Datasets

Stat 151A: Linear Models

The following datasets are used in lectures, homework, and labs.

Datasets eligible for final project

Ames Housing dataset

This dataset can be downloaded from the github repo of Veridical Data Science by Yu and Barter.

- Link to csv
- Source link (downloaded Aug 2024)

Microcredit dataset

Paper abstract:

We use a clustered randomized trial, and over 16,000 household surveys, to estimate impacts at the community level from a group lending expansion at 110% APR by the largest microlender in Mexico. We find no evidence of transformative impacts on 37 outcomes (although some estimates have large confidence intervals), measured at a mean of 27 months post-expansion, across six domains: microentrepreneurship, income, labor supply, expenditures, social status, and subjective well-being. We also examine distributional impacts using quantile regressions, given theory and evidence regarding negative impacts from borrowing at high interest rates, but do not find strong evidence for heterogeneity.

I am using the data as preprocessed by Rachael Meager, downloaded from this repository. You should look at the file import_organise_data_v7.R, which was Rachael's original script, for information about the meaning of the columns. The original source of data is this repository, which I find poorly documented. I post-process Rachael's data using this script.

- Link to paper
- Link to Rdata

Teaching dataset

Paper abstract:

Student evaluations of teaching can be unreliable indicators of effective teaching and affected by implicit bias. We conduct a randomized experi- ment at a selective U.S. liberal arts college in which we vary both the instrument and timing at which we solicit student feedback to assess whether either intervention can mitigate gender disparities in qualitative evaluation comments. External evaluators with expertise in teaching eval- uation scored the positivity, specificity, and constructiveness of student comments in control and treatment groups. We find neither intervention is successful in mitigating gender disparities. Students that receive an alternate prompt which highlights the potential for implicit bias and asks them to articulate and apply their own criteria for effective teaching penalize female faculty at similar rates as those receiving a more stan- dard evaluation prompt. We also find no evidence that delaying solicita- tion of student feedback until the start of the next semester, when students face less stress and fatigue, reduces the gender gap in the pos- itivity of feedback. Instead, we find substantial evidence that the positiv- ity of qualitative student evaluations is affected by student, instructor, and course characteristics across all study conditions. The results suggest that bias in student evaluations may be difficult to mitigate.

- Link to paper
- Link to csv
- Link to data source

Bodyfat dataset

Bodyfat and other physical measurements on a number of individuals.

Measurement stfiveandards are apparently those listed in Benhke and Wilmore (1974), pp. 45-48 where, for instance, the abdomen 2 circumference is measured "laterally, at the level of the iliac crests, and anteriorly, at the umbilicus".

These data are used to produce the predictive equations for lean body weight given in the abstract "Generalized body composition prediction equation for men using simple measurement techniques", K.W. Penrose, A.G. Nelson, A.G. Fisher, FACSM, Human Performance Research Center, Brigham Young University, Provo, Utah 84602 as listed in Medicine and Science in Sports and Exercise, vol. 17, no. 2, April 1985, p. 189.

- Link to paper
- Link to csv

Spotify dataset

This dataset consists of roughly 30,000 Songs from the Spotify API with black-box machine learning quantifications of musical features. No guarantees are made on how the tracks were sampled.

- Link to csv
- Source link (downloaded Dec 2023)

Births dataset

Every year, the US releases to the public a large data set containing information on births recorded in the country. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This is a random sample of 1,000 cases from the data set released in 2014.

Columns

- fage: Father's age in years.
- mage: Mother's age in years.
- mature: Maturity status of mother.
- weeks: Length of pregnancy in weeks.
- premie: Whether the birth was classified as premature (premie) or full-term.
- visits: Number of hospital visits during pregnancy.
- gained: Weight gained by mother during pregnancy in pounds.
- weight: Weight of the baby at birth in pounds.
- lowbirthweight: Whether baby was classified as low birthweight (low) or not (not low).
- sex: Sex of the baby, female or male.
- habit: Status of the mother as a nonsmoker or a smoker.
- marital: Whether mother is married or not married at birth.
- whitemom: Whether mom is white or not white.

Source

United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. Natality Detail File, 2014 United States. Inter-university Consortium for Political and Social Research, 2016-10-07.

- Link to csv
- Source link

Other datasets used in class

Grades dataset

These are assignment and final grades compiled from the last few semesters Prof. Giordano has been taught. Each row is a student in a class, and the columns ar

- `grade`: The grade in the class
- `final`: The grade on the final exam
- `hw`: The overall homework grade
- `quizzes`: The overall quiz grade
- `class`: Which class was being taugh
- `semester`: Which semester the class was taught
- `time`: A chronologically ordered factor

Link to csv

Kleiber dataset

A classic dataset relating animal metabolic rates to weight across a wide range of scales.

- Link to paper
- Link to csv

Heights

A classic dataset relating the heights of father to the heights of sons.

- Link to Galton csv
- Link to Galton paper
- Link to Pearson tsv