Introduction to Applied Science

Rex W. Douglass

11/4/22

Table of contents

1 Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

Part I

Introduction

2 Introduction

This is a book created from markdown and executable code.

See (**knuth84?**) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

3 Modeling Literature

Bayesian Workflow Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, Martin Modrák https://arxiv.org/abs/2011.01808

How to avoid machine learning pitfalls: a guide for academic researchers Michael A. Lones https://arxiv.org/abs/2108.02497

Information geometry and divergences https://franknielsen.github.io/IG/#bookIG

Statistical Rethinking: A Bayesian Course with Examples in R and Stan (& PyMC3 & brms) https://xcelab.net/rm/statistical-rethinking/ https://www.youtube.com/playlist?list=PLDcUM9US4XdMROZ OIRtIK0aOynbgZN

ML Frameworks Interoperability Cheat Sheet http://bl.ocks.org/miguelusque/raw/f44a8e729896a96d0a3e4b07b5

Regression and Other Stories, Andrew Gelman, Jennifer Hill, Aki Vehtari copy of the book https://users.aalto.fi/~ave/ROS.pdf

tidybayes: Bayesian analysis + tidy data + geoms

Graphical Data Analysis with R Antony Unwin

Data Visualization A practical introduction, Kieran Healy

Bayes Rules! An Introduction to Applied Bayesian Modeling, Alicia A. Johnson, Miles Q. Ott, Mine Dogucu, 2021-12-01

Bayesian Statistics Independent readings course on Bayesian statistics with R and Stan, Andrew Heiss and Meng Ye, Fall 2022 https://bayesf22-notebook.classes.andrewheiss.com/rethinking/ https://bayesf22-notebook.classes.andrewheiss.com/bayes-rules/

Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis

An Introduction to Proximal Causal Learning

A Selective Review of Negative Control Methods in Epidemiology

Backpropagation is not just the chain rule%2C%20to%20predict%20y.)

Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs Andrew Gelman &Guido Imbens

R Markdown Cookbook Yihui Xie, Christophe Dervieux, Emily Riederer 2022-11-07 https://bookdown.org/yihui/rmarkdown-cookbook/

Understanding Machine Learning: From Theory to Algorithms https://www.cs.huji.ac.il/w~shais/Understanding machine-learning-theory-algorithms.pdf

https://simplystatistics.org/

Estimation Prediction, Estimation, and Attribution

The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant

A Parsimonious Tour of Bayesian Model Uncertainty

Causal Inference for the Brave and True

https://bayesiancomputationbook.com/welcome.html

Measurement error and the replication crisis The assumption that measurement error always reduces effect sizes is false https://www.science.org/doi/10.1126/science.aal3618

https://journals.sagepub.com/doi/abs/10.1177/00031224211004187#:~:text=The%20estimand%20is%20the%2

Exploring the Dynamics of Latent Variable Models https://www.cambridge.org/core/journals/political-analysis/article/abs/exploring-the-dynamics-of-latent-variable-models/CBE116F37900DAE957B2D7EB53DB0

https://github.com/HenrikBengtsson/matrixStats

Let's Git started

https://github.com/facebookresearch/StarSpace

https://dennybritz.com/posts/wildml/understanding-convolutional-neural-networks-for-nlp/

What's Wrong With My Time Series Blog post by Alex Smolyanskaya ALEX SMOLYANSKAYA February 28, 2017 - San Francisco, CA Tweet this post! Post on LinkedIn What's wrong with my time series? Model validation without a hold-out set https://multithreaded.stitchfix.com/blog/20 wrong-with-my-time-series/

ggRandomForests: Exploring Random Forest Survival https://arxiv.org/pdf/1612.08974.pdf

https://districtdatalabs.silvrback.com/time-maps-visualizing-discrete-events-across-many-timescales

Explained Visually https://setosa.io/ev/

https://github.com/google/BIG-bench/blob/main/docs/paper/BIG-bench.pdf

Two Experiments in Peer Review: Posting Preprints and Citation Bias

Random Walk: A Modern Introduction Gregory F. Lawler and Vlada Limic

Can Transformers be Strong Treatment Effect Estimators? https://arxiv.org/pdf/2202.01336v1.pdf

Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition https://bookdown.org/content/4857/

Patches Are All You Need? https://openreview.net/forum?id=TVHS5Y4dNvM

The validate R-package makes it super-easy to check whether data lives up to expectations you have based on domain knowledge. It works by allowing https://github.com/data-cleaning/validate

Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong https://journals.sagepub.com/doi/10.1080/07388940500339167

autoxgboost https://github.com/ja-thomas/autoxgboost

1,500 scientists lift the lid on reproducibility https://www.nature.com/articles/533452a

Methodology over metrics: current scientific standards are a disservice to patients and society https://www.jclinepi.com/article/S0895-4356(21)00170-0/fulltext

bper: Bayesian Prediction for Ethnicity and Race https://github.com/bwilden/bper

Automatic Differentiation Variational Inference https://www.jmlr.org/papers/volume18/16-107/16-107.pdf

What are the most important statistical ideas of the past 50 years? Andrew Gelman, Aki Vehtari https://arxiv.org/pdf/2012.00174.pdf

Why Propensity Scores Should Not Be Used for Matching https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/

PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R https://cran.r-project.org/web/packages/PRROC/vignettes/PRROC.pdf

On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives https://arxiv.org/abs/1902.10286

['Trust Us': Open Data and Preregistration in Political Science and International Relations] https://osf.io/preprints/metaarxiv/8h2bp/

pals https://cran.r-project.org/web/packages/pals/vignettes/pals_examples.html

Greedy Function Approximation: A Gradient Boosting Machine https://jerryfriedman.su.domains/ftp/trebst.pd

Natural Scales in Geographical Patterns https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5379183/

https://daattali.com/shiny/timevis-demo/

https://www.extremetech.com/computing/151980-inside-ibms-67-billion-sage-the-largest-computer-ever-built

Faux peer-reviewed journals: a threat to research integrity http://deevybee.blogspot.com/2020/12/?m=1

https://github.com/mmxgn/spacy-clausie

http://deevybee.blogspot.com/2020/12/?m=1

http://www.deeplearningbook.org

Statistical Nonsignificance in Empirical Economics https://www.aeaweb.org/articles?id=10.1257/aeri.20190252&

Acquiescence Bias Inflates Estimates of Conspiratorial Beliefs and Political Misperceptions∗ Seth J. Hill† Margaret E. Roberts‡ October 25, 2021 http://www.margaretroberts.net/wp-content/uploads/2021/10/hillroberts_acqbiaspoliticalbeliefs.pdf

The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable https://www.nature.com/articles/s41591-021-01535-y

https://www.math.uzh.ch/pages/varrank/index.html

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing https://arxiv.org/pdf/2107.13586.pdf

How should variable selection be performed with multiply imputed data? https://onlinelibrary.wiley.com/doi/pd

Feature Interactions in XGBoost https://arxiv.org/abs/2007.05758

Landscape of R packages for eXplainable Artificial Intelligence by Szymon Maksymiuk, Alicja Gosiewska, Przemysław Biecek https://arxiv.org/pdf/2009.13248.pdf

Feature Engineering and Selection: A Practical Approach for Predictive Models https://bookdown.org/max/FES/

Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials https://www.ncbi.nlm.nih.gov/pmc/articles/PMC300808/

xgboost.surv https://github.com/bcjaeger/xgboost.surv

DoubleML The Python and R package DoubleML provide an implementation of the double / debiased machine learning framework of Chernozhukov et al. (2018). The Python package is built on top of scikit-learn (Pedregosa et al., 2011) and the R package on top of mlr3 and the mlr3 ecosystem (Lang et al., 2019). https://docs.doubleml.org/stable/index.html

Preplication, Replication: A Proposal to Efficiently Upgrade Journal Replication Standards Get access Arrow Michael Colaresi https://academic.oup.com/isp/article-abstract/17/4/367/2528282?redirectedFrom=fulltext

https://deepmind.com/blog/article/using-jax-to-accelerate-our-research

https://github.com/tidyverts/fable

The Effect: An Introduction to Research Design and Causality https://theeffectbook.net/

https://github.com/dedupeio/dedupe

https://arxiv.org/abs/2205.07407 What GPT Knows About Who is Who Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, Chris Tanner

An Introduction to Ontology Engineering https://people.cs.uct.ac.za/~mkeet/files/OEbook.pdf

R Packages for Item Response Theory Analysis: Descriptions and Features https://www.tandfonline.com/doi/ful

Accuracy vs Explainability of Machine Learning Models [NIPS workshop poster review] https://www.inference.vc/accuracy-vs-explainability-in-machine-learning-models-nips-workshop-poster-review/

https://arxiv-sanity-lite.com/

Attitudes toward amalgamating evidence in statistics∗ Andrew Gelman† Keith O'Rourke‡ http://www.stat.columbia.edu/~gelman/research/unpublished/Amalgamating6.pdf

An overview of gradient descent optimization algorithms https://ruder.io/optimizing-gradient-descent/

https://codeocean.com/

ClustGeo: an R package for hierarchical clustering with spatial constraints https://arxiv.org/pdf/1707.03897.pdf

An Algorithmic Framework for Bias Bounties Ira Globus-Harris, Michael Kearns, Aaron Roth https://arxiv.org/abs/2201.10408

On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis https://arxiv.org/pdf/1707.01780.pdf

Fast TreeSHAP: Accelerating SHAP Value Computation for Trees Jilei Yang https://arxiv.org/abs/2109.09847

Comparing interpretability and explainability for feature selection Jack Dunn, Luca Mingardi, Ying Daisy Zhuo https://arxiv.org/abs/2105.05328

Training Deep Nets with Sublinear Memory Cost Tianqi Chen, Bing Xu, Chiyuan Zhang, Carlos Guestrin https://arxiv.org/abs/1604.06174

ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R https://arxiv.org/pdf/1508.04409.pdf

A Survey of Recent Abstract Summarization Techniques Diyah Puspitaningrum https://arxiv.org/abs/2105.0082

U N D E R S TA N D I N G R A N D O M F O R E S T S from theory to practice https://arxiv.org/pdf/1407.7502.pdf

Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology https://arxiv.org/pdf/1809.03006.pdf

Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO Ray Bai, Veronika Rockova, Edward I. George https://arxiv.org/abs/2010.06451

Representation Tradeoffs for Hyperbolic Embeddings Christopher De Sa‡ Albert Gu† Christopher Re´ † Frederic Sala† https://arxiv.org/pdf/1804.03329.pdf

Ratios: A short guide to confidence limits and proper use V.H. Franz∗ October, 2007 https://arxiv.org/pdf/0710.2024.pdf

The Endogeneity of Historical Data Posted on August 28, 2020 by Adam Slez https://broadstreet.blog/2020/08/2 endogeneity-of-historical-data/

A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251194

Post model-fitting exploration via a "Next-Door" analysis Leying GUAN1* and Robert TIBSHIRANI2 https://tibshirani.su.domains/ftp/nextDoor.pdf

Understanding BERT Transformer: Attention isn't all you need A parsing/composition framework for understanding Transformers https://medium.com/synapse-dev/understanding-bert-transformer-attention-isnt-all-you-need-5839ebd396db

Einstein VI: General and Integrated Stein Variational Inference in NumPyro Ahmad Salim Al-Sibahi, Ola Rønning, Christophe Ley, Thomas Wim Hamelryck https://openreview.net/forum?id=nXSDybDWV

Dream Investigation Results Official Report by the Minecraft Speedrunning Team https://mcspeedrun.com/dream.pdf

Improving Parameter Estimation of Epidemic Models: Likelihood Functions and Kalman Filtering 39 Pages Posted: 8 Aug 2022 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4165188

Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment Published online by Cambridge University Press: 09 March 2021 https://www.cambridge.org/core/journals/political-analysis/article/abs/do-namebased-treatments-violate-information-equivalence-evidence-from-a-correspondence-audit-experiment/56C6846518DDADE6EAF92DAE11552BDF

How Much Should We Trust Staggered Difference-In-Differences Estimates? European Corporate Governance Institute – Finance Working Paper No. 736/2021 Rock Center for Corporate Governance at Stanford University Working Paper No. 246 Journal of Financial Economics (JFE), Forthcoming https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3794018

Building useful models for industry—some tips Jim Savage January 2017 https://khakieconomics.github.io/2017/ useful-models-for-industry.html

An Introduction to Proximal Causal Learning https://arxiv.org/pdf/2009.10982.pdf

First Things First: Assessing Data Quality before Model Quality Anita Gohdes and Megan Price meganp@benetech.orgView all authors and affiliations https://journals.sagepub.com/doi/full/10.1177/0022 kmn9p94f4BFh60b0eH_PE

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans https://www.nature.com/articles/s42256-021-00307-0

Why and How We Should Join the Shift From Significance Testing to Estimation https://www.preprints.org/manuscript/202112.0235/v1

How to make replication the norm https://www.nature.com/articles/d41586-018-02108-9

Applied Bayesian Statistics Using Stan and R https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/bayesian-statistics/

https://seeing-theory.brown.edu/index.html

https://www.brodrigues.co/

FINDING ECONOMIC ARTICLES WITH DATA AND SPECIFIC EMPIRICAL METHODS http://skranz.github.io//r/2021/01/05/FindingEconomicArticles4.html

Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145

Machine vision on historical maps https://weinman.cs.grinnell.edu/research/maps.shtml

Enhancing Validity in Observational Settings When Replication Is Not Possible https://papers.ssrn.com/sol3/pa

1.1 Billion Taxi Rides with SQLite, Parquet & HDFS https://tech.marksblogg.com/billion-nyc-taxi-rides-sqlite-parquet-hdfs.html

Understanding the Bias-Variance Tradeoff http://scott.fortmann-roe.com/docs/BiasVariance.html

Is the LKJ(1) prior uniform? "Yes" http://srmart.in/is-the-lkj1-prior-uniform-yes/

Informative priors for correlation matrices: An easy approach http://srmart.in/informative-priors-for-correlation-matrices-an-easy-approach/

A Tutorial on Spectral Clustering https://arxiv.org/pdf/0711.0189v1.pdf

Automated Geocoding of Textual Documents: A Survey of Current Approaches https://onlinelibrary.wiley.com/

Sparklyr https://spark.rstudio.com/

The AAA Tranche of Subprime Science Andrew Gelman and Eric Loken http://www.stat.columbia.edu/~gelman

Never trust rownames of a dataframe June 16th, 2015 by Ankur Gupta | https://www.perfectlyrandom.org/2015/trust-the-row-names-of-a-dataframe-in-R/

GRAPH ALGORITHMS http://www.martinbroadhurst.com/tag/igraph

Groundhog: Addressing The Threat That R Poses To Reproducible Research http://datacolada.org/95

CS231n Convolutional Neural Networks for Visual Recognition https://cs231n.github.io/neural-networks-3/

Implementing Variational Autoencoders in Keras: Beyond the Quickstart Tutorial http://louistiao.me/posts/implementing-variational-autoencoders-in-keras-beyond-the-quickstart-tutorial/

Hypothesis Testing in Econometrics http://home.uchicago.edu/amshaikh/webfiles/testingreview.pdf

"Why Should I Trust You?" Explaining the Predictions of Any Classifier https://arxiv.org/pdf/1602.04938v3.pdf

Yes, but Did It Work?: Evaluating Variational Inference http://www.stat.columbia.edu/~gelman/research/publi
https://statmodeling.stat.columbia.edu/2018/06/27/yes-work-evaluating-variational-inference/

Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets https://arxiv.org/abs/2103.12028

One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV Joshua Angrist, Michal Kolesár https://arxiv.org/abs/2110.10556

Underspecification Presents Challenges for Credibility in Modern Machine Learning https://arxiv.org/abs/2011.03395

A Survey of Predictive Modelling under Imbalanced Distributions https://arxiv.org/pdf/1505.01658.pdf

Varying Slopes Models and the CholeskyLKJ distribution in TensorFlow Probability https://adamhaber.github.io/post/varying-slopes/

Shapley Decomposition of R-Squared in Machine Learning Models https://arxiv.org/pdf/1908.09718.pdf

Understanding Global Feature Contributions With Additive Importance Measures Ian Covert, Scott Lundberg, Su-In Lee https://arxiv.org/abs/2004.00668

True to the Model or True to the Data? https://arxiv.org/abs/2006.16234

When to Impute? Imputation before and during cross-validation Byron C. Jaeger*1 | Nicholas J. Tierney2 | Noah R. Simon3 https://arxiv.org/pdf/2010.00718.pdf

A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications Hongyun Cai, Vincent W. Zheng, Kevin Chen-Chuan Chang https://arxiv.org/abs/1709.07604

Comparing methods addressing multi-collinearity when developing prediction models https://arxiv.org/abs/2101.01603

Nonparametric causal effects based on incremental propensity score interventions https://arxiv.org/abs/1704.00211

Deep learning generalizes because the parameter-function map is biased towards simple functions Guillermo Valle-Pérez, Chico Q. Camargo, Ard A. Louis https://arxiv.org/abs/1805.08522

Bayesian Item Response Modeling in R with brms and Stan https://arxiv.org/pdf/1905.09501.pdf

Bayesian Inference for a Covariance Matrix https://arxiv.org/pdf/1408.4050.pdf

Cross-validation Confidence Intervals for Test Error Pierre Bayle, Alexandre Bayle, Lucas Janson, Lester Mackey https://arxiv.org/abs/2007.12671

Comparing Published Scientific Journal Articles to Their Pre-print Versions https://arxiv.org/pdf/1604.05363.pd

End-to-End Weak Supervision Salva Rühling Cachay, Benedikt Boecking, Artur Dubrawski https://arxiv.org/abs/2107.02233

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests∗ https://arxiv.org/pdf/1510.04342.pdf

Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift https://arxiv.org/pdf/1801.05134.pdf

A review on outlier/anomaly detection in time series data https://arxiv.org/abs/2002.04236

Entropic Out-of-Distribution Detection: Seamless Detection of Unknown Examples David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano L. I. Oliveira, Teresa Ludermir https://arxiv.org/abs/2006.04005

An Exploratory Characterization of Bugs in COVID-19 Software Projects Akond Rahman, Effat Farhana https://arxiv.org/abs/2006.00586

Be Careful What You Backpropagate: A Case For Linear Output Activations & Gradient Boosting Anders Oland, Aayush Bansal, Roger B. Dannenberg, Bhiksha Raj https://arxiv.org/abs/1707.04199

Introducing Stan2tfp - a lightweight interface for the Stan-to-TensorFlow Probability compiler May 21, 2020 4 min read https://adamhaber.github.io/post/stan2tfp-post1/

L2 Regularization versus Batch and Weight Normalization Twan van Laarhoven https://arxiv.org/abs/1706.0535

Unsupervised Discovery of Temporal Structure in Noisy Data with Dynamical Components Analysis David G. Clark, Jesse A. Livezey, Kristofer E. Bouchard https://arxiv.org/abs/1905.09944

Monte Carlo Gradient Estimation in Machine Learning Shakir Mohamed, Mihaela Rosca, Michael Figurnov, Andriy Mnih https://arxiv.org/abs/1906.10652

Large-scale linear regression: Development of high-performance routines Alvaro Frank, Diego Fabregat-Traver, Paolo Bientinesi https://arxiv.org/abs/1504.07890

The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions Raj Agrawal, Jonathan H. Huggins, Brian Trippe, Tamara Broderick https://arxiv.org/abs/1905.06501

TensorFlow Distributions Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, Rif A. Saurous https://arxiv.org/abs/1711.10604

Asymptotically Exact, Embarrassingly Parallel MCMC Willie Neiswanger, Chong Wang, Eric Xing https://arxiv.org/abs/1311.4780

Python for Data Science https://aeturrell.github.io/python4DS/welcome.html

Using the flextable R package https://ardata-fr.github.io/flextable-book/

Coding for Economists https://aeturrell.github.io/coding-for-economists/intro.html

When Should You Adjust Standard Errors for Clustering? Get access Arrow Alberto Abadie, Susan Athey, Guido W Imbens, Jeffrey M Wooldridge https://academic.oup.com/qje/advance-article-abstract/doi/10.1093/qje/qjac038/6750017

Awesome Deep Learning for Natural Language Processing (NLP) https://github.com/brianspiering/awesome-dl4nlp

R for applied epidemiology and public health https://epirhandbook.com/en/index.html

COVID 19: Reduced forms have gone viral, but what do they tell us?* https://drive.google.com/file/d/1ERjcGX

Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology https://elifesciences.org/articles/67995

Taking Uncertainty Seriously: Bayesian Marginal Structural Models for Causal Inference in Political Science https://github.com/ajnafa/Latent-Bayesian-MSM

Generalized Linear Models https://data.princeton.edu/wws509/notes/c7s4

genieclust: Fast and Robust Hierarchical Clustering with Noise Point Detection https://genieclust.gagolewski.com/

Awesome Graph Classification https://github.com/benedekrozemberczki/awesome-graph-classification

parallelDist https://github.com/alexeckert/parallelDist

Interpretable Machine Learning A Guide for Making Black Box Models Explainable Christoph Molnar https://christophm.github.io/interpretable-ml-book/

The Inverse CDF Method https://dk81.github.io/dkmathstats_site/prob-inverse-cdf.html

HamiltonianMC https://chi-feng.github.io/mcmc-demo/app.html#HamiltonianMC,banana

End-to-End Balancing for Causal Continuous Treatment-Effect Estimation https://assets.amazon.science/5b/71/
to-end-balancing-for-causal-continuous-treatment-effect-estimation.pdf

A tour of probabilistic programming language APIs What does it look like to do MCMC in different frameworks? https://colcarroll.github.io/ppl-api/

Probabilistic Programming & Bayesian Methods for Hackers https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/

Beyond Multiple Linear Regression Applied Generalized Linear Models and Multilevel Models in R https://bookdown.org/roback/bookdown-bysh/

Maybe a section on hyperparameters?

Does batch size matter? https://blog.janestreet.com/does-batch-size-matter/

The Much Quieter Revolution of Synthetic Control: Episode I https://causalinf.substack.com/p/the-much-quieter-revolution-of-synthetic?utm_campaign=post&utm_medium=web&utm_source=

User-friendly introduction to PAC-Bayes bounds https://arxiv.org/pdf/2110.11216.pdf

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results https://journals.sagepub.com/doi/full/10.1177/2515245917747646

The RecordLinkage Package: Detecting Errors in Data https://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf

https://grow.google/certificates/interview-warmup/

The inverse-transform method for generating random variables in R https://heds.nz/posts/inverse-transform/

The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology https://journals.sagepub.com/doi/10.1177/1948550616673876

Evolution of Reporting P Values in the Biomedical Literature, 1990-2015 https://jamanetwork.com/journals/jam

SHAP (SHapley Additive exPlanations) https://github.com/slundberg/shap

The h-index is no longer an effective correlate of scientific reputation https://journals.plos.org/plosone/article?id=

Prior Choice Recommendations Andrew Gelman edited this page on Apr 17, 2020 · 51 revisions https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations#prior-for-a-covariance-matrix

Institute for Replication (I4R) https://i4replication.org/index.html

How Life Sciences Actually Work: Findings of a Year-Long Investigation https://guzey.com/how-life-sciences-actually-work/

Efficient Neural Causal Discovery without Acyclicity Constraints https://github.com/phlippe/ENCO

awesome-text-summarization https://github.com/mathsyouth/awesome-text-summarization

(Ir)Reproducible Machine Learning: A Case Study https://reproducible.cs.princeton.edu/irreproducibility-paper.pdf

THE MYTH OF THE EXPERT REVIEWER https://parameterfree.com/2021/07/06/the-myth-of-the-expert-reviewer/

Understanding and Choosing the Right Probability Distributions https://onlinelibrary.wiley.com/doi/pdf/10.100

Spatial Interdependence and Instrumental Variable Models https://osf.io/preprints/socarxiv/pgrcu/

The case for formal methodology in scientific reform https://royalsocietypublishing.org/doi/10.1098/rsos.200805

Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies https://papers.ssrn.com/sol3/pap

Pandas Comparison with R / R libraries https://pandas.pydata.org/docs/getting_started/comparison/comparis

Non-Standard Errors https://orbilu.uni.lu/bitstream/10993/48686/1/SSRN-id3961574.pdf

Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors https://pubmed.ncbi.nlm.nih.gov/26186114/

The (lack of) impact of retraction on citation networks Charisse R Madlock-Brown 1, David Eichmann https://pubmed.ncbi.nlm.nih.gov/24668038/

The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature https://psyarxiv.com/pbrdk/

Bayesian Estimation of Correlation Matrices of Longitudinal Data Riddhi Pratim Ghosh, Bani Mallick, Mohsen Pourahmadi https://projecteuclid.org/journals/bayesian-analysis/volume-16/issue-3/Bayesian-Estimation-of-Correlation-Matrices-of-Longitudinal-Data/10.1214/20-BA1237.full

Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals https://osf.io/preprints/socarxiv/cfdba/

How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It https://osf.io/preprints/socarxiv/453jk/

Lost in Aggregation: Improving Event Analysis with Report-Level Data Scott J. Cook,Nils B. Weidmann https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12398

Frequentist versus Bayesian approaches to multiple testing Arvid Sjölander & Stijn Vansteelandt https://link.springer.com/article/10.1007/s10654-019-00517-2

Research note: Examining potential bias in large-scale censored data https://misinforeview.hks.harvard.edu/arti
note-examining-potential-bias-in-large-scale-censored-data/

When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? Kosuke Imai,In Song Kim https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12417

Runtime warnings and convergence problems Stan Development Team https://mc-stan.org/misc/warnings.html

Dirichlet Process Gaussian mixture model via the stick-breaking construction in various PPLs This page was last updated on 29 Mar, 2021. https://luiarthur.github.io/TuringBnpBenchmarks/dpsbgmm

xgboost: "Hi I'm Gamma. What can I do for you?" — and the tuning of regularization https://medium.com/data-design/xgboost-hi-im-gamma-what-can-i-do-for-you-and-the-tuning-of-regularization-a42ea17e6ab6

PostGIS In Action https://livebook.manning.com/book/postgis-in-action-second-edition/about-this-book/

Stan User's Guide https://mc-stan.org/docs/stan-users-guide/index.html

Smoothing Terms in GAM Models https://maths-people.anu.edu.au/~johnm/r-book/xtras/autosmooth.pdf

Designing a Deep Learning Project https://medium.com/(**erogol/designing-a-deep-learning-project-9b3698aef127?**)

PyTorch With Baby Steps: From y=x To Training A Convnet https://lelon.io/blog/pytorch-baby-steps

Bayesian inference with Stan: A tutorial on adding custom distributions Jeffrey Annis, Brent J. Miller & Thomas J. Palmeri https://link.springer.com/article/10.3758/s13428-016-0746-9

Bayes Rules! An Introduction to Applied Bayesian Modeling https://www.bayesrulesbook.com/

Graduate Qualitative Methods Training in Political Science: A Disciplinary Crisis Published online by Cambridge University Press: 21 November 2019 https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/graduate-qualitative-methods-training-in-political-science-a-disciplinary-crisis/7B0EEB76E1CC234AFED7EED8DA71BA35

Time Series Analysis by State Space Methods (Oxford Statistical Science Series) https://www.amazon.com/dp/0198523548/ref=cm_sw_r_tw_apa_fabc_0MWV12PSS3K9NW3RF9ZY

Hyperparameters and tuning strategies for random forest https://wires.onlinelibrary.wiley.com/doi/full/10.1002/

Your Cross Validation Error Confidence Intervals are Wrong — here's how to Fix Them https://towardsdatascience.com/your-cross-validation-error-confidence-intervals-are-wrong-heres-how-to-fix-them-abbfe28d390

Probabilistic Programming with Variational Inference: Under the Hood https://willcrichton.net/notes/probabilis programming-under-the-hood/

How to Measure Statistical Causality: A Transfer Entropy Approach with Financial Applications https://towardsdatascience.com/causality-931372313a1c

Kullback-Leibler Divergence Explained https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance https://www.cs.purdue.edu/homes/lintan/publications/variance-ase20.pdf

How (Not) to Reproduce: Practical Considerations to Improve Research Transparency in Political Science https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/abs/how-not-to-reproduce-practical-considerations-to-improve-research-transparency-in-political-science/32E7CF5D975C081BA666D3BD475D7913

Quantifying Bias from Measurable and Unmeasurable Confounders Across Three Domains of Individual Determinants of Political Preferences Published online by Cambridge University Press: 22 February 2022 https://www.cambridge.org/core/journals/political-analysis/article/quantifying-bias-from-measurable-and-unmeasurable-confounders-across-three-domains-of-individual-determinants-of-political-preferences/D1D2DEE9E7180BDCFC592885BE66E9AF

5 Levels of Difficulty — Bayesian Gaussian Random Walk with PyMC3 and Theano https://towardsdatascience.com/5-levels-of-difficulty-bayesian-gaussian-random-walk-with-pymc3-and-theano-34343911c7d2

Single-Parameter Models | Pyro vs. STAN https://towardsdatascience.com/single-parameter-models-pyro-vs-stan-e7e69b45d95c

Partial Identification in Econometrics Elie Tamer https://scholar.harvard.edu/files/tamer/files/pie.pdf

LightGBM for Quantile Regression Understand Quantile Regression https://towardsdatascience.com/lightgbm-for-quantile-regression-4288d0bb23fd

Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach https://strathprints.strath.ac.uk/59463/1/Gallop_Weschle_PSRM_2016_Assessing_the_impact_o

yardstick is a package to estimate how well models are working using tidy data principles. See the package webpage for more information. https://yardstick.tidymodels.org/index.html

The Three Faces of Bayes https://slackprop.wordpress.com/2016/08/28/the-three-faces-of-bayes/

Evaluating Random Forests for Survival Analysis using Prediction Error Curves https://www.ncbi.nlm.nih.gov/p

The role of metadata in reproducible computational research https://www.sciencedirect.com/science/article/pii/

Ecological Inference in the Social Sciences https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885825/

Two Wrongs Make a Right: Addressing Underreporting in Binary Data from Multiple Sources https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667662/

On the low reproducibility of cancer studies https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6599599/

Quarto with Python https://www.meyerperin.com/using-quarto/

Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015 https://www.nature.com/articles/s41562-018-0399-z

Bayesian analysis of tests with unknown specificity and sensitivity∗ Andrew Gelman† and Bob Carpenter‡ https://www.medrxiv.org/content/10.1101/2020.05.22.20108944v3.full.pdf

Notes on the Negative Binomial Distribution https://www.johndcook.com/negative_binomial.pdf

The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!) https://www.joelonsoftware.com/2003/10/08/the-absolute-minimum-every-software-developer-absolutely-positively-must-know-about-unicode-and-character-sets-no-excuses/

Automatic Differentiation Variational Inference https://www.jmlr.org/papers/volume18/16-107/16-107.pdf

IZA DP No. 13233: The Influence of Hidden Researcher Decisions in Applied Microeconomics https://www.iza.org/publications/dp/13233/the-influence-of-hidden-researcher-decisions-in-applied-microeconomics

cdfquantreg: An R Package for CDF-Quantile Regression https://www.jstatsoft.org/article/view/v088i01

https://techdevguide.withgoogle.com/

What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes David M. W. Powers https://arxiv.org/abs/1503.06410

When LOO and other cross-validation approaches are valid https://statmodeling.stat.columbia.edu/2018/08/03/ cross-validation-approaches-valid/

Hamiltonian Monte Carlo explained http://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.ht

Controlling for Unobserved Confounds in Classification Using Correlational Constraints Virgile Landeiro, Aron Culotta https://arxiv.org/abs/1703.01671

The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies Scott E. Maxwell https://statmodeling.stat.columbia.edu/wp-content/uploads/2017/07/maxwell2004.pdf

You need 16 times the sample size to estimate an interaction than to estimate a main effect https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Machine Learning of Sets http://akosiorek.github.io/ml/2020/08/12/machine_learning_of_sets.html

Weak Supervision: A New Programming Paradigm for Machine Learning http://ai.stanford.edu/blog/weak-supervision/

The earth is flat (p>0.05): Significance thresholds and the crisis of unreplicable research https://peerj.com/preprints/2921/

Advanced Natural Language Processing with TensorFlow 2: Build effective real-world NLP applications using NER, RNNs, seq2seq models, Transformers, and more https://www.amazon.com/Advanced-Natural-Language-Processing-TensorFlow/dp/1800200935?*encoding=UT*
20&linkId=4448e1a0cd126f52a2aba844c4bdb78e&language=en_US&ref=as_li_ss_tl

3 reasons why you can't always use predictive performance to choose among models https://statmodeling.stat.columbia.edu/2015/10/23/26857/

Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models Arthur Lewbel https://www.tandfonline.com/doi/full/10.1080/07350015.2012.643126

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift Sergey Ioffe, Christian Szegedy https://arxiv.org/abs/1502.03167

Gradient Boosting explained [demonstration] http://arogozhnikov.github.io/2016/06/24/gradient_boosting_ex

Clustered standard errors vs. multilevel modeling https://statmodeling.stat.columbia.edu/2007/11/28/clustered_

Advanced R https://adv-r.hadley.nz/index.html

Regression to the mean continues to confuse people and lead to errors in published research https://statmodeling.stat.columbia.edu/2018/06/24/regression-mean-continues-confuse-people-lead-errors-published-research/

The statistical significance filter leads to overoptimistic expectations of replicability https://statmodeling.stat.columbia.edu/2018/05/22/statistical-significance-filter-leads-overoptimistic-expectations-replicability/

How to cross-validate PCA, clustering, and matrix decomposition models http://alexhwilliams.info/itsneuronalbl

Inference in Experiments Conditional on Observed Imbalances in Covariates Per JohanssonORCID Icon &Mattias Nordin https://www.tandfonline.com/doi/full/10.1080/00031305.2022.2054859

Scientific progress despite irreproducibility: A seeming paradox Richard M. Shiffrin, Katy Borner, Stephen M. Stigler https://arxiv.org/abs/1710.01946

On Statistical Non-Significance Alberto Abadie https://arxiv.org/abs/1803.00609

On the number of signals in multivariate time series Markus Matilainen, Klaus Nordhausen, Joni Virta https://arxiv.org/abs/1801.04925

Data Science vs. Statistics: Two Cultures? Iain Carmichael, J.S. Marron https://arxiv.org/abs/1801.00371

The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions George Philipp, Dawn Song, Jaime G. Carbonell https://arxiv.org/abs/1712.05577

Theory of Deep Learning III: explaining the non-overfitting puzzle Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, Hrushikesh Mhaskar https://arxiv.org/abs/1801.00173

On overfitting and post-selection uncertainty assessments Liang Hong, Todd A. Kuffner, Ryan Martin https://arxiv.org/abs/1712.02379

A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results Beau Coker, Cynthia Rudin, Gary King https://arxiv.org/abs/1804.08646

Labelling as an unsupervised learning problem Terry Lyons, Imanol Perez Arribas https://arxiv.org/abs/1805.03911

Structural Breaks in Time Series Alessandro Casini, Pierre Perron https://arxiv.org/abs/1805.03807

On consistency and inconsistency of nonparametric tests Mikhail Ermakov https://arxiv.org/abs/1807.09076

A New Angle on L2 Regularization Thomas Tanay, Lewis D Griffin https://arxiv.org/abs/1806.11186

On the Robustness of Interpretability Methods David Alvarez-Melis, Tommi S. Jaakkola https://arxiv.org/abs/1806.08049

Identifying Causal Effects with the R Package causaleffect Santtu Tikka, Juha Karvanen https://arxiv.org/abs/1806.07161

How Does Batch Normalization Help Optimization? Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Madry https://arxiv.org/abs/1805.11604

The effect of the choice of neural network depth and breadth on the size of its hypothesis space Lech Szymanski, Brendan McCane, Michael Albert https://arxiv.org/abs/1806.02460

Is preprocessing of text really worth your time for online comment classification? Fahim Mohammad https://arxiv.org/abs/1806.02908

Geometric Understanding of Deep Learning Na Lei, Zhongxuan Luo, Shing-Tung Yau, David Xianfeng Gu https://arxiv.org/abs/1805.10451

Cross validation residuals for generalised least squares and other correlated data models Ingrid Annette Baade https://arxiv.org/abs/1809.01319

Out-of-Distribution Detection Using an Ensemble of Self Supervised Leave-out Classifiers Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, Theodore L. Willke https://arxiv.org/abs/1809.03576

Handling Imbalanced Dataset in Multi-label Text Categorization using Bagging and Adaptive Boosting Genta Indra Winata, Masayu Leylia Khodra https://arxiv.org/abs/1810.11612

On the Art and Science of Machine Learning Explanations Patrick Hall https://arxiv.org/abs/1810.02909

Causal inference under over-simplified longitudinal causal models Lola Etievant, Vivian Viallon https://arxiv.org/abs/1810.01294

Revisiting the Gelman-Rubin Diagnostic Dootika Vats, Christina Knudson https://arxiv.org/abs/1812.09384

A Survey on Data Collection for Machine Learning: a Big Data – AI Integration Perspective Yuji Roh, Geon Heo, Steven Euijong Whang

A Fundamental Measure of Treatment Effect Heterogeneity Jonathan Levy, Mark van der Laan, Alan Hubbard, Romain Pirracchio https://arxiv.org/abs/1811.03745

Causal Discovery Toolbox: Uncover causal relationships in Python Diviyan Kalainathan, Olivier Goudet https://arxiv.org/abs/1903.02278

Dying ReLU and Initialization: Theory and Numerical Examples Lu Lu, Yeonjong Shin, Yanhui Su, George Em Karniadakis https://arxiv.org/abs/1903.06733

ROC and AUC with a Binary Predictor: a Potentially Misleading Metric John Muschelli https://arxiv.org/abs/1903.04881

Gamification in Science: A Study of Requirements in the Context of Reproducible Research Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, Albrecht Schmidt https://arxiv.org/abs/1903.02446

Matrix factorization for multivariate time series analysis Pierre Alquier, Nicolas Marie https://arxiv.org/abs/1903.05589

On the complexity of logistic regression models Nicola Bulso, Matteo Marsili, Yasser Roudi https://arxiv.org/abs/1903.00386

Part II

Presentation

4 Markdown

R Markdown Cookbook

Part III

Computation

5 Computation

5.1 git

https://git-scm.com/doc

6 R

https://www.r-project.org/other-docs.html

Hands-On Programming with R, Garrett Grolemund

6.0.1 Tidyverse

https://www.tidyverse.org/

R for Data Science

The Tidyverse Cookbook

7 Python

https://docs.python.org/3/

7.0.1 Numpy

https://numpy.org/

7.0.2 Pandas

https://pandas.pydata.org/docs/

Effective Pandas https://store.metasnake.com/effective-pandas-book

8 jax

https://github.com/google/jax

9 Numpyro

https://github.com/pyro-ppl/numpyro

10 Stan

https://mc-stan.org/users/documentation/

10.1 brms

https://github.com/paul-buerkner/brms

11 pyro

https://pyro.ai/

The StatQuest Introduction to PyTorch

12 tensorflow

https://www.tensorflow.org/

13 SQL

postgresql https://www.postgresql.org/docs/

Part IV

Data management

14 Filter

Instance of: Higher-order function

AKA: Subset

Distinct from:

English:

Formalization:

Cites: Wikipedia ; Wikidata

Code

Base

subset: Subsetting Vectors, Matrices and Data Frames

Dplyr

Subset rows using column values

DataTable

Subsetting Rows

14.0.1 Python

Documentation: numpy.mean

Examples:

14.0.2 SQL

Documentation: PostgreSQL AVG Function

```r
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```r
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

14.0.3 Torch

```python
import torch
```

15 Joins

16 Regex

R Regular expressions

17 Fuzzy Recording Matching

Name Match

Part V

Domain

18 Domain

CHANNELLING FISHER: RANDOMIZATION TESTS AND THE STATISTICAL INSIGNIFICANCE OF SEEMINGLY SIGNIFICANT EXPERIMENTAL RESULTS

An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?

Outlier

19 Outliers

An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?

20

21

Part VI

Research Design

22

23 Unit of Analysis

Ecological Inference in the Social Sciences https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885825/

Race to the bottom: Spatial aggregation and event data Scott J. CookORCID Icon&Nils B. Weidmann https://www.tandfonline.com/doi/abs/10.1080/03050629.2022.2025365

Extremal Behavior of Aggregated Data with an Application to Downscaling Sebastian Engelke, Raphael de Fondeville, Marco Oesting https://arxiv.org/abs/1712.09816

24 Estimand

What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory

25 Identification

Partial Identification in Econometrics

26 Garden of Forking Paths

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time∗, Andrew Gelman† and Eric Loken

Achieving Statistical Significance with Covariates and without Transparency

27 Random Control Trials

Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015

CHANNELLING FISHER: RANDOMIZATION TESTS AND THE STATISTICAL INSIGNIFICANCE OF SEEMINGLY SIGNIFICANT EXPERIMENTAL RESULTS

Warnings

- Dicing RCT results up by coverates or in with a regression model instead of doing a simple T test can generate spurious results from a few high leverage outlier observations (Young 2019).

28 Instrumental Variables

How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on Over 60 Replicated Studies

29 Difference in Difference

How Much Should We Trust Differences-In-Differences Estimates?

How Much Should We Trust Staggered Difference-In-Differences Estimates?

When Is Parallel Trends Sensitive to Functional Form?∗

30 Bias Variance Tradeoff

31 Placebo Tests

Do We Really Know the WTO Cures Cancer? https://www.cambridge.org/core/journals/british-journal-of-political-science/article/abs/do-we-really-know-the-wto-cures-cancer/B84A6FCF516FAE3ED7E0C20

Part VII

Estimation

32 Performance

33 Out of Sample Performance

Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure David R. Roberts,Volker Bahn,Simone Ciuti,Mark S. Boyce,Jane Elith,Gurutzeta Guillera-Arroita,Severin Hauenstein,José J. Lahoz-Monfort,Boris Schröder,Wilfried Thuiller,David I. Warton,Brendan A. Wintle,Florian Hartig,Carsten F. Dormann https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.02881

Cross-validation FAQ, Aki Vehtari

When to Impute? Imputation before and during cross-validation

Leakage and the Reproducibility Crisis in ML-based Science

Rescaling and other forms of unsupervised preprocessing introduce bias into cross-validation,Amit Moscovich, Saharon Rosset

Approximate leave-future-out cross-validation for Bayesian time series models

groupdata2

How Cross-Validation Can Go Wrong and What to Do About it.

Moving cross-validation from a research idea to a routine step in Bayesian data analysis

Model selection tutorials and talks, Aki Vehtari Underspecification Presents Challenges for Credibility in Modern Machine Learning (Paper Explained) -Has a neet example of holding out on camera model shows massive degredation in medical mission vision application.

Underspecification Presents Challenges for Credibility in Modern Machine Learning

Cross-validation: what does it estimate and how well does it do it? Stephen Bates, Trevor Hastie, Robert Tibshirani

Consensus features nested cross-validation

Your Cross Validation Error Confidence Intervals are Wrong — here's how to Fix Them

34 Regularization

Ridge Regression Can Produce Misleading Inferences in the Presence of Strong Confounders: The Case of Mass Polarization

Fast Penalized Regression and Cross Validation for Tall Data with the oem Package

35 P Values

Bayesian estimation supersedes the t test

Sequential sampling and testing Safe, anytime-valid inference: confidence sequences, p-values/e-values, and e-processes

Some papers about p values

Part VIII

Mathematical Objects

36 Set

Cites: Wikipedia; Wikidata; PlanetMath

37 List (Sequence)

AKA: Sequence, a_n where n is the nth element, (1,2,3, ….)

Distinct from: Set

Measure of:

Description: A list is a collection of objects with a specific ordering and where the same object can appear more than once. Call each object an element, and its location its index or rank. An index is a natural number counting upward from the first element in the list. Whether counting begins at 0 or 1 depends on local conventions.

Formalization:

Algorithm:

Cites: Wikipedia Wikidata Encyclopedia Of Math Wolfram PlanetMath

37.0.0.1 R

Documentation:

list: Lists – Generic and Dotted Pairs

Examples:

```
example_list = list(1,2,3)
example_list
```

```
[[1]]
[1] 1

[[2]]
[1] 2

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

37.0.0.2 Python

Documentation:

More on Lists

Examples:

```
example_list =  [1,2,3]
example_list
```

```
[1, 2, 3]
```

37.0.0.3 SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
dbListTables(con)
```

```
character(0)
```

```
dbWriteTable(con, "mtcars", mtcars)
dbListTables(con)
```

```
[1] "mtcars"
```

```
create table StatisticalNumbers(
 value int
)
```

```
SELECT * FROM mtcars LIMIT 5;
```

Table 37.1: 5 records

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
21.0	6	160	110	3.90	2.875	17.02	0	1	4	4

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

38 Vector/Matrix/Tensor

Instance of: algebraic object / data structure

AKA: array, matrices

Distinct from: list

English: Vectors, matrices, and tensors are like lists in that they are a collection of objects which are indexed. They differ in that the index can be multi-dimensional, where vectors are 1-d indexed, matrices are 2-d indexed, and tensors are m-d indexed. They also are typically constrained to have objects that share the same type, e.g. numbers or strings.

Formalization:

Cites:

Array:

Wikipedia

3Blue1Brown: Vectors | Chapter 1, Essence of linear algebra 3Blue1Brown: Linear combinations, span, and basis vectors | Chapter 2, Essence of linear algebra

Matrix:

Wikipedia

3Blue1Brown: Linear transformations and matrices | Chapter 3, Essence of linear algebra

Tensor:

Wikipedia

Code

Vector

Note unlike matrix and array, the basic vector function initializes an empty vector and you have to actually use as.vector to coerce something else to vector as the constructor.

vector: Vectors

```
example_vector <- as.vector(c(1,2,3,4))
class(example_vector)
```

```
[1] "numeric"
```

```
example_vector
```

```
[1] 1 2 3 4
```

Matrix

Note we can choose which direction to fill the matrix with, either by row1-col1, row1-col2, row1-col3, row1-col4

matrix: Matrices

```
example_matrix <- matrix(c(1,2,3,4,"A","B","C","D"), nrow = 2, ncol = 4, byrow = TRUE,
               dimnames = list(c("row1", "row2"),
                               c("C.1", "C.2", "C.3", "C.4")))
class(example_matrix)
```

```
[1] "matrix" "array"
```

```
example_matrix
```

```
     C.1 C.2 C.3 C.4
row1 "1" "2" "3" "4"
row2 "A" "B" "C" "D"
```

Arrays

Note array dimensions are ordered, row, column, depth, …, M , and elements are filled row1-col1-depth1, row2-col1-depth1, row1-col2-depth1,… and so on. Note this was coerced to a string because any of the elements were a string.

array: Multi-way Arrays

```
example_tensor= array(c(1,2,3,4,"A","B","C","D","+","-","*","/"),dim=c(2,3,2,2))
class(example_tensor)
```

```
[1] "array"
```

```r
example_tensor
```

```
, , 1, 1

     [,1] [,2] [,3]
[1,] "1"  "3"  "A" 
[2,] "2"  "4"  "B" 

, , 2, 1

     [,1] [,2] [,3]
[1,] "C"  "+"  "*" 
[2,] "D"  "-"  "/" 

, , 1, 2

     [,1] [,2] [,3]
[1,] "1"  "3"  "A" 
[2,] "2"  "4"  "B" 

, , 2, 2

     [,1] [,2] [,3]
[1,] "C"  "+"  "*" 
[2,] "D"  "-"  "/" 
```

38.0.0.1 Python

Documentation:

Examples:

38.0.0.2 SQL

Documentation:

```r
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
```

```
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

38.0.0.3 Jax

38.0.0.4 Torch

```
import torch
```

39 Table

Instance of: arrangement of information or data

AKA: Dataframe

Distinct from:

English: A collection of rows and columns, where rows represent specific instances (AKA records, k-tuple, n-tuple, or a vector), and columns represent features (AKA variables, parameters, properties, attributes, or stanchions). The intersection of a row and column is called a sell.

Formalization:

Cites: Wikipedia Table (information) ; Wikipedia Table Table (database) ; Wikidata ; Wolfram

ML Frameworks Interoperability Cheat Sheet

Code

39.0.0.1 R

Documentation: data.frame: Data Frames

Examples:

```
df=data.frame(a=c(1,2,3,4), b=c('a','b','c','d'))
df
```

```
  a b
1 1 a
2 2 b
3 3 c
4 4 d
```

39.0.0.2 Python

Documentation: pandas.DataFrame

Examples:

```
import pandas as pd
df = pd.DataFrame({'a': [1, 2,3,4], 'b': ['a','b','c','d']})
df
```

```
   a  b
0  1  a
1  2  b
2  3  c
3  4  d
```

39.0.0.3 SQL

Documentation: PostgreSQL AVG Function

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
```

```r
con <- dbConnect(RPostgres::Postgres())
```

```sql
DROP TABLE IF EXISTS df;
```

```sql
CREATE TABLE IF NOT EXISTS df (
    a INTEGER,
    b CHAR
);
```

```sql
INSERT INTO df (a, b)
VALUES
    (1,'a'),
    (2,'b'),
    (3,'c'),
    (4,'d');
```

```sql
SELECT  * FROM  df;
```

Table 39.1: 4 records

a	b
1	a
2	b
3	c
4	d

39.0.0.4 Torch

```python
import torch
```

Part IX

Operations of Arithmetic

40 Addition

Instance of: operation of arithmetic

40.1 Frequentist

AKA: + ; add

Distinct from:

English:

Formalization:

Cites: Wikipedia ; Wikidata ; Wolfram

Code

40.1.0.1 R

Documentation: mean: Arithmetic Mean

Examples:

40.1.0.2 Python

Documentation: numpy.mean

Examples:

40.1.0.3 SQL

Documentation: PostgreSQL AVG Function

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

40.1.0.4 Torch

```
import torch
```

40.2 Bayesian

English: **Formalization**:

Cites:

Code

41 Introduction

Instance of: operation of arithmetic

41.1 Frequentist

AKA: - ; minus

Distinct from:

English:

Formalization:

Cites: Wikipedia ; Wikidata ; Wolfram

Code

41.1.0.1 R

Documentation: mean: Arithmetic Mean

Examples:

41.1.0.2 Python

Documentation: numpy.mean

Examples:

41.1.0.3 SQL

Documentation: PostgreSQL AVG Function

```r
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```r
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

41.1.0.4 Torch

```python
import torch
```

41.2 Bayesian

English: **Formalization**:

Cites:

Code

42 Multiplication

Instance of: operation of arithmetic

42.1 Frequentist

AKA: * ; $\times$; ; multiply

Distinct from:

English:

Formalization:

Cites: Wikipedia ; Wikidata ; Wolfram

3Blue1Brown: Matrix multiplication as composition | Chapter 4, Essence of linear algebra
3Blue1Brown: Cross products in the light of linear transformations | Chapter 11, Essence of linear algebra

Code

42.1.0.1 R

Documentation: mean: Arithmetic Mean

Examples:

42.1.0.2 Python

Documentation: numpy.mean

Examples:

42.1.0.3 SQL

Documentation: PostgreSQL AVG Function

```r
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```r
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

42.1.0.4 Torch

```python
import torch
```

42.2 Bayesian

English: **Formalization**:

Cites:

Code

43 Division

Instance of: operation of arithmetic

43.1 Frequentist

AKA: / ; $\frac{numerator}{denominator}$; ÷

Distinct from:

English:

Formalization:

Cites: Wikipedia ; Wikidata ; Wolfram

Code

43.1.0.1 R

Documentation: mean: Arithmetic Mean

Examples:

43.1.0.2 Python

Documentation: numpy.mean

Examples:

43.1.0.3 SQL

Documentation: PostgreSQL AVG Function

```r
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```r
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

43.1.0.4 Torch

```python
import torch
```

43.2 Bayesian

English: **Formalization**:

Cites:

Code

Part X

Operations of Algebra

44 Dot product

Instance of: algebraic operation

AKA: scalar product; inner product ; projection product ; $ · $

Distinct from:

English:

Formalization:

$$a \cdot b$$

Cites: Wikipedia ; Wikidata ; Wolfram

3Blue1Brown: Dot products and duality | Chapter 9, Essence of linear algebra

Code

44.0.0.1 R

Documentation:

Examples:

44.0.0.2 Python

Documentation: numpy.mean

Examples:

44.0.0.3 SQL

Documentation: PostgreSQL AVG Function

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
```

```
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

44.0.0.4 Torch

```
import torch
```

44.1 Bayesian

English: **Formalization**:

Cites:

Code

Part XI

Moments of a Distribution

45 Mean

Measure of: Central tendency

45.1 Frequentist

AKA: Arithmetic mean; average; $\bar{x}$ (sample mean); μ (population mean); μ_x (population mean)

Distinct from: Geometric mean (GM); Harmonic mean (HM); generalized mean/ Power mean; weighted arithmetic mean

English: Take a list of numbers, sum those numbers, and then divide by the number of numbers.

Formalization:

$$\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} x_i) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Cites: Wikipedia ; Wikidata ; Wolfram

Code

45.1.0.1 R

Documentation: mean: Arithmetic Mean

Examples:

```
x = c(1,2,3,4)
x
```

```
[1] 1 2 3 4
```

```
#Algorithm
x_bar = sum(x, na.rm=T)/length(x)
x_bar
```

```
[1] 2.5
```

```
#Base Function
x_bar = mean(x, na.rm=T)
x_bar
```

```
[1] 2.5
```

45.1.0.2 Python

Documentation: numpy.mean

Examples:

```
x = [1,2,3,4]
print(x)
```

```
[1, 2, 3, 4]
```

```
#Algorithm
x_bar= sum(x)/len(x)
x_bar
```

```
2.5
```

```
#statistics Function
import statistics
x_bar = statistics.mean(x)
x_bar
```

```
2.5
```

```
#scipy Function
#<string>:1: DeprecationWarning: scipy.mean is deprecated and will be removed in SciPy 2.0
import scipy
x_bar = scipy.mean(x)
```

```
<string>:1: DeprecationWarning: scipy.mean is deprecated and will be removed in SciPy 2.0.0,
```

```
x_bar
```

```
2.5
```

```
#numpy Function
import numpy as np
x = np.array(x)
x_bar = x.mean()
x_bar
```

```
2.5
```

45.1.0.3 SQL

Documentation: PostgreSQL AVG Function

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

```
Loading required package: RPostgres
```

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched m
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-0
con <- dbConnect(RPostgres::Postgres())
```

```
DROP TABLE IF EXISTS t1;
```

```
CREATE TABLE IF NOT EXISTS t1 (
    id serial PRIMARY KEY,
    amount INTEGER
);
```

```
INSERT INTO t1 (amount)
VALUES
    (10),
    (NULL),
    (30);
```

```
SELECT  * FROM  t1;
```

Table 45.1: 3 records

id	amount
1	10
2	NA
3	30

```
SELECT AVG(amount)::numeric(10,2)
FROM t1;
```

Table 45.2: 1 records

avg
20

45.1.0.4 Torch

```
import torch
```

45.2 Bayesian

Bayesian average; Solving an age-old problem using Bayesian Average; Of bayesian average and star ratings; Bayesian Average Ratings ;

English: The Bayesian average is the weighted average of a prior and the observed sample average. When would you want this? When you have strong beliefs about the true mean, or when sample size is too small to reliable calculate a mean. For example a movie rating website where a movie may have only a single 5 star rating and so would rank higher than the Godfather with over a 100 almost all 5 star ratings.

Formalization:

$$\bar{x} = \frac{C * m + (\sum_{i=1}^{n} x_i)}{c + n}$$

Where m is a prior for true mean, and C is a constant representing how many elements would be necessary to reliably estimate a sample mean.

Code

Part XII

Supervised Learning

46

47

48

49

Videos

StatQuest with Josh Starmer Gradient Boost Part 1 (of 4): Regression Main Ideas XGBoost Part 1 (of 4): Regression

XGBoost LightGBM

50 Gaussian Processes

GPJax "GPJax aims to provide a low-level interface to Gaussian process (GP) models in Jax, structured to give researchers maximum flexibility in extending the code to suit their own needs."

Part XIII

Unsupervised Learning

51

References

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*." *The Quarterly Journal of Economics* 134 (2): 557–98. https://doi.org/10.1093/qje/qjy029.