ORIE 4741: Learning with Big Messy Data

Exploratory Data Analysis

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Announcements

- If you're taking lecture async: remember to submit participation post after each class!
- Otherwise, register your iClicker.
- Sections start next week. They are optional, attend any one you prefer. Only two (on Tuesday and Wednesday) will be live.
- Office hours: Zoom links or rooms and times are posted on course website.
- Gradescope is open for submission of hw0, due Thursday 9:30am.
- First quiz this week! It should occupy about 20 minutes; you'll have up to half an hour to complete it. Start it anytime between 10am Friday and noon Saturday.

Questions from zulip

- enrollment: yes, we expect you'll get in!
- protocol:
  - use the right stream (eg, general, homework, project, ...) and a good subject line
  - search for your question before posting new question

Our programming language policy

- we'll do demos and provide homework starter code in python
- you're welcome to use any language you like (that your TAs can read) for homework or project
- TAs will only support python

Topics to review

We will cover (most of) these in section, too:

- Linear algebra: invertible matrices, rank, norm, basic matrix identities. When is a matrix invertible?
- QR factorization
- Gradients (multivariate derivative)
- Projections
- SVD
- Maximum likelihood estimation
- Union bound
- Computational complexity

Why look at the data?

- detect errors in data
- check assumptions
- select appropriate models
- understand relationships among the features
- understand relationships between features and labels

How to look at the data?

- inspect raw data
- summary statistics
- visualize

American community survey

2013 ACS:

- 3M respondents, 87 economic/demographic survey questions
  - income
  - cost of utilities (water, gas, electric)
  - weeks worked per year
  - hours worked per week
  - home ownership
  - looking for work
  - use foodstamps
  - education level
  - state of residence
  - ...
- 1/3 of responses missing

find it at `https://people.orie.cornell.edu/mru8/orie4741/data/acs_2013.csv`

How do computers work?

on a laptop:

- hard disk: usually $\leq$ 500 GB
- memory (RAM): usually $\leq$ 16 GB
- many programs (*e.g.*, Excel): substantially more limited

How do computers work?

on a laptop:

- hard disk: usually $\leq$ 500 GB
- memory (RAM): usually $\leq$ 16 GB
- many programs (*e.g.*, Excel): substantially more limited

don't load a giant file into memory.
your computer will crash.

How do computers work?

on a laptop:

- hard disk: usually $\leq$ 500 GB
- memory (RAM): usually $\leq$ 16 GB
- many programs (*e.g.*, Excel): substantially more limited

don't load a giant file into memory.
your computer will crash.

how big is ACS data?
3M respondents $\times$ 100 questions $=$ 300M numbers $\approx$ 300MB

Inspect raw data

solution for large files: technology from the 70s!

bash shell:

- "how big are these files?": ls -lh
- "show me some lines from the file": head, tail, less
- "how many lines are in the file?": wc -l

American Community Survey

Variable	Description	Type
HHTYPE	household type	categorical
STATEICP	state	categorical
OWNERSHP	own home	Boolean
COMMUSE	commercial use	Boolean
ACREHOUS	house on $\geq$ 10 acres	Boolean
HHINCOME	household income	real
COSTELEC	monthly electricity bill	real
COSTWATR	monthly water bill	real
COSTGAS	monthly gas bill	real
FOODSTMP	on food stamps	Boolean
HCOVANY	have health insurance	Boolean
SCHOOL	currently in school	Boolean
EDUC	highest level of education	ordinal
GRADEATT	highest grade level attained	ordinal
EMPSTAT	employment status	categorical
LABFORCE	in labor force	Boolean
CLASSWKR	class of worker	Boolean
WKSWORK2	weeks worked per year	ordinal
UHRSWORK	usual hours worked per week	real
LOOKING	looking for work	Boolean
MIGRATE1	migration status	categorical

Python and Jupyter

- Python is a programming language:
  it parses human-readable code to machine-readable code,
  executes it, returns the answer
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as `.ipynb` files.
- Jupyter notebooks display inputs and outputs in a browser
- Google Colab is an interface to a webserver running Python

Python and Jupyter

- Python is a programming language:
  it parses human-readable code to machine-readable code,
  executes it, returns the answer
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as `.ipynb` files.
- Jupyter notebooks display inputs and outputs in a browser
- Google Colab is an interface to a webserver running Python

how to access?

- install Python with Anaconda distribution (versions 3.7 or 3.8 are fine)
- use Google Colab

Summary statistics

univariate

- mean, median, mode
- max, min, range
- variance
- ...

explore via Python + Jupyter notebook

`https://github.com/ORIE4741/demos/blob/master/eda.ipynb`

Summary statistics

univariate

- mean, median, mode
- max, min, range
- variance
- ...

explore via Python + Jupyter notebook

`https://github.com/ORIE4741/demos/blob/master/eda.ipynb`

multi- (but usualy just bi-)variate

- correlation, covariance
- ...

The perils of summary statistics

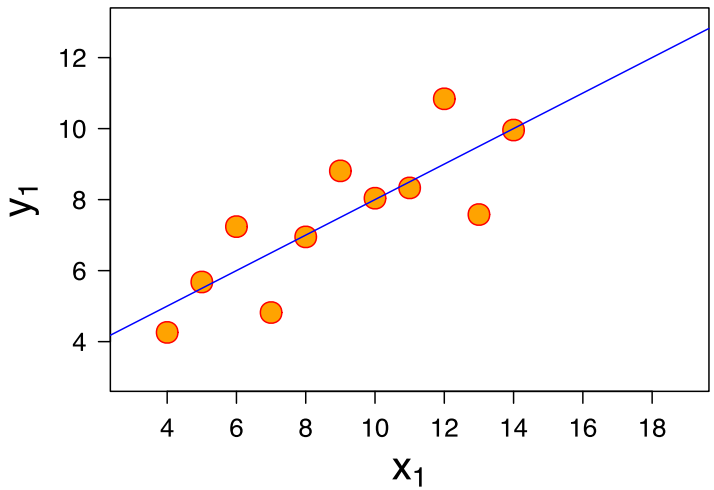
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

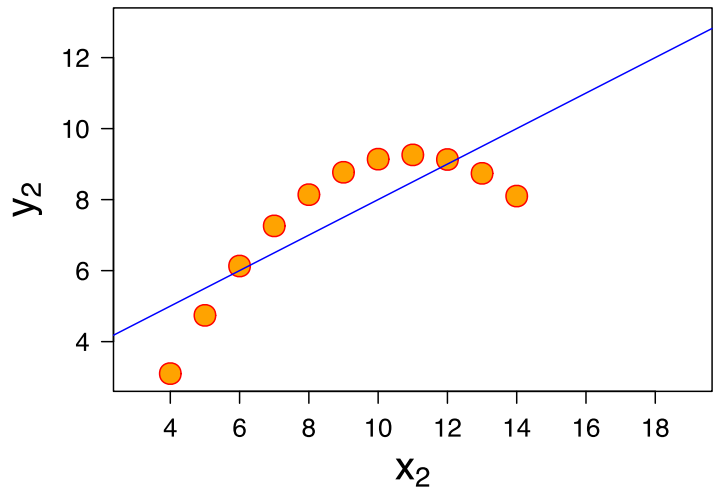
The perils of summary statistics

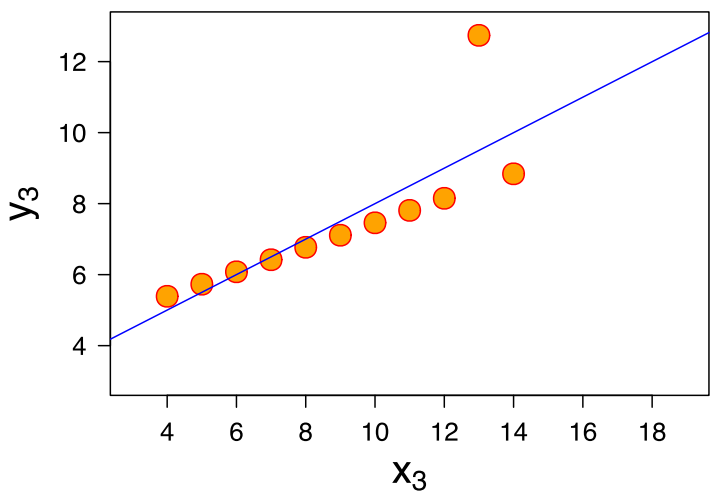
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

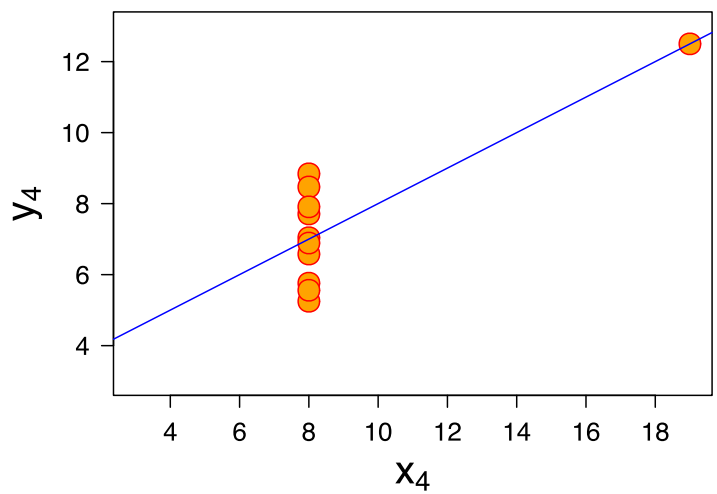
same mean, variance, correlation, line of best fit. . .

The perils of summary statistics









The perils of summary statistics: modern update

https://www.autodeskresearch.com/publications/samestats

What to visualize?

- examples across all features (usually not)
- plot features across all examples (much more common)

Best practices

- Always label your axes.
- Ensure all marks on plot are meaningful.
- Beware of pie charts; bar charts are often easier to read.
- Beware of line plots; if your data is not continuous, try scatter plot instead.
- Consider the scale of your axes. Log scale or not?
- Consider which curves to plot on same axes. Make comparisons easy!

Beware of bad data

Label: Number of Days Physical Health Not Good
Section Name: Healthy Days — Health Related Quality of Life
Core Section Number: 2
Question Number: 1
Column: 91-92
Type of Variable: Num
SAS Variable Name: PHYSHLTH
Question Prologue:
Question: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 30	Number of days	159,327	36.43	35.59
88	None	269,145	61.53	62.53
77	Don't know/Not sure	7,602	1.74	1.58
99	Refused	1,336	0.31	0.30
BLANK	Not asked or Missing	26	.	.

Take away

- always look at (some of) your data
- decide what question you want to answer

Questions?

https://docs.google.com/spreadsheets/d/
1vLbwi0WCOn0wU6cU_r0RHAnY7C0fDZ1F8Yq09pqYYuk