender), geographic (e.g., full residential address), anthropometric (e.g., weight, height), and work/study

routine-related data. A deactivated version of the questionnaire can be seen at https://bit.ly/brchrono-form.

6.4.3 Sample

The sample is made up of $73,825$ Brazilian subjects. It was obtained in 2017 from October 15th to 21st by a broadcast of the online research questionnaire on a popular Sunday TV show with national reach (Globo 2017). This amount of data collected in such a short time gave the sample a population cross-sectional characteristic.

A survey made in 2019 by the Brazilian Institute of Geography and Statistics (IBGE)27 saw that 82.17% of Brazilians' homes had access to an internet connection. Hence, this sample can have a good diversity of Brazil's population. Only Brazilian residents in states with UTC-3 timezone and with an age equal to or greater than 18 years old were included in the final sample.

In order to verify if the sample size was adequate for the study of the phenomenon under investigation, a power analysis was conducted for nested multiple regression models using the G*Power software (Faul et al. 2007). The analysis used the parameters presented in Leocadio-Miguel et al. (2017) article for a multiple linear regression with 10 tested predictors and only 10 conceived predictors, considering a significance level of 0.05 (α) and a power of 0.95 ($1 - \beta$). The result showed that a sample of $5,895$ individuals would be necessary to test the hypothesis.

Daylight Saving Time (DST) began in Brazil at midnight on November 15th, 2017. Residents from the Midwest, Southeast, and South regions were instructed to set the clock forward by 1 hour. We believe that this event did not contaminate the data since it started on the same day of the data collection. It's important to notice that MCTQ asks subjects to relate their routine behavior, not how they behaved in the last few days. A possible effect of the DST on the sample is the production of an even later chronotype for populations near the planet's poles, amplifying a possible latitude effect. However, this was not shown on the hypothesis test.

Based on the 2010 census29, Brazil had 51.793% of females and 48.207% of males with an age equal to or greater than 18 years old. The sample is skewed for female subjects, with 66.341% of females and 33.659% of male subjects.

The subject's mean age is 32.017 years (SD $= 9.242$; Max. $= 58.786$). Female subjects have a mean age of $31.770 years$ (SD $= 9.340$; Max. $= 58.786$) and male subjects 32.504 years (SD $= 9.026$; Max. $= 58.772$). For comparison, based on the 2010 census29, Brazil's population with an age equal to or greater than 18 years old had a mean age of 41.032 years (SD $= 9.242$), with a mean age of 41.645 years (SD $= 16.907$) for female subjects and a mean age of 40.373 years (SD $= 16.200$) for male subjects. Data related to age and sex from the last Brazil census (2022) were not available before this article was submitted.

Considering the 5 major regions of Brazil, the sample is mostly skewed for the Southeast, the most populated region. According to Brazil's 2022 census, the Southeast region is home to 41.784% of Brazil's population, followed by the Northeast (26.910%), South (14.741%), North (8.544%), and Midwest (8.021%) regions. 62.454% of the sample is located in the Southeast region, 11.797% in the Northeast, 17.861% in the South, 1.682% in the North, and 6.205% in the Midwest region. Note that a lack of subjects in the North and Midwest region is justified by the sample timezone inclusion criteria (UTC-3).

The sample latitudinal range was 30.211 decimal degrees (Min. $= -30.109$; Max. $= 0.10177$) with a longitudinal span of 16.378 decimal degrees (Min. $= -51.342$; Max. $= -34.964$). For comparison, Brazil has a latitudinal range of 39.024 decimal degrees (Min. $= -33.752$; Max. $= 5.2719$) and a longitudinal span of 39.198 decimal degrees (Min. $= -34.793$; Max. $= -73.991$).

The results shown in this article are just a preliminary view of the data analysis. The latitudes and longitudes of each subject are represented by the coordinates of his/her state's capital (a low resolution). The final results will have the latitude and longitude coordinates based on the subject's postal codes and will also use a balanced dataset following the latest Brazil census.

6.4.4 Analysis

The data wrangling and analysis followed the data science program proposed by Hadley Wickham and Garrett Grolemund (Wickham and Grolemund 2016). All processes were made with the help of the R programming language (R Core Team 2023),

RStudio IDE (Posit Team 2023), and several R packages. The tidyverse and rOpenSci package ecosystem and other R packages adherents of the tidy tools manifesto (Wickham and Bryan 2023) were prioritized. The MCTQ data was analyzed using the `mctq` rOpenSci peer-reviewed package (Vartanian 2023d). All processes were made in order to provide result reproducibility and to be in accordance with the FAIR principles (Wilkinson et al. 2016).

The study hypothesis was tested using nested models of multiple linear regressions. The main idea of nested models is to verify the effect of the inclusion of one or more predictors in the model variance explanation (i.e., the R^2) (Allen 1997). This can be made by creating a restricted model and then comparing it with a full model. Hence, the hypothesis can be schematized as follows.

$$\begin{cases} H_0 : R^2_{res} >= R^2_{full} \\ H_a : R^2_{res} < R^2_{full} \end{cases}$$

In order to test a possible latitude association in predicting the chronotype, the full model was the restricted model with the addition of the latitude variable. The restricted model had the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}) as the response variable, MCTQ proxy for the chronotype, with sex and age as predictors.

A residual analysis was made to ensure the validity of the models before the hypothesis test. The hypothesis was tested using a 0.05 (α) significance level.

To favor the alternative hypothesis (H_a), not only the R^2 of the full model must be significantly larger than the R^2 of the restricted model, but the effect size must be at least considered small. To evaluate the effect size, Cohen's f^2 and his categorical parameters for size were used (Cohen 1992). That means that, in order to favor (H_a), the effect size must be at least equal to or greater than 0.0219.

No blinding procedures were used during the analysis.

6.4.5 Data availability

The data that support the findings of this study are available from the corresponding author [DV]. Restrictions apply to the availability of these data, which were used

under the approval of a Research Ethics Committee (REC) linked to the Brazilian National Research Ethics Committee (CONEP), hence it cannot be publicly shared. Data are, however, available from the author upon reasonable request and with CONEP approval.

6.4.6 Code availability

An R package representing the research compendium of the project, with all the code used accompanied by extensive documentation, will be available under the MIT license at https://github.com/danielvartan/mastersthesis when the research is completed. The code has all the steps from the raw data to the hypothesis test results.

6.5 Acknowledgments

Financial support was provided by the Coordination for the Improvement of Higher Education Personnel (CAPES) and by the University of Sao Paulo (USP) (Grant number: 88887.703720/2022-00).

6.6 Ethics declarations

6.6.1 Competing interests

The author declares that the study was carried out without any commercial or financial connections that could be seen as a possible competing interest.

6.7 Additional information

This manuscript shows only preliminary results and should not be considered a document ready for journal submission.

There is no supplementary information.

Correspondence can be sent to Daniel Vartanian (danvartan@gmail.com).

6.8 Rights and permissions

The author does not give permission to share this manuscript while the research is ongoing. An exception is made for education purposes. In the latter case, the manuscript must be accompanied by a warning statement, located at the top of the first page, warning the reader about the need for confidentiality. After the research process ends, something that just a written communication by the author can provide, this article will be licensed under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as be given appropriate credit to the original author and the source, provide a link to the Creative Commons license, and indicate if changes were made.

7 Discussion and conclusions

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

7.1 Discussion

7.2 Limitations

7.3 Conclusions

7.4 Future perspectives

REFERENCES [1]

Allen, Michael Patrick. 1997. *Understanding Regression Analysis*. New York: Plenum Press.

Aschoff, Jürgen, ed. 1981. *Biological Rhythms*. Vol. 4. Handbook of Behavioral Neurobiology. New York, NY: Plenum Press. https://doi.org/10.1007/978-1-4615-6552-9.

———. 1989. "Temporal Orientation: Circadian Clocks in Animals and Humans." *Animal Behaviour* 37 (June): 881–96. https://doi.org/10.1016/0003-3472(89)90132-2.

Aschoff, Jürgen, K. Klotter, and R. Wever. 1965. "Circadian Vocabulary: A Recommended Terminology with Definitions." In *Circadian Clocks*. North-Holland.

Bertalanffy, Ludwig von. 1968. *General System Theory: Foundations, Development, Applications*. New York, NY: George Braziller.

Bussab, Wilton de Oliveira. 1988. *Análise de variância e de regressão: uma introdução*. 2nd ed. Métodos quantitativos. São Paulo: Atlas.

Cambridge University Press. n.d. "Cambridge Dictionary." Accessed September 21, 2023. https://dictionary.cambridge.org/.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155–59. https://doi.org/10.1037/0033-2909.112.1.155.

Dalpiaz, David. n.d. *Applied Statistics with R*. https://book.stat420.org/.

DeGroot, Morris H., and Mark J. Schervish. 2012. *Probability and Statistics*. 4th ed. Boston: Addison-Wesley.

Dudek, Bruce. 2020. *Linear Models with R: Emphasis on 2-IV Models: Basics of Multiple Regression*. https://bcdudek.net/regression1/.

Ehret, Charles F. 1974. "The Sense of Time: Evidence for Its Molecular Basis in the Eukaryotic Gene-Action System." In *Advances in Biological and Medical Physics*, 15:47–77. Elsevier. https://doi.org/10.1016/B978-0-12-005215-8.50009-7.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91. https://doi.org/10.3758/BF03193146.

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Thousand Oaks, CA: Sage.

[1] According to the APA style - American Psychological Association.

Frey, Bruce B., ed. 2022. *The SAGE Encyclopedia of Research Design*. 2nd ed. Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781071812082.

Globo. 2017. “Metade Da População Se Sente Mal No Horário de Verão, Revela Pesquisa.” Fantástico. October 15, 2017. https://globoplay.globo.com/v/6219513/.

Hair, Joseph F. 2019. *Multivariate Data Analysis*. 8th ed. Andover, Hampshire: Cengage.

Horne, J. A., and O. Ostberg. 1976. “A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms.” *International Journal of Chronobiology* 4 (2): 97–110.

Horzum, Mehmet Barış, Christoph Randler, Ercan Masal, Şenol Beşoluk, İsmail Önder, and Christian Vollmer. 2015. “Morningness–Eveningness and the Environment Hypothesis – a Cross-Cultural Comparison of Turkish and German Adolescents.” *Chronobiology International* 32 (6): 814–21. https://doi.org/10.3109/07420528.2015.1041598.

Hut, Roelof A., Silvia Paolucci, Roi Dor, Charalambos P. Kyriacou, and Serge Daan. 2013. “Latitudinal Clines: An Evolutionary View on Biological Rhythms.” *Proceedings of the Royal Society B: Biological Sciences* 280 (1765): 20130433. https://doi.org/10.1098/rspb.2013.0433.

Johnson, Richard, and Dean Wichern. 2013. *Applied Multivariate Statistical Analysis: Pearson New International Edition*. 6th ed. Harlow, UK: Pearson.

Khalsa, Sat Bir S., Megan E. Jewett, Christian Cajochen, and Charles A. Czeisler. 2003. “A Phase Response Curve to Single Bright Light Pulses in Human Subjects.” *The Journal of Physiology* 549 (3): 945–52. https://doi.org/10.1113/jphysiol.2003.040477.

Kuhlman, Sandra J., L. Michon Craig, and Jeanne F. Duffy. 2018. “Introduction to Chronobiology.” *Cold Spring Harbor Perspectives in Biology* 10 (9): a033613. https://doi.org/10.1101/cshperspect.a033613.

Kuhn, Max, and Julia Silge. 2022. *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. Sebastopol, CA: O’Reilly Media. https://www.tmwr.org/.

Latinitium. n.d. “Latin Dictionaries.” Latinitium. Accessed September 21, 2023. https://latinitium.com/latin-dictionaries/.

Leocadio-Miguel, Mario André, Fernando Mazzili Louzada, Leandro Lourenção Duarte, Roberta Peixoto Areas, Marilene Alam, Marcelo Ventura Freire, John Fontenele-

Araujo, Luiz Menna-Barreto, and Mario Pedrazzoli. 2017. “Latitudinal Cline of Chronotype.” *Scientific Reports* 7 (1): 5437. https://doi.org/10.1038/s41598-017-05797-w.

Leocadio-Miguel, Mario André, Valéria Clarisse De Oliveira, Danyella Pereira, and Mario Pedrazzoli. 2014. “Detecting Chronotype Differences Associated to Latitude: A Comparison Between Horne--Östberg and Munich Chronotype Questionnaires.” *Annals of Human Biology* 41 (2): 107–10. https://doi.org/10.3109/03014460.2013.832795.

Levins, Richard. 1998. “Dialectics and Systems Theory.” *Science & Society* 62 (3): 375–99. https://www.jstor.org/stable/40403729.

Marques, Mirian David, and Gisele Oda. 2012. “Glossário.” *Revista da Biologia* 9 (3). https://www.revistas.usp.br/revbiologia/article/view/114816.

Minors, David S., James M. Waterhouse, and Anna Wirz-Justice. 1991. “A Human Phase-Response Curve to Light.” *Neuroscience Letters* 133 (1): 36–40. https://doi.org/10.1016/0304-3940(91)90051-T.

Mitchell, Melanie. 2013. “Introduction to Complexity.” Online course. 2013. https://www.complexityexplorer.org/courses/1-https://www.complexityexplorer.org/courses/1.

Paranjpe, Dhanashree A, and Vijay Kumar Sharma. 2005. “Evolution of Temporal Order in Living Organisms.” *Journal of Circadian Rhythms* 3 (May). https://doi.org/10.1186/1740-3391-3-7.

Pittendrigh, C. S. 1960. “Circadian Rhythms and the Circadian Organization of Living Systems.” *Cold Spring Harbor Symposia on Quantitative Biology* 25: 159–84. https://doi.org/10.1101/SQB.1960.025.01.015.

———. 1993. “Temporal Organization: Reflections of a Darwinian Clock-Watcher.” *Annual Review of Physiology* 55 (1): 17–54. https://doi.org/10.1146/annurev.ph.55.030193.000313.

Pittendrigh, Colin S., Walter T. Kyner, and Tsuguhiko Takamura. 1991. “The Amplitude of Circadian Oscillations: Temperature Dependence, Latitudinal Clines, and the Photoperiodic Time Measurement.” *Journal of Biological Rhythms* 6 (4): 299–313. https://doi.org/10.1177/074873049100600402.

Popper, Karl R. 1979. *Objective Knowledge: An Evolutionary Approach*. Rev. ed. Oxford University Press.

Posit Team. 2023. “RStudio: Integrated Development Environment for R.” C. Posit Software. http://www.posit.co.

R Core Team. 2023. “R: A Language and Environment for Statistical Computing.” R. R Foundation for Statistical Computing. https://www.R-project.org.

Randler, Christoph. 2008. “Morningness□eveningness Comparison in Adolescents from Different Countries Around the World.” *Chronobiology International* 25 (6): 1017–28. https://doi.org/10.1080/07420520802551519.

Randler, Christoph, and Arash Rahafar. 2017. “Latitude Affects Morningness-Eveningness: Evidence for the Environment Hypothesis Based on a Systematic Review.” *Scientific Reports* 7 (1): 39976. https://doi.org/10.1038/srep39976.

Reis, Cátia. 2020. “Sleep patterns in portugal.” Tese de doutorado, Lisboa: Universidade de Lisboa. Repositório da Universidade de Lisboa. http://hdl.handle.net/10451/54147.

Roenneberg, Till. 2012. “What Is Chronotype?” *Sleep and Biological Rhythms* 10 (2): 75–76. https://doi.org/10.1111/j.1479-8425.2012.00541.x.

Roenneberg, Till, Karla V. Allebrandt, Martha Merrow, and Céline Vetter. 2012. “Social Jetlag and Obesity.” *Current Biology* 22 (10): 939–43. https://doi.org/10.1016/j.cub.2012.03.038.

Roenneberg, Till, Serge Daan, and Martha Merrow. 2003. “The Art of Entrainment.” *Journal of Biological Rhythms* 18 (3): 183–94. https://doi.org/10.1177/0748730403018003001.

Roenneberg, Till, Roelof Hut, Serge Daan, and Martha Merrow. 2010. “Entrainment Concepts Revisited.” *Journal of Biological Rhythms* 25 (5): 329–39. https://doi.org/10.1177/0748730410379082.

Roenneberg, Till, Tim Kuehnle, Myriam Juda, Thomas Kantermann, Karla Allebrandt, Marijke Gordijn, and Martha Merrow. 2007. “Epidemiology of the Human Circadian Clock.” *Sleep Medicine Reviews* 11 (6): 429–38. https://doi.org/10.1016/j.smrv.2007.07.005.

Roenneberg, Till, and Martha Merrow. 2006. “EUCLOCK: Portuguese MCTQ.” 2006. http://web.archive.org/web/20141115175303/https://www.bioinfo.mpg.de/mctq/core_work_life/core/core.jsp?language=por_b.

Roenneberg, Till, Eva C. Winnebeck, and Elizabeth B. Klerman. 2019. “Daylight Saving Time and Artificial Time Zones – a Battle Between Biological and Social Times.” *Frontiers in Physiology* 10 (August): 944. https://doi.org/10.3389/fphys.2019.00944.

Roenneberg, Till, Anna Wirz-Justice, and Martha Merrow. 2003. “Life Between Clocks: Daily Temporal Patterns of Human Chronotypes.” *Journal of Biological Rhythms* 18 (1): 80–90. https://doi.org/10.1177/0748730402239679.

Skeldon, Anne C., and Derk□Jan Dijk. 2021. “Weekly and Seasonal Variation in the Circadian Melatonin Rhythm in Humans: Entrained to Local Clock Time, Social Time, Light Exposure or Sun Time?” *Journal of Pineal Research* 71 (1): e12746. https://doi.org/10.1111/jpi.12746.

The Royal Society. n.d. “History of the Royal Society.” Accessed September 9, 2023. https://royalsociety.org/about-us/history/.

Vartanian, Daniel. 2022a. “{Entrainment}: A Rule-Based Model of the 24h Light/Dark Cycle Entrainment Phenomenon.” Python. https://github.com/danielvartan/entrainment.

———. 2022b. “Ecology of Sleep and Circadian Phenotypes of the Brazilian Population.” São Paulo, November 19. https://doi.org/10.13140/RG.2.2.25343.07840.

———. 2023a. “Ecology of Sleep and Circadian Phenotypes of the Brazilian Population [Research Compendium].” Research compedium. R/TeX/Quarto. https://github.com/danielvartan/tesesusp.

———. 2023b. “{Lockr}: Easily Encrypt/Decrypt Files.” R. https://github.com/danielvartan/lockr.

———. 2023c. “{Lubritime}: An Extension for the Lubridate Package.” R. https://github.com/danielvartan/lubritime.

———. 2023d. “{Mctq}: Tools to Process the Munich ChronoType Questionnaire (MCTQ).” R. https://docs.ropensci.org/mctq/.

———. 2023e. “{Tesesusp}: A Quarto Format for USP Theses and Dissertation.” TeX. https://github.com/danielvartan/tesesusp.

———. 2023f. “Plano de ensino: ACH0042 - Resolução de Problemas II.” São Paulo. https://doi.org/10.13140/RG.2.2.33335.50086.

———. 2023g. “Ecologia do sono e de fenótipos circadianos da população brasileira [Data Management Plan].” DMPHub. https://doi.org/10.48321/D1DW8P.

Vartanian, Daniel, and Mario Pedrazzoli. 2017. “Questionário de Cronotipo: Baseado No Munich ChronoType Questionnaire (MCTQ).” 2017. https://web.archive.org/web/20171018043514/each.usp.br/gipso/mctq.

Viana-Mendes, Juliana, Ana Amélia Benedito-Silva, Maria Augusta Medeiros Andrade, Daniel Vartanian, Bruno da Silva Brandão Gonçalves, José Cipolla-Neto, and Mario Pedrazzoli. 2023. “Actigraphic Characterization of Sleep and Circadian Phenotypes of PER3 Gene VNTR Genotypes.” *Chronobiology International*, September. https://doi.org/10.1080/07420528.2023.2256858.

Wang, Jiapeng, and Yihong Dong. 2020. “Measurement of Text Similarity: A Survey.” *Information* 11 (9): 421. https://doi.org/10.3390/info11090421.

Watson, Nathaniel F., M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, et al. 2015. “Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society.” *Journal of Clinical Sleep Medicine* 11 (6): 591–92. https://doi.org/10.5664/jcsm.4758.

Wickham, Hadley, and Jennifer Bryan. 2023. *R Packages*. 2nd ed. O’Reilly. https://r-pkgs.org/.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science*. O’Reilly. https://r4ds.had.co.nz.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

A Appendice: Chapter 2 supplemental material

> **Note**
>
> You are reading the work-in-progress of this thesis. This chapter should be readable but is currently undergoing final polishing.

A.1 Load and embed texts

See Vartanian and Pedrazzoli (2017) to visualize the data questionnaire.

See Roenneberg and Merrow (2006) to visualize the EUCLOCK portuguese questionnaire.

See Reis (2020) to learn more about the MCTQ^{PT} questionnaire.

```
data_text <- c(
  "Você vai para a cama às ___ horas.",
  "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
  "Depois de ir para a cama, você decide dormir às ___ horas.",
  "Você precisa de ___ para dormir.",
  "Você acorda às ___ horas.",
  "Você se levanta ___ depois de despertar.",
  "Você vai para a cama às ___ horas.",
  "",
  "Depois de ir para a cama, você decide dormir às ___ horas.",
  "Você precisa de ___ para dormir.",
  "Você acorda às ___ horas.",
  "Você se levanta ___ depois de despertar."
)

euclock_text <- c(
  "vou para a cama às ___ horas.",
  "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
  "às ___ horas, decido dormir.",
```

```
  "Eu necessito ___ minutos para adormecer.",
  "acordo às ___ horas,",
  "passados ___ minutos, me levanto.",
  "vou para a cama às ___ horas.",
  "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
  "às ___ horas, decido dormir.",
  "Eu necessito ___ minutos para adormecer.",
  "acordo às ___ horas,",
  "passados ___ minutos, me acordo."
)

mctq_pt_text <- c(
  "Vou para a cama às ___ horas.",
  "Algumas pessoas permanecem algum tempo acordadas depois de estarem na cama.",
  "Às ___ horas estou pronto para adormecer.",
  "Necessito de ___ minutos para adormecer.",
  "Acordo às ___ horas.",
  "Após ___ minutos, levanto-me.",
  "Vou para a cama às ___ horas.",
  "Algumas pessoas permanecem algum tempo acordadas depois de estarem na cama.",
  "Às ___ horas estou pronto para adormecer.",
  "Necessito de ___ minutos para adormecer.",
  "Acordo às ___ horas.",
  "Após ___ minutos, levanto-me."
)
```

```
# library(textreuse)

data_text_textreuse <-
  textreuse::TextReuseTextDocument(
    text = data_text,
    meta = list(id = "data")
```

```
  )

euclock_text_textreuse <-
  textreuse::TextReuseTextDocument(
    text = euclock_text,
    meta = list(id = "euclock")
  )

mctq_pt_text_textreuse <-
  textreuse::TextReuseTextDocument(
    text = mctq_pt_text,
    meta = list(id = "mctq_pt")
  )
```

```
# See
# <https://huggingface.co/neuralmind/bert-base-portuguese-cased>
# to learn more.

# library(checkmate)
# library(text)
# library(rutils)

rutils:::assert_internet()

text_embed <- function(text) {
  checkmate::assert_character(text)

  text |>
    text::textEmbed(
      model = "neuralmind/bert-base-portuguese-cased",
      layers = - 2,
      dim_name = TRUE,
```

```
        aggregation_from_layers_to_tokens = "concatenate",
        aggregation_from_tokens_to_texts = "mean",
        aggregation_from_tokens_to_word_types = NULL,
        keep_token_embeddings = TRUE,
        tokens_select = NULL,
        tokens_deselect = NULL,
        decontextualize = FALSE,
        model_max_length = NULL,
        max_token_to_sentence = 4,
        tokenizer_parallelism = FALSE,
        device = "gpu",
        logging_level = "error"
      )
}

data_text_textembed <- text_embed(data_text)
euclock_text_textembed <- text_embed(euclock_text)
mctq_pt_text_textembed <- text_embed(mctq_pt_text)
```

A.2 Text similarity

See Wang and Dong (2020) to learn more.

For a quick explanation, see https://youtu.be/e9U0QAFbfLI.

```
# See `?text::textSimilarity` to learn more.

# library(checkmate)
# library(cli)
# library(text)

text_distance <- function(x, y) {
  checkmate::assert_list(x, len = 2)
```

```r
  checkmate::assert_list(y, len = 2)

  methods <- c(
    "binary", "cosine", "canberra", "euclidean", "manhattan", "maximum",
    "minkowski", "pearson"
  )

  for (i in methods) {
    cli::cli_alert_info(paste0(
      "Method: {.strong {stringr::str_to_title(i)}}"
      ))

    test <-
      text::textSimilarity(
        x$texts$texts,
        y$texts$texts,
        method = i,
        center = TRUE,
        scale = FALSE
      )

    cli::cli_bullets(c(">" = "Line by line"))
    print(test)

    cli::cli_bullets(c(">" = "Overall mean"))
    print(mean(test))

    cli::cat_line()
  }
}
```

```r
# See `?textreuse::jaccard_similarity` to learn more.

# library(checkmate)
# library(cli)
# library(textreuse)

text_representation <- function(x, y) {
  checkmate::assert_class(x, "TextReuseTextDocument")
  checkmate::assert_class(y, "TextReuseTextDocument")

  cli::cli_alert_info(paste0("Method: {.strong Jaccard similarity}"))
  print(textreuse::jaccard_similarity(x, y))
  cli::cat_line()

  cli::cli_alert_info(paste0("Method: {.strong Jaccard bag similarity}"))
  print(textreuse::jaccard_bag_similarity(x, y))
  cli::cat_line()
}
```

A.2.1 How similar is the *data questionnaire* when compared to the *EUCLOCK questionnaire*?

Text distance | Embedded (Semantic) test

```r
# See `?text::textSimilarity` to learn more.

text_distance(data_text_textembed, euclock_text_textembed)
#> i Method: Binary
#> > Line by line
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1
#> > Overall mean
#> [1] 1
#> i Method: Cosine
```

```
#> > Line by line
#>  [1] 0.9911730 1.0000000 0.9639984 0.9662432 0.9604119 0.9557896 0.9911730
#>  [8] 0.1559853 0.9639984 0.9662432 0.9604119 0.9497428
#> > Overall mean
#> [1] 0.9020976
#> i Method: Canberra
#> > Line by line
#>  [1] -218.3367    1.0000 -335.1895 -318.9419 -335.9000 -373.7989 -218.3367
#>  [8] -642.5522 -335.1895 -318.9419 -335.9000 -381.4067
#> > Overall mean
#> [1] -317.7912
#> i Method: Euclidean
#> > Line by line
#>  [1]  -1.504474   1.000000  -4.058168  -3.976768  -4.144779  -4.705653
#>  [7]  -1.504474 -19.453535  -4.058168  -3.976768  -4.144779  -5.071087
#> > Overall mean
#> [1] -4.633221
#> i Method: Manhattan
#> > Line by line
#>  [1]  -53.58509    1.00000 -108.96057 -105.53439 -111.46668 -124.16400
#>  [7]  -53.58509 -198.94244 -108.96057 -105.53439 -111.46668 -131.16054
#> > Overall mean
#> [1] -101.03
#> i Method: Maximum
#> > Line by line
#>  [1]   0.52944993   1.00000000   0.37664773   0.07405534   0.11978415
#>  [6]   0.31058226   0.52944993 -14.91353795   0.37664773   0.07405534
#> [11]   0.11978415   0.16830817
#> > Overall mean
#> [1] -0.9362311
#> i Method: Minkowski
#> > Line by line
```

```
#>  [1]  -1.504474   1.000000  -4.058168  -3.976768  -4.144779  -4.705653
#>  [7]  -1.504474 -19.453535  -4.058168  -3.976768  -4.144779  -5.071087
#> > Overall mean
#> [1] -4.633221
#> i Method: Pearson
#> > Line by line
#>  [1] 0.9911730 1.0000000 0.9639984 0.9662432 0.9604119 0.9557896 0.9911730
#>  [8] 0.1559853 0.9639984 0.9662432 0.9604119 0.9497428
#> > Overall mean
#> [1] 0.9020976
```

Text representation

```
# The maximum value for the Jaccard bag similarity is 0.5.

text_representation(euclock_text_textreuse, data_text_textreuse)
#> i Method: Jaccard similarity
#> [1] 0.2173913
#> i Method: Jaccard bag similarity
#> [1] 0.1446541
```

A.2.2 How similar is the *data questionnaire* when compared to the *MCTQ^{PT} questionnaire*?

Text distance | Embedded (Semantic) test

```
# See `?text::textSimilarity` to learn more.

text_distance(data_text_textembed, mctq_pt_text_textembed)
#> i Method: Binary
#> > Line by line
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1
```

```
#> > Overall mean
#> [1] 1
#> i Method: Cosine
#> > Line by line
#>  [1] 0.9901437 0.9898982 0.9687513 0.9575260 0.9882873 0.9598702 0.9901437
#>  [8] 0.1601500 0.9687513 0.9575260 0.9882873 0.9598702
#> > Overall mean
#> [1] 0.9066005
#> i Method: Canberra
#> > Line by line
#>  [1] -227.9938 -247.6044 -335.9297 -349.9493 -225.4809 -353.5263 -227.9938
#>  [8] -631.6228 -335.9297 -349.9493 -225.4809 -353.5263
#> > Overall mean
#> [1] -322.0823
#> i Method: Euclidean
#> > Line by line
#>  [1]  -1.662187  -1.729814  -3.807963  -4.603019  -1.810249  -4.458696
#>  [7]  -1.662187 -19.380367  -3.807963  -4.603019  -1.810249  -4.458696
#> > Overall mean
#> [1] -4.482867
#> i Method: Manhattan
#> > Line by line
#>  [1]  -57.28537  -58.97985 -102.26048 -119.55311  -59.76706 -117.55850
#>  [7]  -57.28537 -193.79964 -102.26048 -119.55311  -59.76706 -117.55850
#> > Overall mean
#> [1] -97.13571
#> i Method: Maximum
#> > Line by line
#>  [1]   0.60554241   0.60396075   0.41795957   0.01856209   0.39220500
#>  [6]   0.28932291   0.60554241 -14.95654231   0.41795957   0.01856209
#> [11]   0.39220500   0.28932291
#> > Overall mean
```

```
#> [1] -0.9087831
#> i Method: Minkowski
#> > Line by line
#>  [1]  -1.662187  -1.729814  -3.807963  -4.603019  -1.810249  -4.458696
#>  [7]  -1.662187 -19.380367  -3.807963  -4.603019  -1.810249  -4.458696
#> > Overall mean
#> [1] -4.482867
#> i Method: Pearson
#> > Line by line
#>  [1] 0.9901437 0.9898982 0.9687513 0.9575260 0.9882873 0.9598702 0.9901437
#>  [8] 0.1601500 0.9687513 0.9575260 0.9882873 0.9598702
#> > Overall mean
#> [1] 0.9066005
```

Text representation

```
# The maximum value for the Jaccard bag similarity is 0.5.

text_representation(mctq_pt_text_textreuse, data_text_textreuse)
#> i Method: Jaccard similarity
#> [1] 0.1052632
#> i Method: Jaccard bag similarity
#> [1] 0.09815951
```

B Appendice: Chapter 3 supplemental material

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

C Appendice: Chapter 4 supplemental material

> **Note**
>
> You are reading the work-in-progress of this thesis. This chapter should be readable but is currently undergoing final polishing.

C.1 Data wrangling

The data wrangling processes were performed using the `targets` R package. A graphical pipeline of the process can be visualized below.

The analyses are 100% reproducible. You can find all the code used in this thesis along with the computational notebooks at https://github.com/danielvartan/mastersthesis/.

```
# library(dplyr)
# library(here)
library(targets)
# library(tidyr)

data <-
  targets::tar_read("geocoded_data", store = here::here("_targets")) |>
  dplyr::select(
    age, sex, state, region, latitude, longitude, height, weight, work, study,
    msf_sc, sjl, le_week,
    ) |>
  tidyr::drop_na(msf_sc, age, sex, latitude)
```

C.2 Distribution of main variables

```r
# library(here)
# library(hms)
# library(magrittr)

source(here::here("R/test_normality.R"))
source(here::here("R/utils.R"))

col <- "age"

stats <- data |>
  magrittr::extract2(col) |>
  test_normality(
    name = col,
    threshold = hms::parse_hms("12:00:00"),
    remove_outliers = FALSE,
    iqr_mult = 1.5,
    log_transform = FALSE,
    density_line = TRUE,
    text_size = text_size,
    print = TRUE
  )
#> # A tibble: 14 x 2
#>   name    value
#>   <chr>   <chr>
#> 1 n       79198
#> 2 n_rm_na 79198
#> 3 n_na    0
#> 4 mean    31.9838074965417
#> 5 var     85.2414919292643
#> 6 sd      9.23263190695179
```

```
#> # i 8 more rows

stats$stats |> list_as_tibble()
```

Figure 3 – **?(caption)**

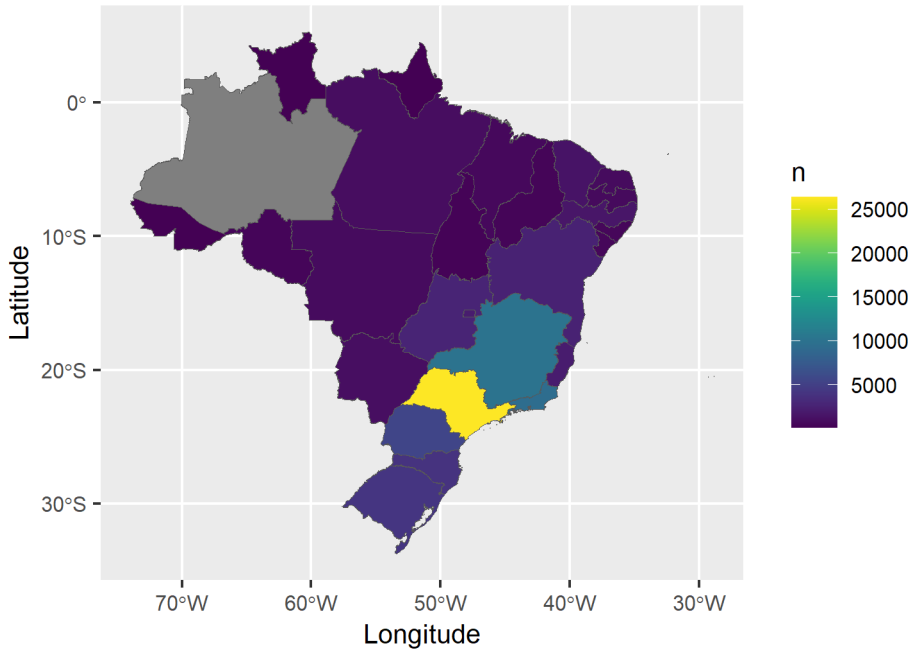
name	value
n	79198
n_rm_na	79198
n_na	0
mean	31.9838074965417
var	85.2414919292643
sd	9.23263190695179
min	18
q_1	24.7222222222222
median	30.5388888888889
q_3	37.61875
max	58.7861111111111
iqr	12.8965277777778
skewness	0.665751526654394
kurtosis	2.82381488030798

Figure 4 – **?(caption)**

C.3 Geographic distribution

```r
# library(here)
# library(rutils)

source(here::here("R/plot_brazil_uf_map.R"))

rutils:::assert_internet()

brazil_uf_map <-
  data |>
  plot_brazil_uf_map(option = "viridis", text_size = 10)
```

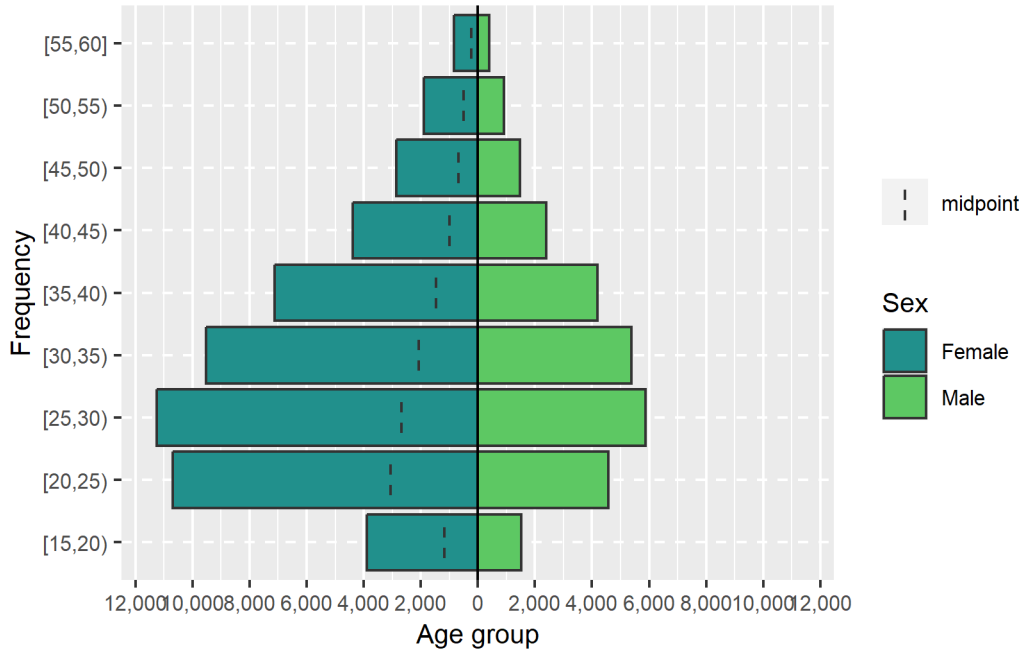
Figure 5 – **?(caption)**



C.4 Age pyramid

```r
# library(here)

source(here::here("R/plot_age_pyramid.R"))

age_pyramid <-
  data |>
  plot_age_pyramid(
    interval = 10,
    na_rm = TRUE,
    text_size = text_size
  )
```

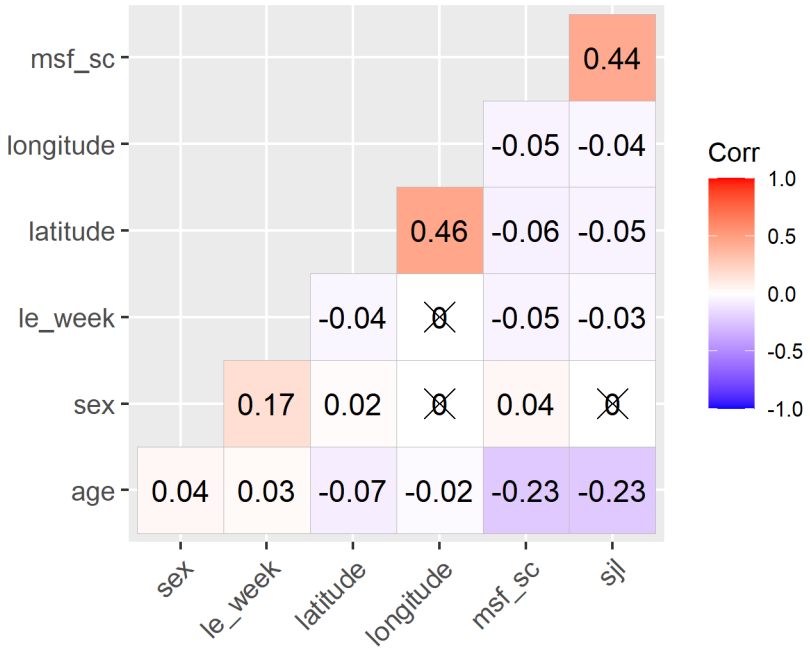
Figure 6 – **?(caption)**



C.5 Correlation matrix

```r
# library(here)

source(here::here("R/plot_ggcorrplot.R"))

cols <- c("sex", "age", "latitude", "longitude", "msf_sc", "sjl", "le_week")

ggcorrplot <-
  data |>
  plot_ggcorrplot(
    cols = cols,
    na_rm = TRUE,
    text_size = text_size,
    hc_order = TRUE
    )
```

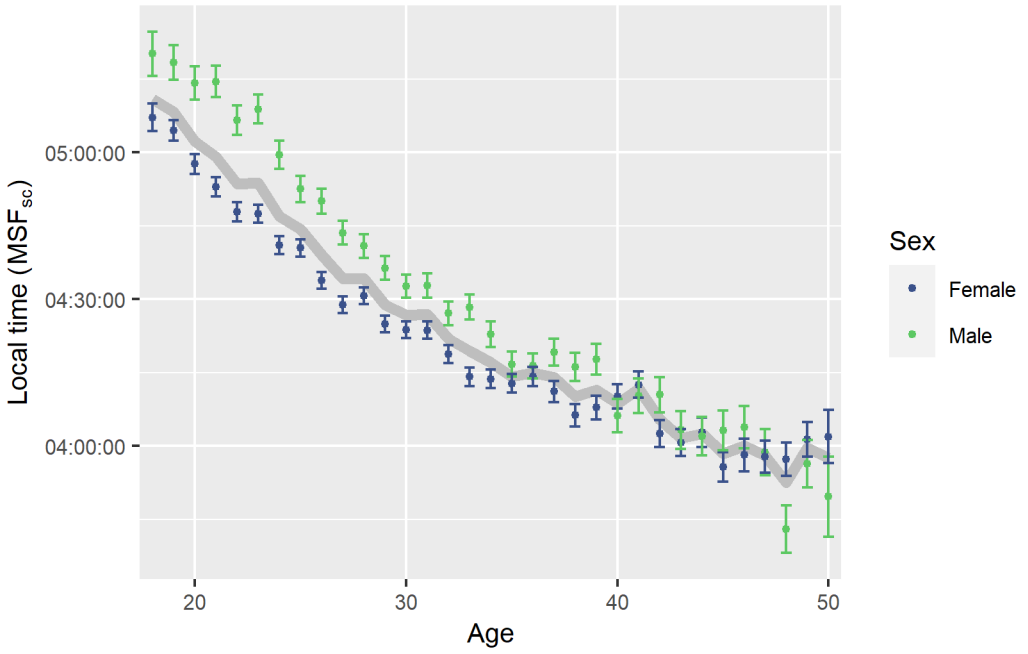
Figure 7 – **?(caption)**



C.6 Age series

```r
# library(here)
# library(latex2exp)

source(here::here("R/plot_age_series.R"))

col <- "msf_sc"
y_lab <- latex2exp::TeX("Local time ($MSF_{sc}$)")

data |>
  dplyr::filter(age <= 50) |>
  plot_age_series(
    col = col,
    y_lab = y_lab,
    line_width = 2,
    boundary = 0.5,
    point_size = 1,
    error_bar_width = 0.5,
    error_bar_linewidth = 0.5,
    error_bar = TRUE,
    text_size = text_size
    )
```

Figure 8 – **?(caption)**

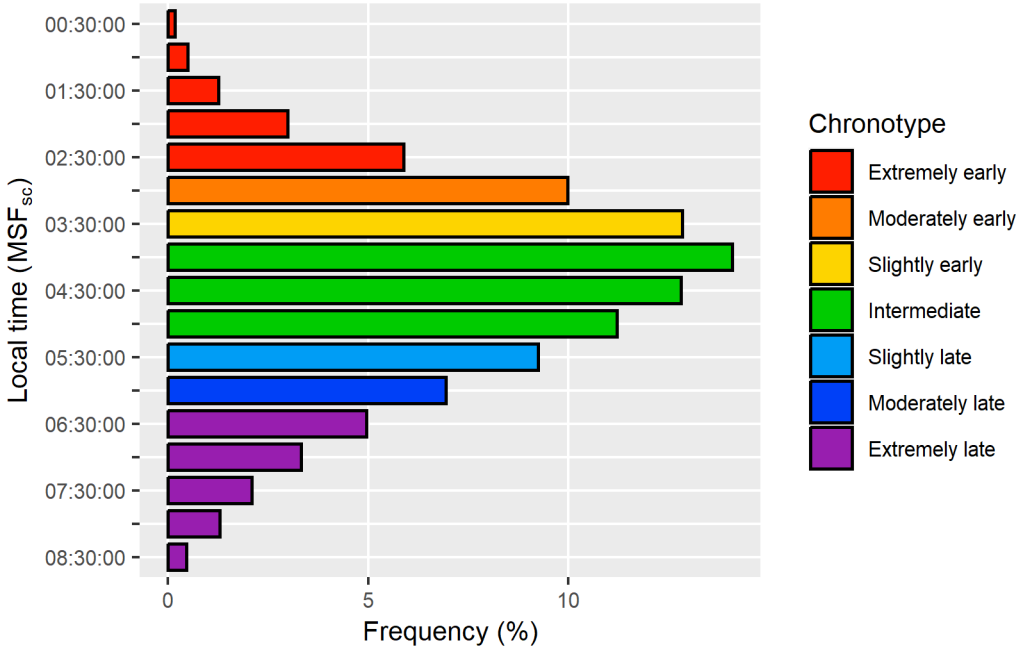


C.7 Chronotype

```
# library(here)
# library(latex2exp)

source(here::here("R/plot_chronotype.R"))

col <- "msf_sc"
y_lab <- latex2exp::TeX("Local time ($MSF_{sc}$)")

data |>
  plot_chronotype(
    col = col,
    x_lab = "Frequency (%)",
    y_lab = y_lab,
    col_width = 0.8,
    col_border = 0.6,
    text_size = text_size,
```

```
    legend_position = "right",
    chronotype_cuts = FALSE
  )
```

Figure 9 – **?(caption)**

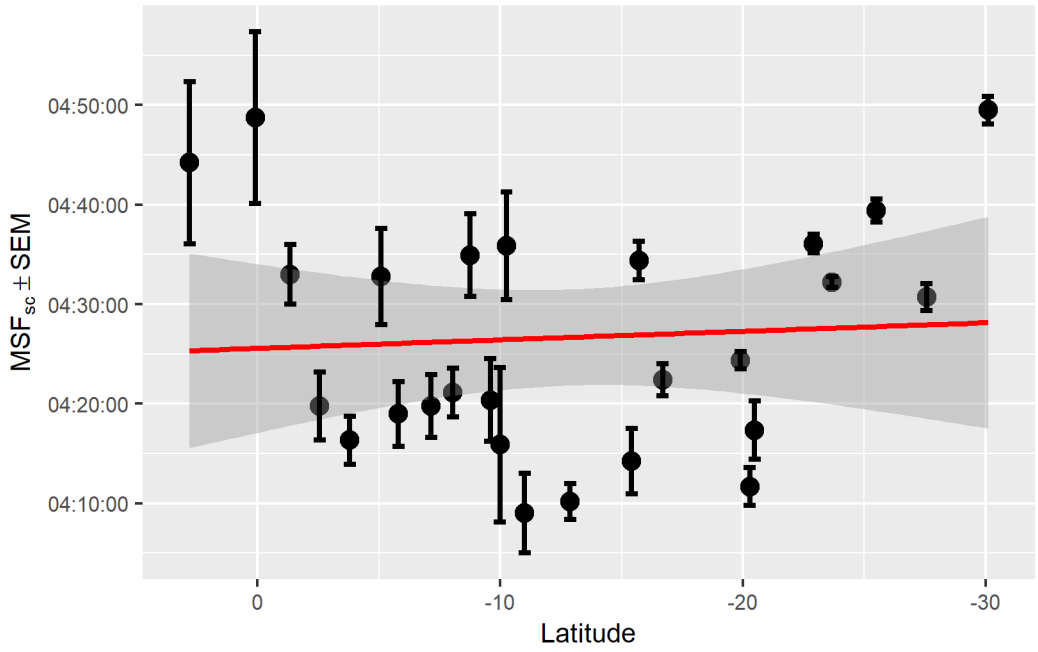


C.8 Latitude series

```
# library(here)
# library(latex2exp)

source(here::here("R/plot_latitude_series.R"))

col <- "msf_sc"
y_lab <- latex2exp::TeX("$MSF_{sc} \\pm SEM$")

data |>
  dplyr::filter(age <= 50) |>
  plot_latitude_series(
    col = col,
```

```
    y_lab = y_lab,
    line_width = 2,
    point_size = 3,
    error_bar_width = 0.5,
    error_bar_linewidth = 1,
    error_bar = TRUE,
    text_size = text_size
  )
```

Figure 10 – **?(caption)**



C.9 Statistics

C.9.1 Numerical variables

```
# library(here)
# library(magrittr)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

```

```r
col <- "age"

data |>
  magrittr::extract2(col) |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

name	value
n	79198
n_rm_na	79198
n_na	0
mean	31.9838074965417
var	85.2414919292643
sd	9.23263190695179
min	18
q_1	24.7222222222222
median	30.5388888888889
q_3	37.61875
max	58.7861111111111
iqr	12.8965277777778
skewness	0.665751526654394
kurtosis	2.82381488030798

C.9.2 Sex

```r
# See <https://sidra.ibge.gov.br> to learn more.

# library(dplyr)
library(magrittr)
# library(sidrar)
# library(stringr)
# library(rutils)

rutils:::assert_internet()

# Brazil's 2010 census data
```

```r
census_data <-
  sidrar::get_sidra(x = 1378) %>% # Don't change the pipe
  dplyr::filter(
    Sexo %in% c("Homens", "Mulheres", "Total"),
    stringr::str_detect(Idade, "^(1[8-9]|[2-9][0-9]+) (ano|anos)$"),
    .[[13]] == "Total",
    .[[19]] == "Total"
    ) |>
  dplyr::transmute(
    sex = dplyr::case_when(
      Sexo == "Homens" ~ "Male",
      Sexo == "Mulheres" ~ "Female",
      Sexo == "Total" ~ "Total"
    ),
    value = Valor
  ) |>
  dplyr::group_by(sex) |>
  dplyr::summarise(n = sum(value)) |>
  dplyr::ungroup()

census_data <-
  dplyr::bind_rows(
    census_data |>
      dplyr::filter(sex != "Total") |>
      dplyr::mutate(
        n_rel = n / sum(n[sex != "Total"]),
        n_per = round(n_rel * 100, 3)
      ),
    census_data |>
      dplyr::filter(sex == "Total") |>
      dplyr::mutate(n_rel = 1, n_per = 100)
  ) |>
```

```r
  dplyr::as_tibble() |>
  dplyr::arrange(sex)

count <- data |>
  dplyr::select(sex) |>
  dplyr::group_by(sex) |>
  dplyr::summarise(n = dplyr::n()) |>
  dplyr::ungroup() |>
  dplyr::mutate(
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
  ) |>
  dplyr::arrange(dplyr::desc(n_rel)) |>
  dplyr::bind_rows(
    dplyr::tibble(
      sex = "Total",
      n = nrow(tidyr::drop_na(data, sex)),
      n_rel = 1,
      n_per = 100
    )
  )

count <-
  dplyr::left_join(
    count, census_data,
    by = "sex",
    suffix = c("_sample", "_census")
  ) |>
  dplyr::mutate(
    n_rel_diff = n_rel_sample - n_rel_census,
    n_per_diff = n_per_sample - n_per_census
  ) |>
```

```
  dplyr::relocate(
    sex, n_sample, n_census, n_rel_sample, n_rel_census, n_rel_diff,
    n_per_sample, n_per_census, n_per_diff
  )

## Difference from the sample to the census
count
```

sex	n_sample	n_census	n_rel_sample	n_rel_census	n_rel_diff	n_per_sample
Female	52463	69631672	0.662	0.518	0.144	66.243
Male	26735	64809723	0.338	0.482	-0.144	33.757
Total	79198	134441395	1.000	1.000	0.000	100.000

```
count |> dplyr::select(sex, n_per_sample, n_per_census, n_per_diff)
```

sex	n_per_sample	n_per_census	n_per_diff
Female	66.243	51.793	14.45
Male	33.757	48.207	-14.45
Total	100.000	100.000	0.00

```
sum(count$n_per_diff)
#> [1] -7.105427e-15
```

C.9.3 Sex and Age

```
# library(here)
# library(dplyr)
# library(magrittr)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

value <- "Male"

```

```r
data |>
  dplyr::filter(sex == value) |>
  magrittr::extract2("age") |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

name	value
n	26735
n_rm_na	26735
n_na	0
mean	32.4343759740665
var	80.9906211885464
sd	8.99947893983571
min	18
q_1	25.5388888888889
median	31.2583333333333
q_3	37.9319444444444
max	58.7722222222222
iqr	12.3930555555556
skewness	0.617696405622681
kurtosis	2.84390555184727

```r
# See <https://sidra.ibge.gov.br> to learn more.

# library(dplyr)
library(magrittr)
# library(sidrar)
# library(stats)
# library(stringr)
# library(rutils)

rutils:::assert_internet()

# Brazil's 2010 census data
census_data <-
  sidrar::get_sidra(x = 1378) %>% # Don't change the pipe
  dplyr::filter(
```

```r
    Sexo %in% c("Homens", "Mulheres", "Total"),
    stringr::str_detect(Idade, "^(1[8-9]|[2-9][0-9]+) (ano|anos)$"),
    .[[13]] == "Total",
    .[[19]] == "Total"
    ) |>
  dplyr::transmute(
    sex = dplyr::case_when(
      Sexo == "Homens" ~ "Male",
      Sexo == "Mulheres" ~ "Female",
      Sexo == "Total" ~ "Total"
    ),
    age = as.numeric(stringr::str_extract(Idade, "\\d+")),
    value = Valor
  ) |>
  dplyr::group_by(sex) |>
  dplyr::summarise(
    mean = stats::weighted.mean(age, value),
    sd = sqrt(Hmisc::wtd.var(age, value))
  ) |>
  dplyr::ungroup() |>
  dplyr::mutate(
    min = c(18, 18, 18),
    max = c(100, 100, 100)
  ) |>
  dplyr::relocate(sex, mean, sd, min, max) |>
  dplyr::as_tibble()

count <- data |>
  dplyr::select(sex, age) |>
  dplyr::group_by(sex) |>
  dplyr::mutate(sex = as.character(sex)) |>
  dplyr::summarise(
```

```r
      mean = mean(age, na.rm = TRUE),
      sd = stats::sd(age, na.rm = TRUE),
      min = min(age, na.rm = TRUE),
      max = max(age, na.rm = TRUE)
      ) |>
    dplyr::ungroup() |>
    dplyr::bind_rows(
      dplyr::tibble(
        sex = "Total",
        mean = mean(data$age, na.rm = TRUE),
        sd = stats::sd(data$age, na.rm = TRUE),
        min = min(data$age, na.rm = TRUE),
        max = max(data$age, na.rm = TRUE)
      )
    )

  count <-
    dplyr::left_join(
      count,
      census_data,
      by = "sex",
      suffix = c("_sample", "_census")
    ) |>
    dplyr::mutate(mean_diff = mean_sample - mean_census) |>
    dplyr::relocate(
      sex, mean_sample, mean_census, mean_diff, sd_sample, sd_census,
      min_sample, min_census, max_sample, max_census
    )

  count
```

sex	mean_sample	mean_census	mean_diff	sd_sample	sd_census	min_sample
Female	31.754	41.645	-9.890	9.341	16.907	18
Male	32.434	40.373	-7.938	8.999	16.200	18
Total	31.984	41.032	-9.048	9.233	16.582	18

```

count |>
  dplyr::select(
    sex, mean_sample, mean_census, mean_diff, sd_sample, sd_census
    )
```

sex	mean_sample	mean_census	mean_diff	sd_sample	sd_census
Female	31.754	41.645	-9.890	9.341	16.907
Male	32.434	40.373	-7.938	8.999	16.200
Total	31.984	41.032	-9.048	9.233	16.582

```

sum(count$mean_diff)
#> [1] -26.87654
```

C.9.4 Longitudinal range

Sample

```
# library(here)
# library(dplyr)
# library(magrittr)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

stats <-
  data |>
  magrittr::extract2("longitude") |>
  stats_sum(print = FALSE)
```

```r

abs(stats$max - stats$min)
#> [1] 33.023
stats |> list_as_tibble()
```

name	value
n	79198
n_rm_na	79198
n_na	0
mean	-45.9455401815147
var	18.9406905927715
sd	4.35209037047388
min	-67.9869962
q_1	-48.4296364
median	-46.9249578
q_3	-43.7756411
max	-34.9639996
iqr	4.6539953
skewness	0.0156480710174436
kurtosis	5.78918700160139

Brazil

```r
# library(measurements)

change_sign <- function(x) x * (-1)

## Ponta do Seixas, PB (7° 09′ 18″ S, 34° 47′ 34″ O)
min <-
  measurements::conv_unit("34 47 34", from = "deg_min_sec", to = "dec_deg") |>
  as.numeric() |>
  change_sign()

## Nascente do rio Moa, AC (7° 32′ 09″ S, 73° 59′ 26″ O)
max <-
  measurements::conv_unit("73 59 26", from = "deg_min_sec", to = "dec_deg") |>
```

```
  as.numeric() |>
  change_sign()

min
#> [1] -34.79278
max
#> [1] -73.99056
abs(max - min)
#> [1] 39.19778
```

C.9.5 Latitudinal range

Sample

```
# library(here)
# library(dplyr)
# library(magrittr)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

stats <-
  data |>
  magrittr::extract2("latitude") |>
  stats_sum(print = FALSE)

abs(stats$max - stats$min)
#> [1] 32.91596
stats |> list_as_tibble()
```

name	value
n	79198
n_rm_na	79198
n_na	0
mean	-20.8338507528991
var	40.2956396934244
sd	6.34788466289554
min	-30.1087672
q_1	-23.6820636
median	-23.6820636
q_3	-19.9026404
max	2.8071961
iqr	3.7794232
skewness	1.40629570823769
kurtosis	4.67433697579443

Brazil

```r
# library(measurements)

change_sign <- function(x) x * (-1)

## Arroio Chuí, RS (33° 45′ 07″ S, 53° 23′ 50″ O)
min <-
  measurements::conv_unit("33 45 07", from = "deg_min_sec", to = "dec_deg") |>
  as.numeric() |>
  change_sign()

## Nascente do rio Ailã, RR (5° 16′ 19″ N, 60° 12′ 45″ O)
max <-
  measurements::conv_unit("5 16 19", from = "deg_min_sec", to = "dec_deg") |>
  as.numeric()

min
#> [1] -33.75194
max
#> [1] 5.271944
```

```r
abs(max - min)
#> [1] 39.02389
```

C.9.6 Region

```r
# See <https://sidra.ibge.gov.br> to learn more.

# library(dplyr)
# library(sidrar)
# library(stats)
# library(stringr)
# library(rutils)

rutils:::assert_internet()

# Brazil's 2022 census data
census_data <-
  sidrar::get_sidra(x = 4714, variable = 93, geo = "Region") |>
  dplyr::select(dplyr::all_of(c("Valor", "Grande Região"))) |>
  dplyr::transmute(
    col = `Grande Região`,
    n = Valor,
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
    ) |>
  dplyr::mutate(
    col = dplyr::case_when(
      col == "Norte" ~ "North",
      col == "Nordeste" ~ "Northeast",
      col == "Centro-Oeste" ~ "Midwest",
      col == "Sudeste" ~ "Southeast",
      col == "Sul" ~ "South"
```

```r
    )
  ) |>
  dplyr::as_tibble() |>
  dplyr::arrange(dplyr::desc(n_rel))

count <- data |>
  magrittr::extract2("region") |>
  stats_sum(print = FALSE) |>
  magrittr::extract2("count") |>
  dplyr::mutate(
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
    ) |>
  dplyr::arrange(dplyr::desc(n_rel))

count <-
  dplyr::left_join(
    count, census_data, by = "col", suffix = c("_sample", "_census")
  ) |>
  dplyr::mutate(
    n_rel_diff = n_rel_sample - n_rel_census,
    n_per_diff = n_per_sample - n_per_census
  ) |>
  dplyr::relocate(
    col, n_sample, n_census, n_rel_sample, n_rel_census, n_rel_diff,
    n_per_sample, n_per_census, n_per_diff
  )

## Difference from the sample to the census
count
```

col	n_sample	n_census	n_rel_sample	n_rel_census	n_rel_diff	n_per_sample
Southeast	47966	84847187	0.606	0.418	0.188	60.565
South	13560	29933315	0.171	0.147	0.024	17.122
Northeast	9138	54644582	0.115	0.269	-0.154	11.538
Midwest	6563	16287809	0.083	0.080	0.003	8.287
North	1971	17349619	0.025	0.085	-0.061	2.489

```
count |> dplyr::select(col, n_per_sample, n_per_census, n_per_diff)
```

col	n_per_sample	n_per_census	n_per_diff
Southeast	60.565	41.784	18.781
South	17.122	14.741	2.381
Northeast	11.538	26.910	-15.372
Midwest	8.287	8.021	0.266
North	2.489	8.544	-6.055

```
sum(count$n_per_diff)
#> [1] 0.001
```

C.9.7 State

> **TODO**
>
> Compare with 2022 census.

```
# library(here)
# library(dplyr)
# library(magrittr)

source(here::here("R/stats_sum.R"))

data |>
  magrittr::extract2("state") |>
  stats_sum(print = FALSE) |>
  magrittr::extract2("count") |>
  dplyr::mutate(
    n_rel = n / sum(n),
```

```
    n_per = round(n_rel * 100, 3)
    ) |>
  dplyr::arrange(dplyr::desc(n_rel))
```

col	n	n_rel	n_per
São Paulo	26379	0.333	33.308
Minas Gerais	10115	0.128	12.772
Rio de Janeiro	9381	0.118	11.845
Paraná	5517	0.070	6.966
Rio Grande do Sul	4097	0.052	5.173
Santa Catarina	3946	0.050	4.982
Goiás	2674	0.034	3.376
Bahia	2522	0.032	3.184
Espírito Santo	2091	0.026	2.640
Distrito Federal	2087	0.026	2.635
Pernambuco	1550	0.020	1.957
Ceará	1398	0.018	1.765
Mato Grosso do Sul	1014	0.013	1.280
Pará	938	0.012	1.184
Rio Grande do Norte	789	0.010	0.996
Mato Grosso	788	0.010	0.995
Paraíba	773	0.010	0.976
Maranhão	652	0.008	0.823
Sergipe	533	0.007	0.673
Alagoas	526	0.007	0.664
Rondônia	401	0.005	0.506
Piauí	395	0.005	0.499
Tocantins	268	0.003	0.338
Acre	132	0.002	0.167
Roraima	119	0.002	0.150
Amapá	113	0.001	0.143

D Appendice: Chapter 5 supplemental material

> **!** Important
>
> You are reading the work-in-progress of this thesis. This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

E Appendice: Chapter 6 supplemental material

> **Note**
>
> You are reading the work-in-progress of this thesis. This chapter should be readable but is currently undergoing final polishing.

E.1 Hypothesis

Populations residing near the equator (latitude 0°) exhibit, on average, a shorter/morning circadian phenotype when compared to populations residing near the poles of the planet (Horzum et al. 2015; Hut et al. 2013; Leocadio-Miguel et al. 2017, 2014; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler and Rahafar 2017).

The study hypothesis was tested using nested models of multiple linear regressions. The main idea of nested models is to verify the effect of the inclusion of one or more predictors in the model variance explanation (i.e., the R^2) (Allen 1997). This can be made by creating a restricted model and then comparing it with a full model. Hence, the hypothesis can be schematized as follows.

$$\begin{cases} H_0 : R^2_{res} >= R^2_{full} \\ H_a : R^2_{res} < R^2_{full} \end{cases}$$

The general equation for the F-test (Allen 1997, 113) :

$$F = \frac{R^2_F - R^2_R/(k_F - k_R)}{(1 - R^2_F)/(N - k_F - 1)}$$

Where:

- R^2_F = Coefficient of determination for the **full** model
- R^2_R = Coefficient of determination for the **restricted** model
- k_F = Number of independent variables in the full model
- k_R = Number of independent variables in the restricted model
- N = Number of observations in the sample

$$\mathrm{F} = \frac{\text{Additional Var. Explained}/\text{Additional d.f. Expended}}{\text{Var. unexplained}/\text{d.f. Remaining}}$$

E.2 Assumptions

See DeGroot and Schervish (2012, 736–38) to learn more.

> ⚠ Warning
>
> **The predictor is known**. Either the vectors $z_1, \dots, z_n$ are known ahead of time, or they are the observed values of random vectors $Z_1, \dots, Z_n$ on whose values we condition before computing the joint distribution of $(Y_1, \dots, Y_n)$.

> ⚠ Warning
>
> **Normality**. For $i = 1, \dots, n$, the conditional distribution of Y_i given the vectors $z_1, \dots, z_n$ is a normal distribution (**normality assumption**).

> ⚠ Warning
>
> **Linear mean**. There is a vector of parameters $\beta = (\beta_0, \dots, \beta_{p-1})$ such that the conditional mean of Y_i given the values $z_1, \dots, z_n$ has the form
>
> $$z_{i0}\beta_0 + z_{i1}\beta_1 + \cdots + z_{ip-1}\beta_{p-1}$$
>
> for $i = 1, \dots, n$ (**zero error mean assumption**).

> ⚠ Warning
>
> **Common variance**. The observations $Y_1, \dots, Y_n$ have the same variance σ^2 (**homoscedasticity assumption**).

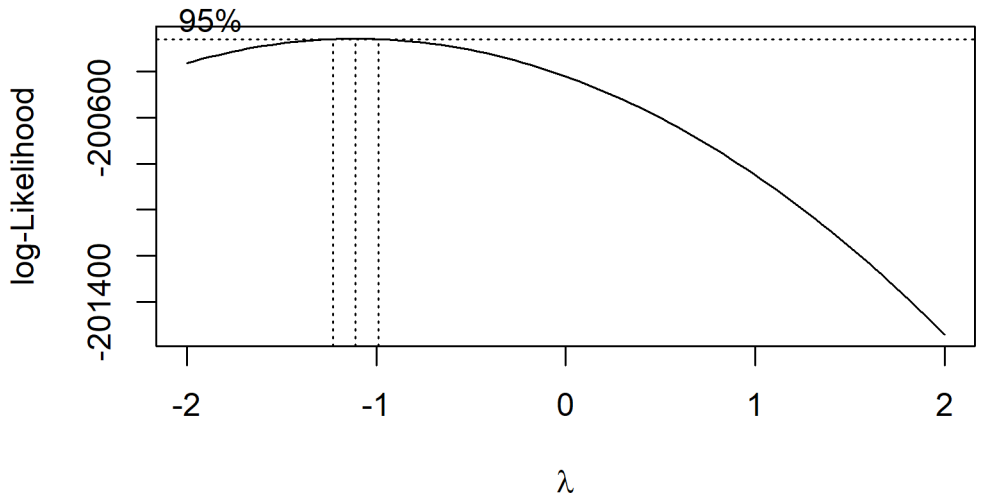
> ⚠ Warning
>
> **Independence**. The random variables $Y_1, \dots, Y_n$ are independent given the observed $z_1, \dots, z_n$ (**independent errors assumption**).

E.3 Data preparation

```r
# library(dplyr)
# library(here)
library(targets)
# library(tidyr)

source(here::here("R/utils.R"))

utc_minus_3_states <- c(
  "Amapá", "Pará", "Maranhão", "Tocantins", "Piauí", "Ceará",
  "Rio Grande do Norte", "Paraíba", "Pernambuco", "Alagoas", "Sergipe",
  "Bahia", "Distrito Federal", "Goiás", "Minas Gerais", "Espírito Santo",
  "Rio de Janeiro", "São Paulo", "Paraná", "Santa Catarina",
  "Rio Grande do Sul"
)

data <-
  targets::tar_read("geocoded_data", store = here::here("_targets")) |>
  dplyr::filter(state %in% utc_minus_3_states) |>
  dplyr::select(msf_sc, age, sex, state, latitude, longitude) |>
  dplyr::mutate(msf_sc = transform_time(msf_sc)) |>
  tidyr::drop_na(msf_sc, age, sex, latitude)
```

E.4 Restricted model

```r
# library(MASS)

box_cox <- MASS::boxcox(msf_sc ~ age + sex, data = data)
```




```r
lambda <- box_cox$x[which.max(box_cox$y)]

lambda
#> [1] -1.1111
```

```r
# library(stats)

res_model <- stats::lm(
  ((msf_sc^lambda - 1) / lambda) ~ age + sex, data = data
  )
```

```r
# library(broom)

broom::tidy(res_model)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.9	0	513579298.250	0
age	0.0	0	-65.128	0
sexMale	0.0	0	13.020	0

```r
# library(broom)
# library(dplyr)
# library(tidyr)

broom::glance(res_model) |> tidyr::pivot_longer(cols = dplyr::everything())
```

name	value
r.squared	0.054
adj.r.squared	0.054
sigma	0.000
statistic	2178.876
p.value	0.000
df	2.000
logLik	1106194.897
AIC	-2212381.794
BIC	-2212344.801
deviance	0.000
df.residual	76741.000
nobs	76744.000

```r
# library(olsrr)

# res_model |> olsrr::ols_regress()
res_model |> summary()
#>
#> Call:
#> stats::lm(formula = ((msf_sc^lambda - 1)/lambda) ~ age + sex,
#>     data = data)
#>
#> Residuals:
#>           Min            1Q        Median            3Q           Max
#> -0.0000004859 -0.0000000911 -0.0000000031  0.0000000916  0.0000004204
#>
#> Coefficients:
#>                     Estimate       Std. Error     t value Pr(>|t|)
#> (Intercept)  0.8999976603602  0.0000000017524 513579298.2   <2e-16 ***
```

```
#> age         -0.0000000033812  0.0000000000519       -65.1   <2e-16 ***
#> sexMale      0.0000000132309  0.0000000010162        13.0   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.000000133 on 76741 degrees of freedom
#> Multiple R-squared:  0.0537, Adjusted R-squared:  0.0537
#> F-statistic: 2.18e+03 on 2 and 76741 DF,  p-value: <2e-16
```

E.4.1 Residual diagnostics

Normality and zero mean error assumption.

```
# library(here)
# library(stats)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

res_model |>
  stats::residuals() |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

name	value
n	76744
n_rm_na	76744
n_na	0
mean	6.60699976667332e-23
var	0.000000000000017685286682 6985
sd	0.000000132986039427823
min	-0.000000485865195534305
q_1	-0.000000091113801656790 8
median	-0.000000003135303247871 35
q_3	0.000000091553820345483
max	0.000000420368932360539
iqr	0.000000182667622002274
skewness	-0.0105262146639209
kurtosis	2.82813923301771

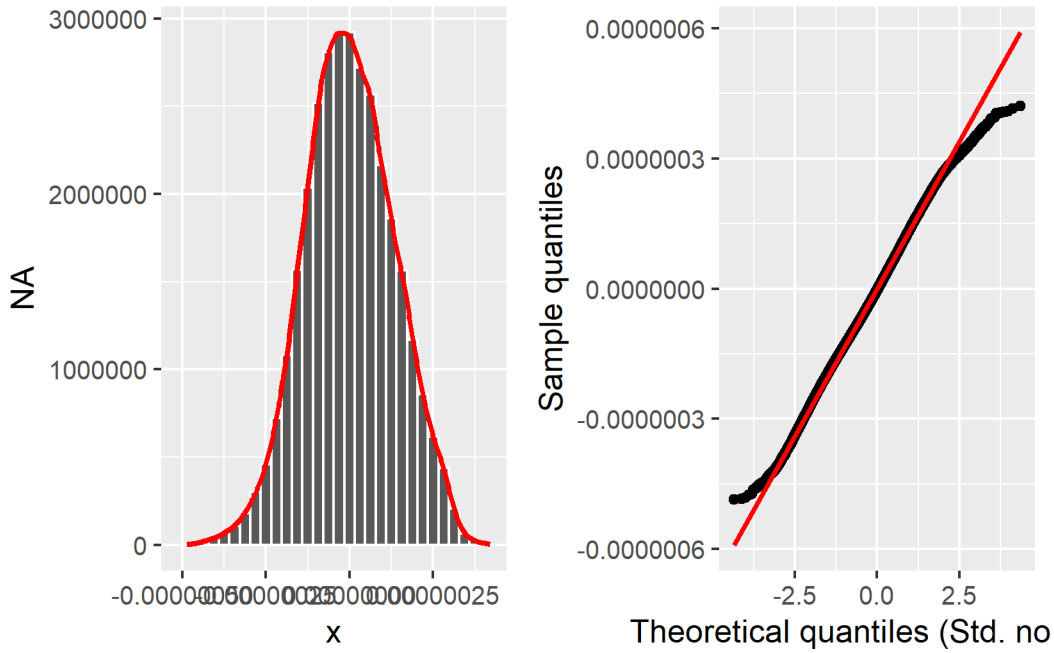
```
# See `?moments::agostino.test` & `?fBasics::dagoTest()` to learn more.

# library(fBasics)
# library(moments)

source(here::here("R/normality_sum.R"))

res_model |>
  stats::residuals() |>
  normality_sum()
```

test	p_value
Anderson-Darling	0.000
Bonett-Seier	0.000
Cramer-von Mises	0.000
D'Agostino Omnibus Test	NA
D'Agostino Skewness Test	0.234
D'Agostino Kurtosis Test	NA
Jarque–Bera	0.000
Lilliefors (K-S)	0.000
Pearson chi-square	0.000
Shapiro-Francia	NA
Shapiro-Wilk	NA

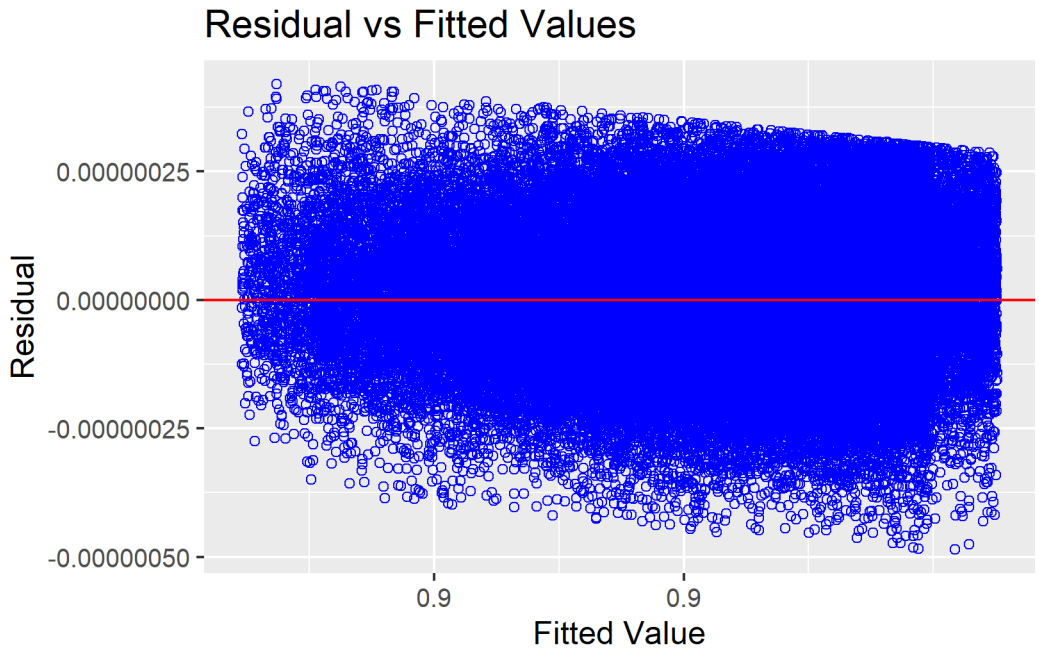
Correlation between observed residuals and expected residuals under normality.

```r
# library(olsrr)

res_model |> olsrr::ols_test_correlation()
#> [1] 0.99929
```

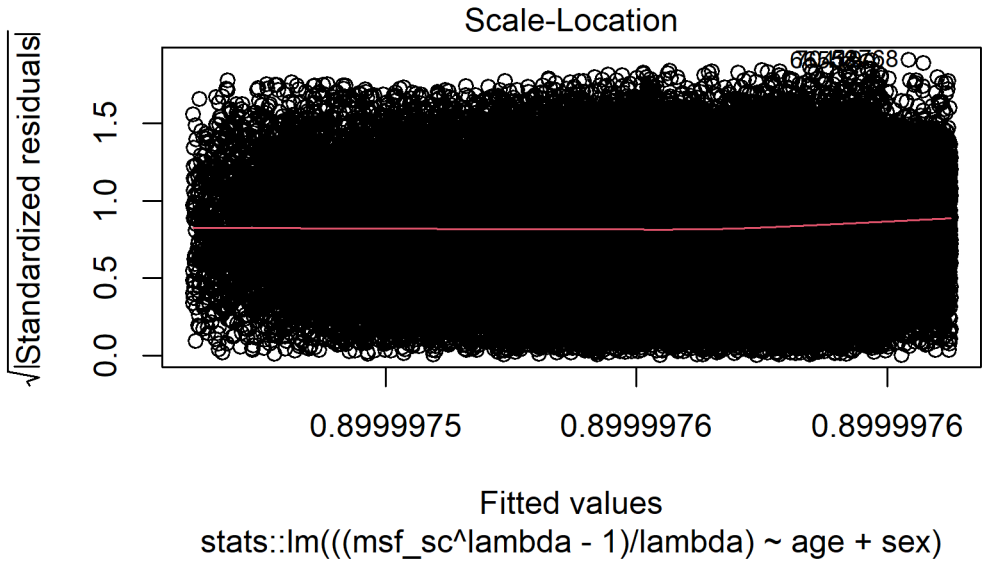
```r
# library(cowplot)
# library(olsrr)
# library(stats)

source(here::here("R/test_normality.R"))

# res_model |> olsrr::ols_plot_resid_qq()

qq_plot <- res_model |>
  stats::residuals() |>
  plot_qq(print = FALSE)

hist_plot <- res_model |>
  stats::residuals() |>
  plot_hist(print = FALSE)

cowplot::plot_grid(hist_plot, qq_plot, ncol = 2, nrow = 1)
```




```r
# library(olsrr)

# Linear mean assumption

res_model |> olsrr::ols_plot_resid_fit()
```




```r
res_model |> plot(3)
```



E.4.2 Heteroskedasticity

Homoscedasticity assumption.

```
# library(olsrr)

# "It test whether variance of errors from a regression is dependent on the values of a indep

res_model |> olsrr::ols_test_breusch_pagan()
#>
#>  Breusch Pagan Test for Heteroskedasticity
#>  -----------------------------------------
#>  Ho: the variance is constant
#>  Ha: the variance is not constant
#>
#>                           Data
#>  ----------------------------------------------------------
#>  Response : ((msf_sc^lambda - 1)/lambda)
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
```

```
#>         Test Summary
#>  ----------------------------
#>  DF            =    1
#>  Chi2          =    70149.3586
#>  Prob > Chi2   =    0.0000
```

```
# library(olsrr)

res_model |> olsrr::ols_test_score()
#>
#>  Score Test for Heteroskedasticity
#>  ---------------------------------
#>  Ho: Variance is homogenous
#>  Ha: Variance is not homogenous
#>
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>       Test Summary
#>  ------------------------
#>  DF            =    1
#>  Chi2          =    0.000
#>  Prob > Chi2   =    1.000
```

E.4.3 Collinearity diagnostics

Independence assumption.

```
# library(olsrr)

res_model |> olsrr::ols_coll_diag()
#> Tolerance and Variance Inflation Factor
#> ------------------------------------
#>   Variables Tolerance    VIF
```

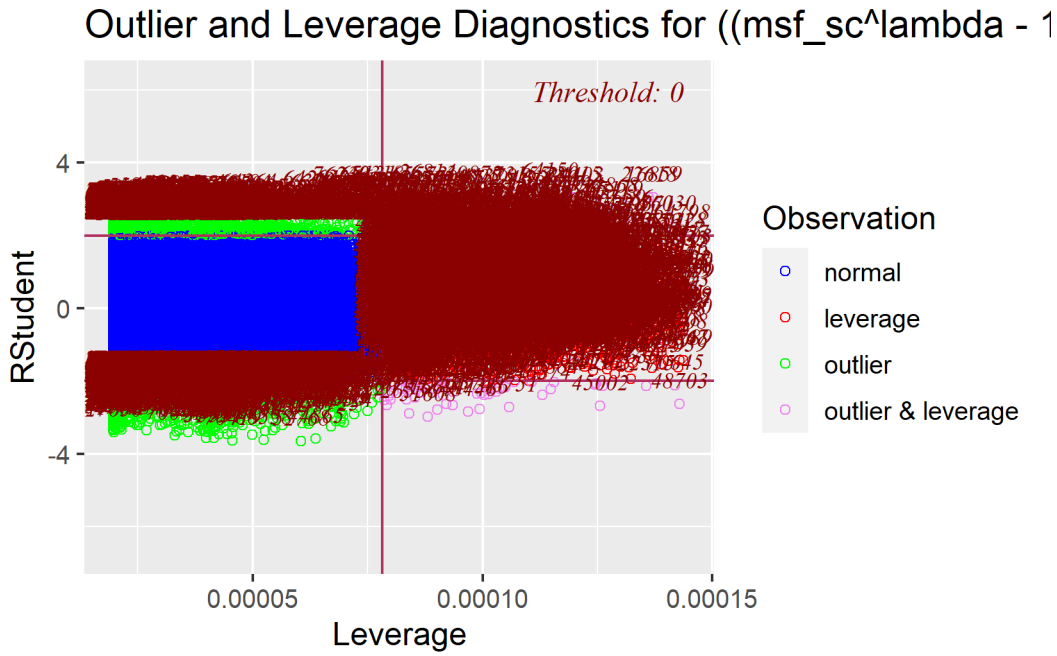
```
#> 1       age    0.9988 1.0012
#> 2   sexMale    0.9988 1.0012
#>
#>
#> Eigenvalue and Condition Index
#> ------------------------------
#>   Eigenvalue Condition Index intercept      age   sexMale
#> 1   2.422418          1.0000  0.011753 0.011936 0.0669897
#> 2   0.538450          2.1211  0.015824 0.018848 0.9280439
#> 3   0.039132          7.8679  0.972423 0.969216 0.0049664
```

The variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables (e.g., VIF equal to 1 = variables are not correlated).

```
# library(car)

res_model |> car::vif()
#>    age    sex
#> 1.0012 1.0012
```

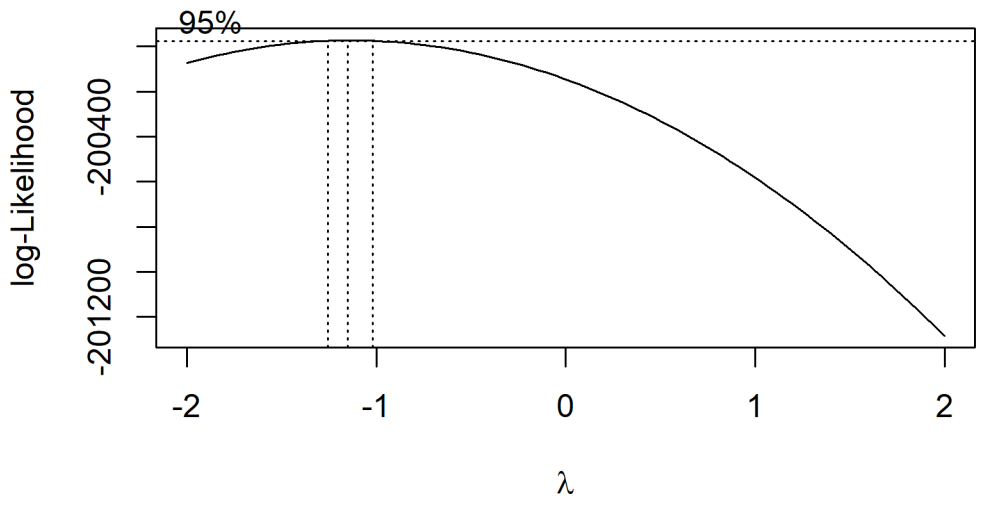
E.4.4 Measures of influence

```
# library(olsrr)

res_model |> olsrr::ols_plot_resid_lev()
```



E.5 Full model

```
# library(MASS)

box_cox <- MASS::boxcox(
  msf_sc ~ age + sex + latitude, data = data
  )
```



```r

box_cox$x[which.max(box_cox$y)] # lambda
#> [1] -1.1515
```

```r
lambda # The same lambda of the restricted model
#> [1] -1.1111
```

```r
# library(stats)

full_model <- stats::lm(
  ((msf_sc^lambda - 1) / lambda) ~ age + sex + latitude,
  data = data
  )
```

```r
# library(broom)

# ?broom::tidy.lm
broom::tidy(full_model)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.9	0	391908052.847	0
age	0.0	0	-66.928	0
sexMale	0.0	0	13.558	0
latitude	0.0	0	-23.852	0

```r
# library(broom)
# library(dplyr)
# library(tidyr)

# ?broom::glance.lm
broom::glance(full_model) |>
  tidyr::pivot_longer(cols = dplyr::everything())
```

name	value
r.squared	0.061
adj.r.squared	0.061
sigma	0.000
statistic	1652.979
p.value	0.000
df	3.000
logLik	1106478.331
AIC	-2212946.661
BIC	-2212900.420
deviance	0.000
df.residual	76740.000
nobs	76744.000

```
# full_model |> olsrr::ols_regress()
full_model |> summary()
#>
#> Call:
#> stats::lm(formula = ((msf_sc^lambda - 1)/lambda) ~ age + sex +
#>     latitude, data = data)
#>
#> Residuals:
#>           Min            1Q        Median            3Q           Max
#> -0.0000004874 -0.0000000911 -0.0000000034  0.0000000912  0.0000004328
#>
#> Coefficients:
#>                     Estimate       Std. Error     t value Pr(>|t|)
#> (Intercept)  0.8999976247783  0.0000000022965 391908052.9   <2e-16 ***
#> age         -0.0000000034710  0.0000000000519       -66.9   <2e-16 ***
#> sexMale      0.0000000137296  0.0000000010127        13.6   <2e-16 ***
#> latitude    -0.0000000018222  0.0000000000764       -23.9   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.000000132 on 76740 degrees of freedom
```

```
#> Multiple R-squared:  0.0607, Adjusted R-squared:  0.0607
#> F-statistic: 1.65e+03 on 3 and 76740 DF,  p-value: <2e-16
```

E.5.1 Residual diagnostics

Normality and zero mean error assumption.

```
# library(here)
# library(stats)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

full_model |>
  stats::residuals() |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

name	value
n	76744
n_rm_na	76744
n_na	0
mean	4.85272564733669e-24
var	0.000000000000001755513613045 61
sd	0.000000132495796652 03
min	-0.000000487410752460545
q_1	-0.0000000910649425186321
median	-0.000000003374344652286
q_3	0.000000091189958883 9585
max	0.000000432826012898983
iqr	0.000000182254901402591
skewness	0.000655994107765645
kurtosis	2.82688323293117

```
# library(here)
# library(stats)

```

```
source(here::here("R/normality_sum.R"))

full_model |>
  stats::residuals() |>
  normality_sum()
```

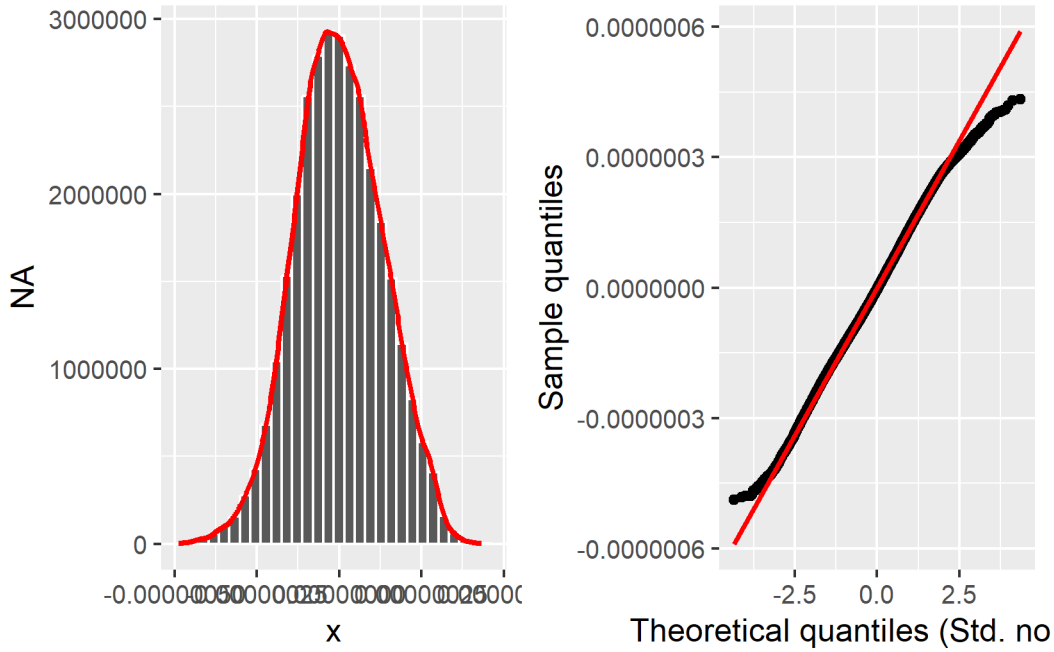
test	p_value
Anderson-Darling	0.000
Bonett-Seier	0.000
Cramer-von Mises	0.000
D'Agostino Omnibus Test	NA
D'Agostino Skewness Test	0.941
D'Agostino Kurtosis Test	NA
Jarque–Bera	0.000
Lilliefors (K-S)	0.000
Pearson chi-square	0.000
Shapiro-Francia	NA
Shapiro-Wilk	NA

Correlation between observed residuals and expected residuals under normality.

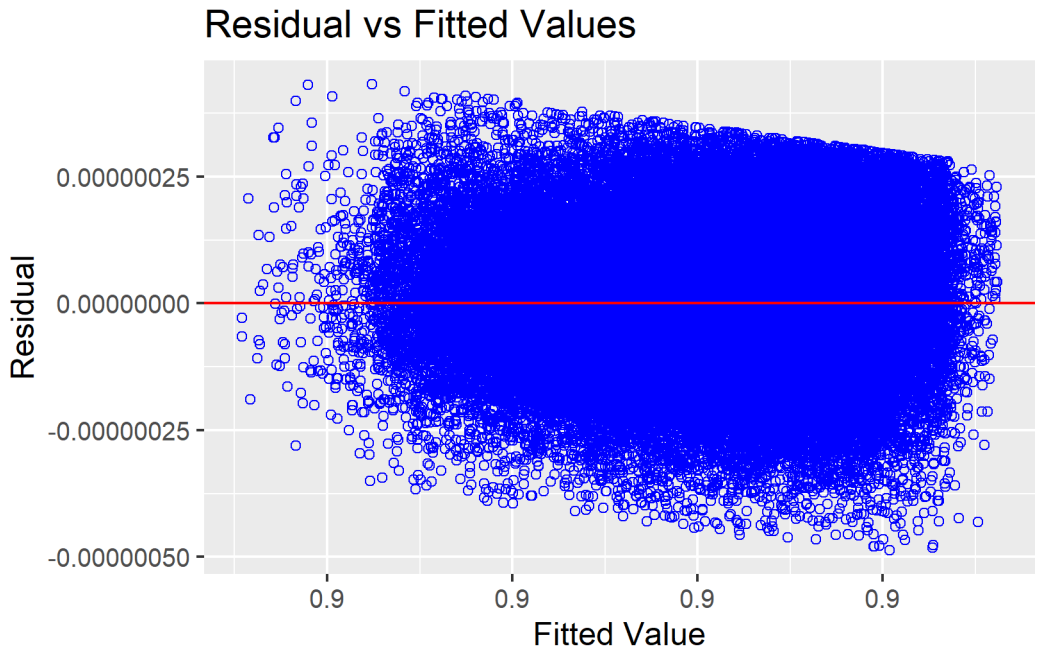
```
# library(olsrr)

full_model |> olsrr::ols_test_correlation()
#> [1] 0.99929
```

```
# library(here)
# library(cowplot)
# library(olsrr)
# library(stats)

source(here::here("R/test_normality.R"))

# full_model |> olsrr::ols_plot_resid_qq()

hist_plot <- full_model |>
  stats::residuals() |>
```

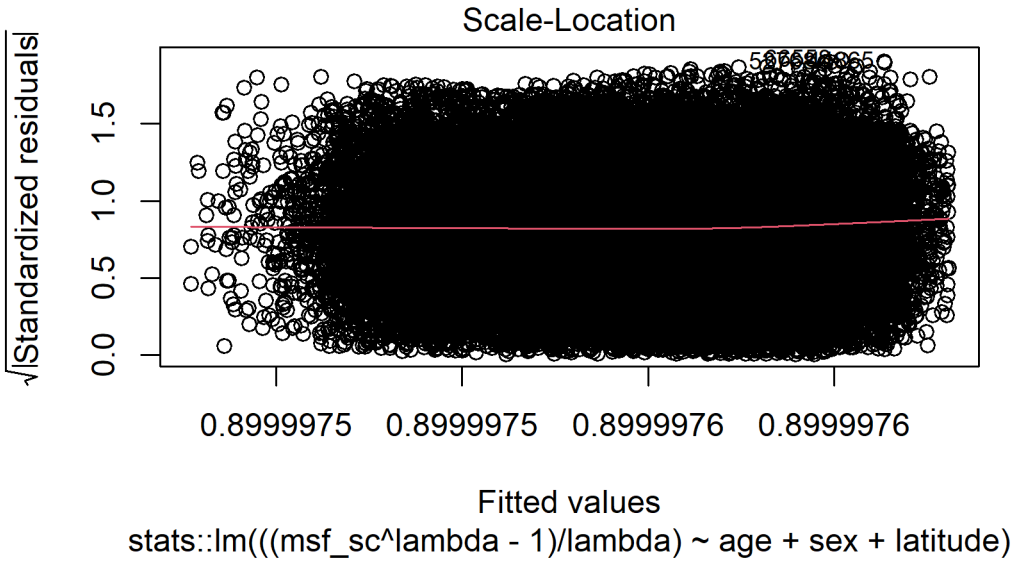
```r
  plot_hist(print = FALSE)

qq_plot <- full_model |>
  stats::residuals() |>
  plot_qq(print = FALSE)

cowplot::plot_grid(hist_plot, qq_plot, ncol = 2, nrow = 1)
```




```r
# library(olsrr)

full_model |> olsrr::ols_plot_resid_fit()
```




```
full_model |> plot(3)
```



E.5.2 Heteroskedasticity

Homoscedasticity assumption.

```r
# library(olsrr)

full_model |> olsrr::ols_test_breusch_pagan()
#>
#>  Breusch Pagan Test for Heteroskedasticity
#>  -----------------------------------------
#>  Ho: the variance is constant
#>  Ha: the variance is not constant
#>
#>                           Data
#>  ----------------------------------------------------------
#>  Response : ((msf_sc^lambda - 1)/lambda)
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>         Test Summary
#>  ----------------------------
#>  DF            =    1
#>  Chi2          =    70101.1634
#>  Prob > Chi2   =    0.0000
```

```r
# library(olsrr)

full_model |> olsrr::ols_test_score()
#>
#>  Score Test for Heteroskedasticity
#>  ---------------------------------
#>  Ho: Variance is homogenous
#>  Ha: Variance is not homogenous
#>
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>       Test Summary
```

```
#>  ------------------------
#>  DF            =    1
#>  Chi2          =    0.000
#>  Prob > Chi2   =    1.000
```

E.5.3 Collinearity diagnostics

Independence assumption.

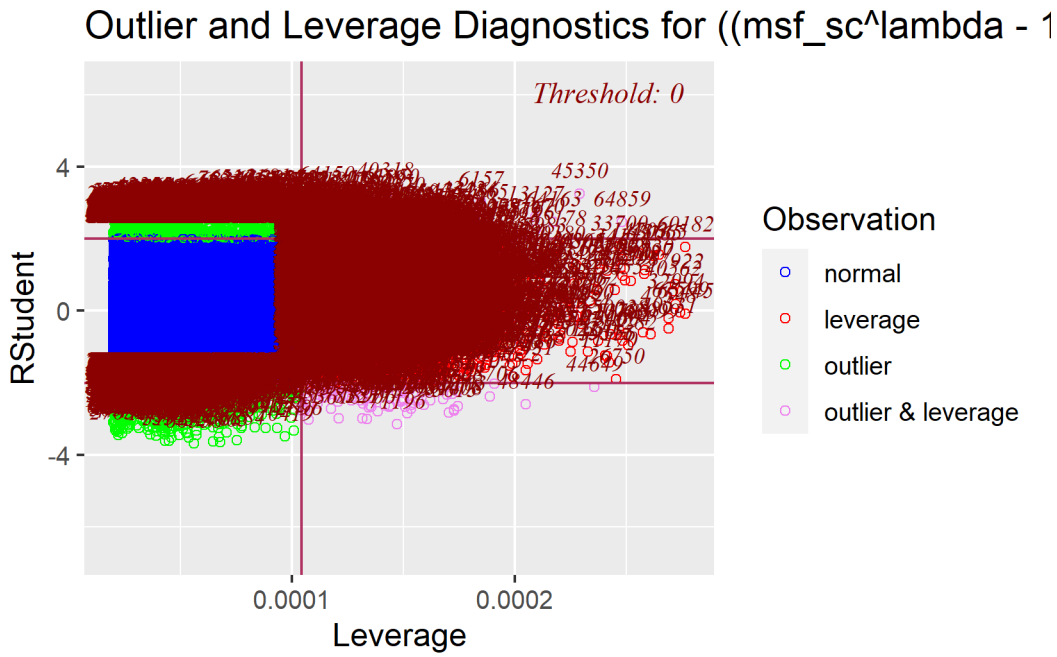
```
# library(olsrr)

full_model |> olsrr::ols_coll_diag()
#> Tolerance and Variance Inflation Factor
#> ------------------------------------
#>   Variables Tolerance    VIF
#> 1       age   0.99354 1.0065
#> 2   sexMale   0.99838 1.0016
#> 3  latitude   0.99441 1.0056
#>
#>
#> Eigenvalue and Condition Index
#> ----------------------------
#>   Eigenvalue Condition Index  intercept       age   sexMale  latitude
#> 1   3.312504          1.0000 0.00377395 0.0064918 0.0304493 0.0068553
#> 2   0.584652          2.3803 0.00328127 0.0064143 0.9588857 0.0083393
#> 3   0.073700          6.7042 0.00040414 0.5063551 0.0023826 0.5659326
#> 4   0.029145         10.6609 0.99254063 0.4807389 0.0082824 0.4188728
```

The variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables (e.g., VIF equal to 1 = variables are not correlated).

```r
# library(car)

full_model |> car::vif()
#>       age       sex latitude
#>    1.0065    1.0016    1.0056
```

E.5.4 Measures of influence

```r
# library(olsrr)

full_model |> olsrr::ols_plot_resid_lev()
```



E.6 Nested regression models test

$$\begin{cases} H_0 : R^2_{res} >= R^2_{full} \\ H_a : R^2_{res} < R^2_{full} \end{cases}$$

$$F = \frac{R^2_F - R^2_R/(k_F - k_R)}{(1 - R^2_F)/(N - k_F - 1)}$$

$$F = \frac{\text{Additional Var. Explained} / \text{Additional d.f. Expended}}{\text{Var. unexplained} / \text{d.f. Remaining}}$$

```r
# library(dplyr)
# library(here)

source(here::here("R/utils-stats.R"))

dplyr::tibble(
  name = c("r_squared_res", "r_squared_full", "diff"),
  value = c(
  r_squared(res_model), r_squared(full_model),
  r_squared(full_model) - r_squared(res_model)
  )
)
```

name	value
r_squared_res	0.054
r_squared_full	0.061
diff	0.007

```r
# library(stats)

stats::anova(res_model, full_model)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
76741	0	NA	NA	NA	NA
76740	0	1	0	568.94	0

```r
# library(stats)
# library(here)

source(here::here("R/utils-stats.R"))

```

```r
n <- nrow(data)
k_res <- length(stats::coefficients(res_model)) - 1
k_full <- length(stats::coefficients(full_model)) - 1

((r_squared(full_model) - r_squared(res_model)) / (k_full - k_res)) / ((1 - r_squared(full_m
#> [1] 568.94
```

$$f^2 = \frac{\mathrm{R}_F^2 - \mathrm{R}_R^2}{1 - \mathrm{R}_F^2}$$

$$f^2 = \frac{\text{Additional Var. Explained}}{\text{Var. unexplained}}$$

```r
# library(here)

source(here::here("R/cohens_f_squared.R"))
source(here::here("R/utils-stats.R"))

cohens_f_squared_summary(
  adj_r_squared(res_model),
  adj_r_squared(full_model)
  )
```

name	value
f_squared	0.00740068896515648
effect_size	Negligible

E.7 Group test

$$\begin{cases} \mathrm{H}_0 : \mathrm{MSF}_{\mathrm{sc}}^{0^o} >= \mathrm{MSF}_{\mathrm{sc}}^{30^o} \\ \mathrm{H}_a : \mathrm{MSF}_{\mathrm{sc}}^{0^o} < \mathrm{MSF}_{\mathrm{sc}}^{30^o} \end{cases}$$

```r
# library(dplyr)
# library(here)
# library(magrittr)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

group_1 <- "Amapá" # Boa vista (0° 2′ 18.84″ N, 51° 3′ 59.1″ W)

msf_sc_group_1 <-
  data |>
  dplyr::filter(state == group_1) |>
  magrittr::extract2("msf_sc")

stats_sum_group_1 <-
  data |>
  dplyr::filter(state == group_1) |>
  magrittr::extract2("msf_sc") |>
  stats_sum(print = FALSE)

stats_sum_group_1 |> list_as_tibble()
```

name	value
n	113
n_rm_na	113
n_na	0
mean	103564.475347661
var	28720417.6933099
sd	5359.14337308771
min	92100
q_1	99000
median	104164.285714286
q_3	107142.857142857
max	114450
iqr	8142.85714285714
skewness	-0.0419134073448402
kurtosis	2.19330502090315

```r
# library(dplyr)
# library(here)
# library(magrittr)

source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

group_2 <- "Rio Grande do Sul" # Porto Alegre (30° 01' 58" S, 51° 13' 48" O)

msf_sc_group_2 <- data |>
  dplyr::filter(state == group_2) |>
  magrittr::extract2("msf_sc")

stats_sum_group_2 <-
  data |>
  dplyr::filter(state == group_2) |>
  magrittr::extract2("msf_sc") |>
  stats_sum(print = FALSE)

stats_sum_group_2 |> list_as_tibble()
```

name	value
n	4097
n_rm_na	4097
n_na	0
mean	103651.485407441
var	26397872.5737977
sd	5137.88600241361
min	88092.8571428571
q_1	99857.1428571429
median	103328.571428571
q_3	107057.142857143
max	117064.285714286
iqr	7200
skewness	0.263604552812467
kurtosis	2.69168212740063

```r
# library(stats)

stats::t.test(msf_sc_group_1, msf_sc_group_2, alternative = "less")
#>
#>  Welch Two Sample t-test
#>
#> data:  msf_sc_group_1 and msf_sc_group_2
#> t = -0.17, df = 118, p-value = 0.43
#> alternative hypothesis: true difference in means is less than 0
#> 95 percent confidence interval:
#>    -Inf 759.34
#> sample estimates:
#> mean of x mean of y
#>    103564    103651
```

```r
# library(dplyr)

dplyr::tibble(
  name = c("mean_group_1", "mean_group_2", "diff"),
  value = c(
  stats_sum_group_1$mean, stats_sum_group_2$mean,
  stats_sum_group_2$mean - stats_sum_group_1$mean
  )
)
```

name	value
mean_group_1	103564.48
mean_group_2	103651.49
diff	87.01

See Frey (2022, 224–27) to learn more.

$$d = \frac{\mu_1 - \mu_2}{\sigma_e}$$

```r
# library(effsize)

effsize::cohen.d(msf_sc_group_1, msf_sc_group_2)
#>
#> Cohen's d
#>
#> d estimate: -0.016915 (negligible)
#> 95 percent confidence interval:
#>    lower    upper
#> -0.20387  0.17004
```