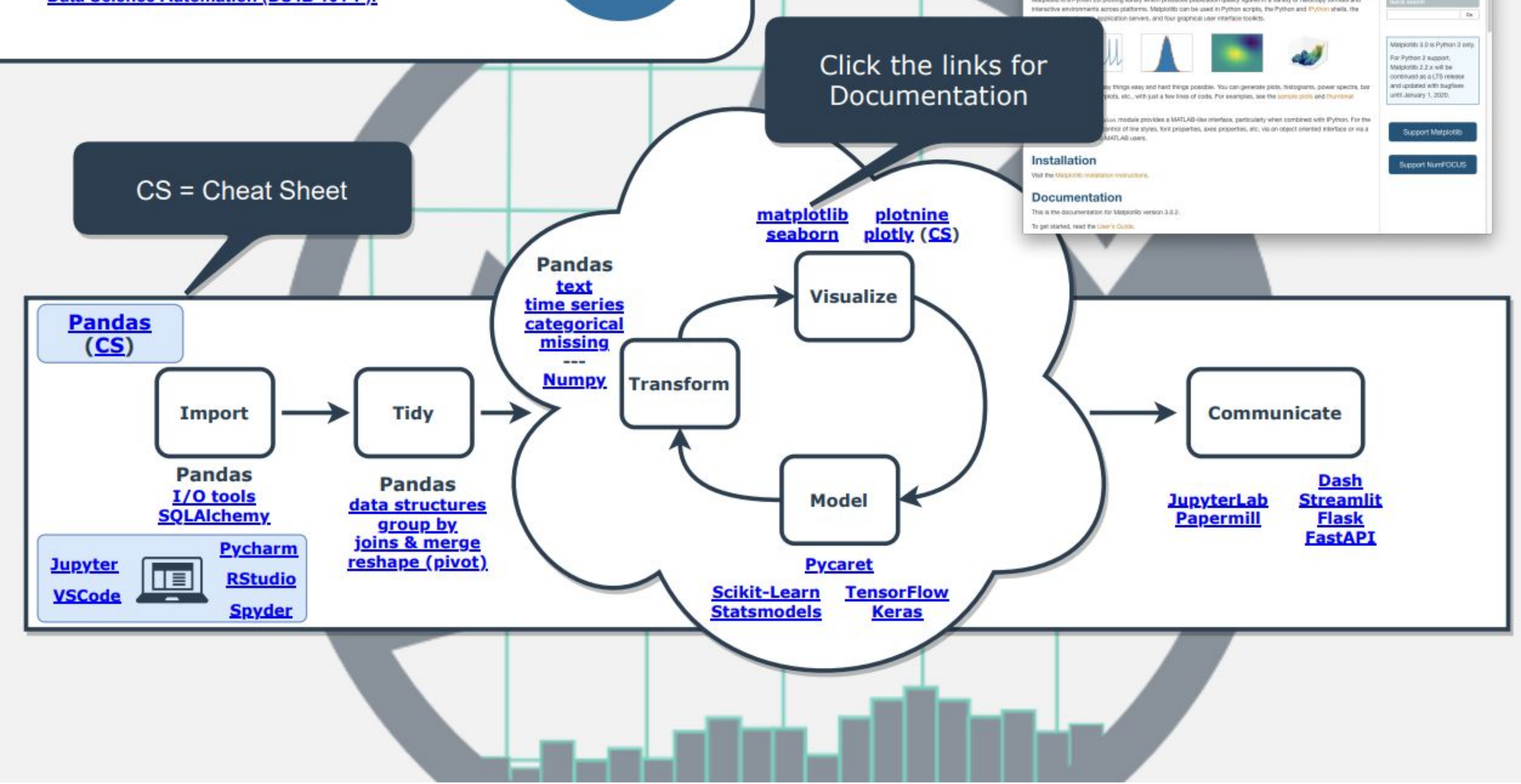
Selected Topics in Data Science

Vsevolod Suschevskiy
(Seva)



[Ultimate Python Data Science Cheat Sheet (business-science.io)]

Based on

MS&E 226: “Small” Data — https://web.stanford.edu/~rjohari/teaching/syllabus/226_syllabus_2018.pdf

CS109A Introduction to Data Science — https://harvard-iacs.github.io/2021-CS109A

Plus some random twitter threads on data science

Who is

Vsevolod Suscheskiy — me

Alena Suvorova — Candidate of Sciences, Computer Science

Ilya Musabirov — PhD student in the Department of Computer Science at the University of Toronto

What about

“Big” data — data at unprecedented levels of granularity.

- Billions of: Facebook posts, tweets, medical tests, power meter readings...
- Often arriving faster than we can store and analyze it

Key features of “big” data:

- **Can’t be analyzed on a single machine.**
- Requires new algorithms and tools to store, query, and analyze the data.

“Small” data

Data that can be analyzed, processed, etc., on a single machine.

- Advances in technology means even “small” data is getting bigger (e.g., 32GB of RAM even on home PCs)
- Most analysis of “big” data starts by understanding “small” data

This class is a user’s manual for “small” data analysis.

In the process you will learn skills that should help you for any data analysis.

“Small” data

Data that can be analyzed, processed, etc., on a single machine.

- Advances in technology means even “small” data is getting bigger (e.g., 32GB of RAM even on home PCs)
- Most analysis of “big” data starts by understanding “small” data

This class is a user’s manual for “small” data analysis.

In the process you will learn skills that should help you for any data analysis.

And to successfully pass an interviews for some kind of DS position

Selected topics

1. Summarization
   a. Given a single data set, how do we summarize it?
   b. Basic sample statistics; models; linear and logistic regression; in-sample fit (R2 and residuals).
2. Prediction
   a. How do we generalize our understanding of a data set to new samples?
   b. Binary classification; linear regression and logistic regression as approaches to prediction; model complexity and the bias-variance decomposition; out-of-sample validation.

Selected topics

3. Inference
    a. How do we generalize our understanding of a data set to draw inferences about the population or system from which the data came?
    b. Frequentist estimation and hypothesis testing; application to linear regression; bootstrap; multiple hypothesis testing. Comparison to Bayesian approaches.
4. Causality
    a. How do we determine the effect that changing a system will have?
    b. The Rubin causal model, potential outcomes, and counterfactuals; randomized experiments; causal inference from observational data; data-driven decision making.

Summarization, Prediction, Inference, or Causality

Customer profile report

Evaluate the usefulness of the new feature

Evaluate the success of the marketing campaign on the new TV channel

Recommend a movie to watch for the evening

Prescribe treatment for cancer

To find out if a new mobile app helps students learn better

Customer profile report

- Summarizatoin
- Prediction
- Inference
- Causality

Customer profile report

Summarizatoin

Prediction

Inference

Causality

Customer profile report

Summarizatoin

Prediction

Inference

Causality

Evaluate the usefulness of the new feature

- Summarizatoin
- Prediction
- Inference
- Causality

To 0

Evaluate the usefulness of the new feature

Summarizatoin

Prediction

Inference

Causality

Evaluate the usefulness of the new feature

Summarizatoin

Prediction

Inference

Causality

Evaluate the success of the marketing campaign on the new TV channel

Summarizatoin

Prediction

Inference

Causality

Recommend a movie to watch for the evening

Summarizatoin

Prediction

Inference

Causality

Classical first class on Data Science

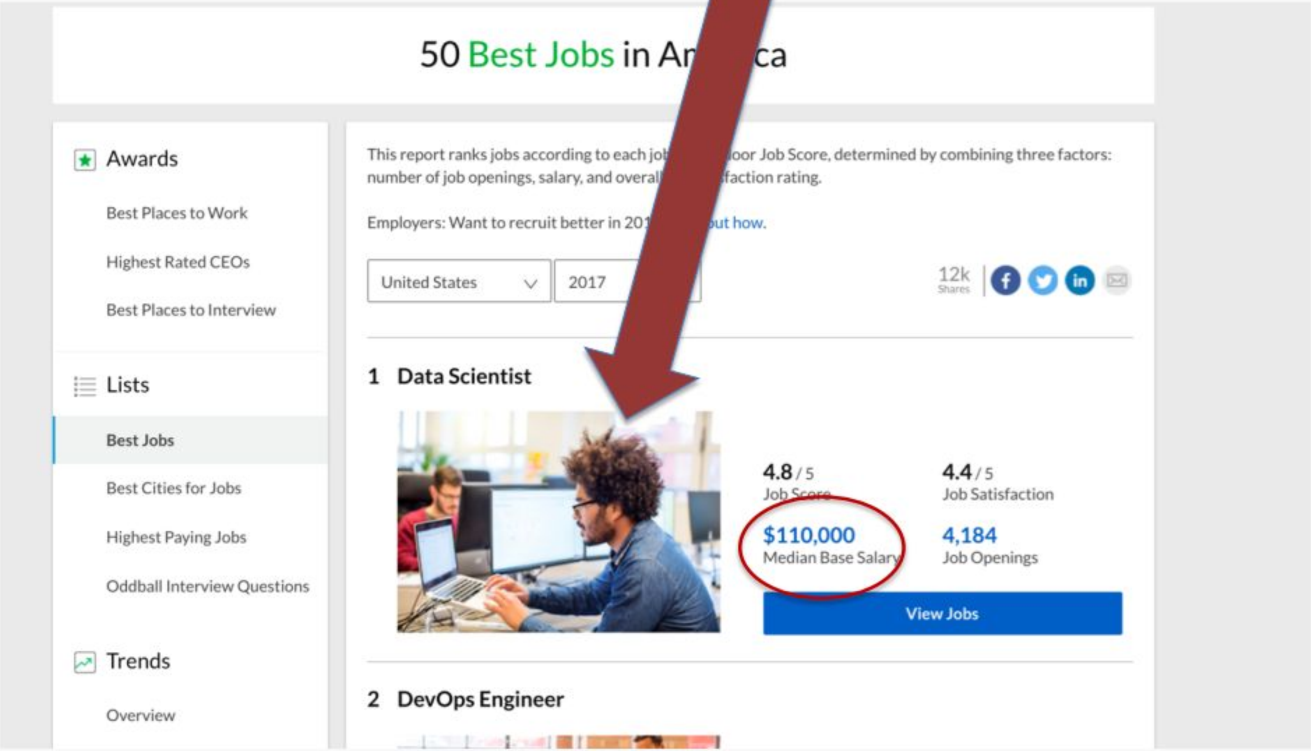
50 Best Jobs in America for 2020

Best Jobs | 2020 | United States

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Front End Engineer	$105,240	3.9/5	13,122	View Jobs
#2	Java Developer	$83,589	3.9/5	16,136	View Jobs
#3	Data Scientist	$107,801	4.0/5	6,542	View Jobs
#4	Product Manager	$117,713	3.8/5	12,173	View Jobs
#5	DevOps Engineer	$107,310	3.9/5	6,603	View Jobs
#6	Data Engineer	$102,472	3.9/5	6,941	View Jobs
#7	Software Engineer	$105,563	3.6/5	50,438	View Jobs

Why?

Jobs!



50 Best Jobs in America
Awards
Best Places to Work
Highest Rated CEOs
Best Places to Interview
Lists
Best Jobs
Best Cities for Jobs
Highest Paying Jobs
Oddball Interview Questions
Trends
Overview
Employers: Want to recruit better in 201
United States
2017
12k Shares
1 Data Scientist
4.8 / 5
4.4 / 5
Job Satisfaction
$110,000
Median Base Salary
4,184
Job Openings
View Jobs
2 DevOps Engineer

History

Long time ago (thousands of years) science was only empirical and people counted stars

History (cont)

Long time ago (thousands of years) science was only empirical and people counted stars or crops

History (cont)

Long time ago (thousands of years) science was only empirical and people counted stars or crops and used the data to create machines to describe the phenomena

Few hundred years: theoretical approaches, try to derive equations to describe general phenomena.

1. $\nabla \cdot \mathbf{D} = \rho_V$
2. $\nabla \cdot \mathbf{B} = 0$
3. $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$

$$T^2 = \frac{4\pi^2}{GM} a^3$$

can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

T Earth years

a Astronomical units AU (a = 1 AU for Earth)

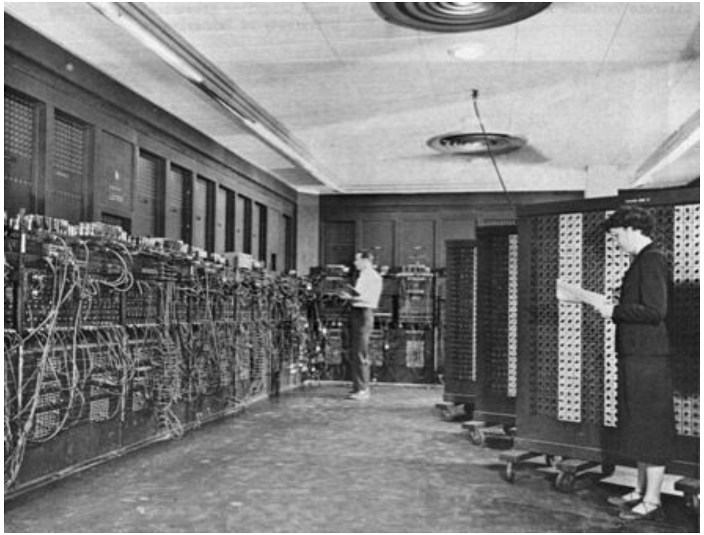
M Solar masses $M_\odot$

Then $\frac{4\pi^2}{G} = 1$

$$H(t)|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t}|\psi(t)\rangle$$

History (cont)

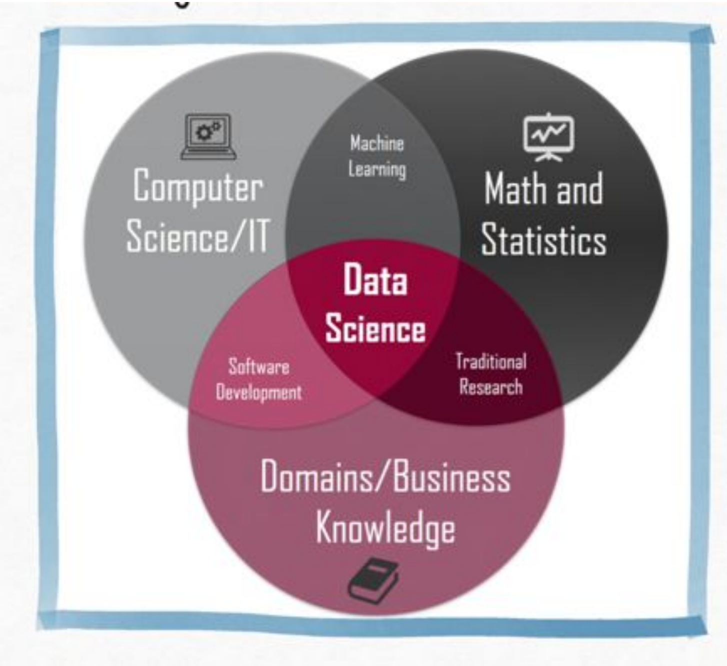
About a hundred years ago: computational approaches



History (cont)

And then …. data science

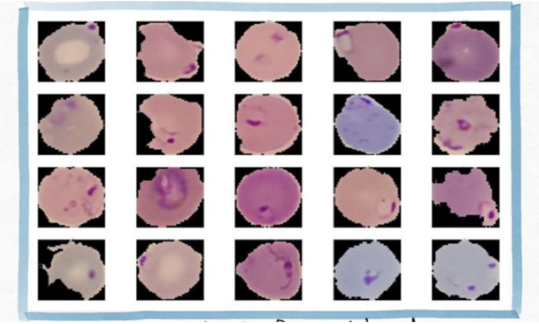
In both data science and machine learning we extract pattern and insights from data.



- Inter-disciplinary
- Data and task focused
- Resource aware
- Adaptable to changes in the environment and needs

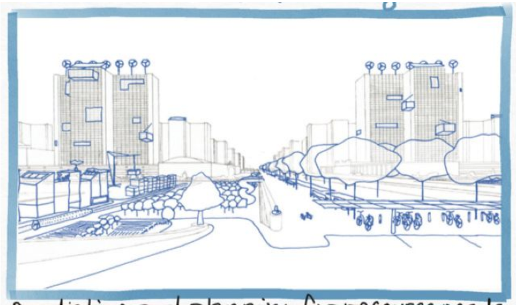
The Potential of Data Science

Disease Diagnosis



Detecting malaria from blood smears

Urban Planning



Predicting and planning for resource needs

Drug Discovery

Quickly discovering new drugs for COVID

Agriculture

Precision agriculture

The Potential of Data Science

Gender Bias

Some DS models for evaluate job applications show bias in favor of male candidate

Racial Bias

Risk models used in US courts have shown to be biased against non-white defendants

What?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

What?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

What is the scientific goal?

What would you do if you had **all** of the data?

What do you want to predict or estimate?

What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

How were the data sampled?

Which data are relevant?

Are there privacy issues?

What?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

Plot the data.

Are there anomalies or egregious issues?

Are there patterns?

What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Build a model.

Fit the model.

Validate the model.

What?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

What did we learn?

Do the results make sense?

Can we effectively tell a story?

Goal of the course

Theory

1. Key Machine Learning concept
2. Important metrics for evaluation
3. Handling different kinds of data
4. Extracting insights from analysis of the models

Practice

1. Implement ML and deep learning models using python libraries
2. Using free online tools and resources for data science

Impact

1. Solving real-life problems using DS
2. Evaluating the social impact of DS

Hidden goal

Theory

1. Let other people who studied DS understand that you learned DS

Practice

1. Being able to solve the test assignment for the analyst position

Impact

1. Use some of Data Analysis techniques in your thesis

Which data science skills are important (To get a $50,000 increase in salary)

[Which data science skills are important (To get a $50,000 increase in salary) (business-science.io)]

Plan	Skills
Machine Learning	Supervised Classification, Supervised Regression, Unsupervised Clustering, Dimensionality Reduction, Local Interpretable Model Explanation - H2O Automatic Machine Learning, parsnip (XGBoost, SVM, Random Forest, GLM), K-Means, UMAP, recipes, lime
Data Visualization	Interactive and Static Visualizations, ggplot2 and plotly
Data Wrangling & Cleaning	Working with outliers, missing data, reshaping data, aggregation, filtering, selecting, calculating, and many more critical operations, dplyr and tidyr packages
Data Preprocessing & Feature Engineering	Preparing data for machine learning, Engineering Features (dates, text, aggregates), Recipes package
Time Series	Working with date/datetime data, aggregating, transforming, visualizing time series, timetk package
Forecasting	ARIMA, Exponential Smoothing, Prophet, Machine Learning (XGBoost, Random Forest, GLMnet, etc), Deep Learning (GluonTS), Ensembles, Hyperparameter Tuning, Scaling to 1000s of forecasts, Modeltime package
Text	Working with text data, Stringr
NLP	Machine learning, Text Features
Functional Progamming	Making reusable functions, sourcing code
Iteration	Loops and Mapping, using Purrr package
Reporting	Rmarkdown, Interactive HTML, Static PDF
Applications	Building Shiny web applications, Flexdashboard, Bootstrap
Deployment	Cloud (AWS, Azure, GCP), Docker, Git
Databases	SQL (for data import), MongoDB (for apps)

The first break somewhere here

Lets meet each other

https://forms.office.com/Pages/ResponsePage.aspx?id=JGzyIZMHB0unPVY80uwjX9YIlgF3uQBOpHh4x5ymTLdUNVhHT1Y2WDhORjhHSEhPUzlXUzhUN1I1MS4u

Summarizing a sample

We have a data

1. Salary of Data Scientists
2. Height of students
3. Democratic values

We have a mean

1. Salary of Data Scientists — 87963.01982372967787
2. Height of students — 169.21345
3. Intelligence scores— 3.3234

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              kid_score   R-squared:                       0.215
Model:                            OLS   Adj. R-squared:                  0.208
Method:                 Least Squares   F-statistic:                     29.38
Date:                Wed, 06 Apr 2022   Prob (F-statistic):           1.31e-21
Time:                        23:44:35   Log-Likelihood:                -1871.7
No. Observations:                 434   AIC:                             3753.
Df Residuals:                     429   BIC:                             3774.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     20.8226      9.188      2.266      0.024       2.764      38.881
mom_iq         0.5621      0.061      9.249      0.000       0.443       0.682
mom_hs         5.5612      2.313      2.404      0.017       1.014      10.108
mom_work       0.1337      0.768      0.174      0.862      -1.375       1.643
mom_age        0.2199      0.332      0.662      0.509      -0.433       0.873
==============================================================================
Omnibus:                        7.277   Durbin-Watson:                   1.623
Prob(Omnibus):                  0.026   Jarque-Bera (JB):                7.480
Skew:                          -0.313   Prob(JB):                       0.0238
Kurtosis:                       2.851   Cond. No.                     1.09e+03
==============================================================================


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.09e+03. This might indicate that there are
```

Placeholder for some code

https://github.com/vvseva/DS_DAPS

Relationship

Modeling relationships

We focus on a particular type of summarization:

Modeling the relationship between observations.

Formally:

- Let Y_i, $i = 1, \ldots, n$, be the i'th observed (real-valued) outcome.
  - Let $Y = (Y_1, \ldots, Y_n)$
- Let X_{ij} , $i = 1, \ldots, n$, $j = 1, \ldots, p$ be the i'th observation of the j'th (real-valued) covariate.
  - Let $X_i = (X_{i1}, \ldots, X_{ip})$.
  - Let X be the matrix whose rows are X_i

.

Pictures and names

How to visualize Y and X?

Names for the Yi ’s: *outcomes, response variables, target variables, dependent variables*

Names for the Xij ’s: *covariates, features, regressors, predictors, explanatory variables, independent variables*

X is also called *the design matrix*

Data

The kidiq dataset loaded earlier contains the following columns:

- `kid_score` Child's score on IQ test
- `mom_hs` Did mom complete high school?
- `mom_iq` Mother's score on IQ test
- `mom_work` Working mother?
- `mom_age` Mother's age at birth of child

[Note: Always question how variables are defined!]

Reasonable question: How is `kid_score` related to the other variables?

Continuous variables

Variables such as `kid_score` and `mom_iq` are continuous variables: they are naturally real-valued.

For now we only consider outcome variables that are continuous (like `kid_score`). Note: even continuous variables can be constrained:

- Both `kid_score` and `mom_iq` must be positive.
- `mom_age` must be a positive integer.

Categorical variables

Other variables take on only finitely many values, e.g.:

- `mom_hs` is 0 (resp., 1) if mom did (resp., did not) attend high school
- `mom_work` is a code that ranges from 1 to 4:
    - 1 = did not work in first three years of child's life
    - 2 = worked in 2nd or 3rd year of child's life
    - 3 = worked part-time in first year of child's life
    - 4 = worked full-time in first year of child's life

These are categorical variables (or factors).

Modeling relationships

Goal: Find a functional relationship f such that:

$Y_i \approx f(X_i)$

This is our first example of a “model.”

We use models for lots of things:

- Associations and correlations
- Predictions
- Causal relationships

Linear regression models

Linear relationships

We first focus on modeling the relationship between outcomes and covariates as linear.

In other words: find coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$ such that: [1]

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}.$$

This is a linear regression model.

"hats" on variables denote quantities computed from data. In this case, whatever the coefficients are, they will have to be computed from the data we were given.

How to choose β^?

There are many ways to choose $\hat{\boldsymbol{\beta}}$.

We focus primarily on *ordinary least squares* (OLS):

Choose $\hat{\beta}$ so that

$$\text{SSE} = \text{sum of squared errors} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

is minimized, where

$$\hat{Y}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_j X_{ij}$$

is the *fitted* value of the i'th observation.

Questions to ask

Here are some important questions to be asking:

- Is the resulting model a good fit?
- Does it make sense to use a linear model?
- Is minimizing SSE the right objective?

Homework?