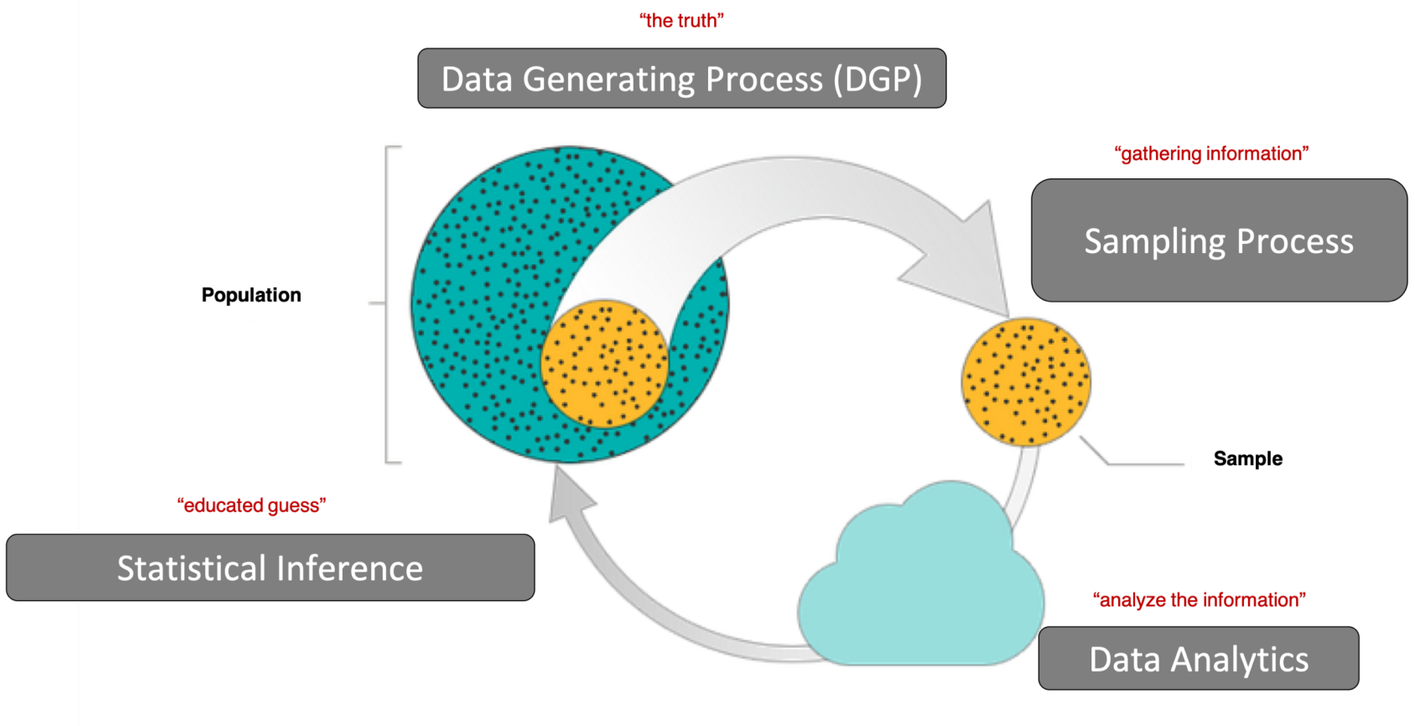
URP 6931. Introduction to Urban Analytics

Lecture 03: Univariate linear regression

Instructor: Shenhao Wang
Assistant Professor, Director of Urban AI Lab
Department of Urban and Regional Planning
University of Florida

Recap Lecture 02

Review Probability
P(x)



Review Vector
X

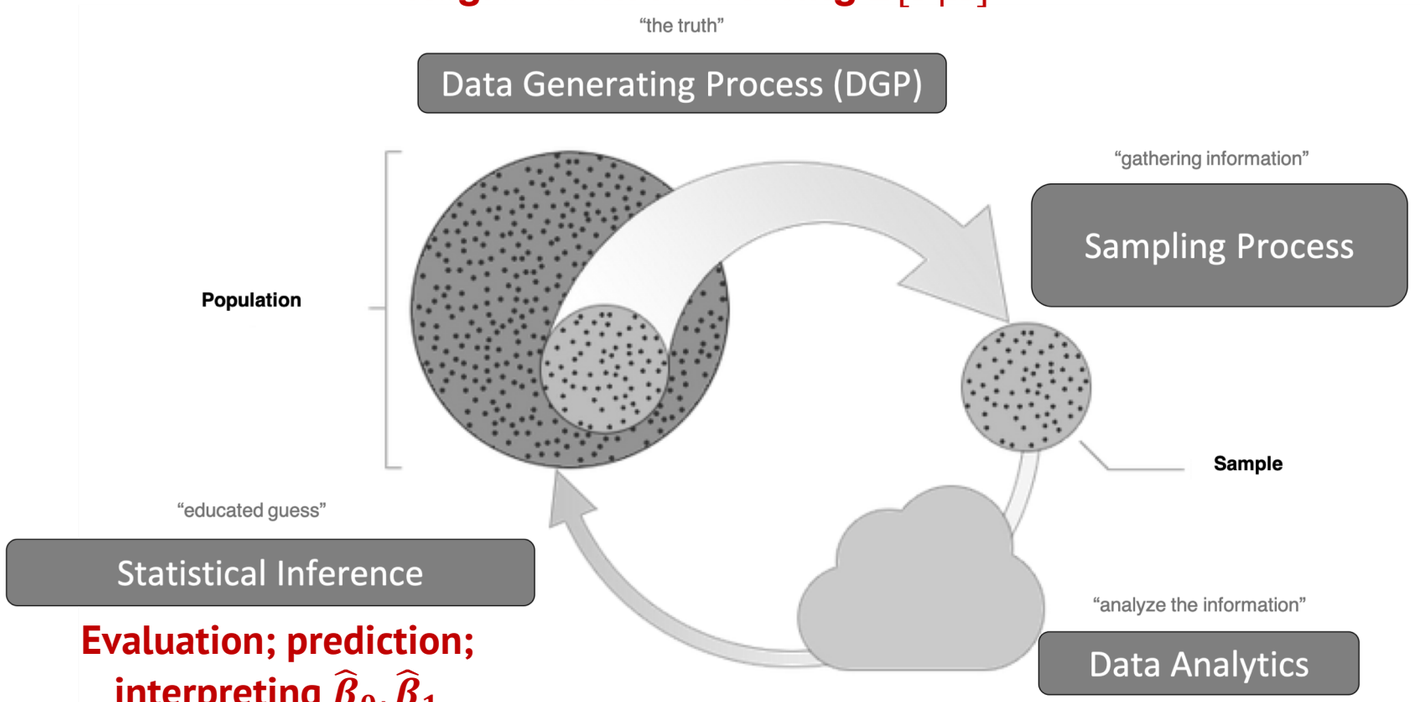
Review Statistics
$\widehat{P}(x)$

Lecture 03. Univariate linear regression

1. What is regression?
   Recovering the conditional mean function
   $E[Y|X]$
2. Ordinary least square
   Univariate linear regression
   $\hat{E}[Y|X] = \hat{\beta}_0 + \hat{\beta}_1 x_i$
3. Use the model
   Evaluation, interpretation, and prediction
4. The general paradigm
   A general analytical process
5. Lab session
   Matplotlib, Statsmodels.

Mapping Lecture 03 to the General Diagram

Regression: recovering $E[Y|X]$



i.i.d.
Y: continuous

Evaluation; prediction; interpreting $\widehat{\beta}_0, \widehat{\beta}_1$

Regression: $\widehat{E}[Y|X] = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$
Estimation/Training/Learning

Part 1. What is regression?

What is a regression?

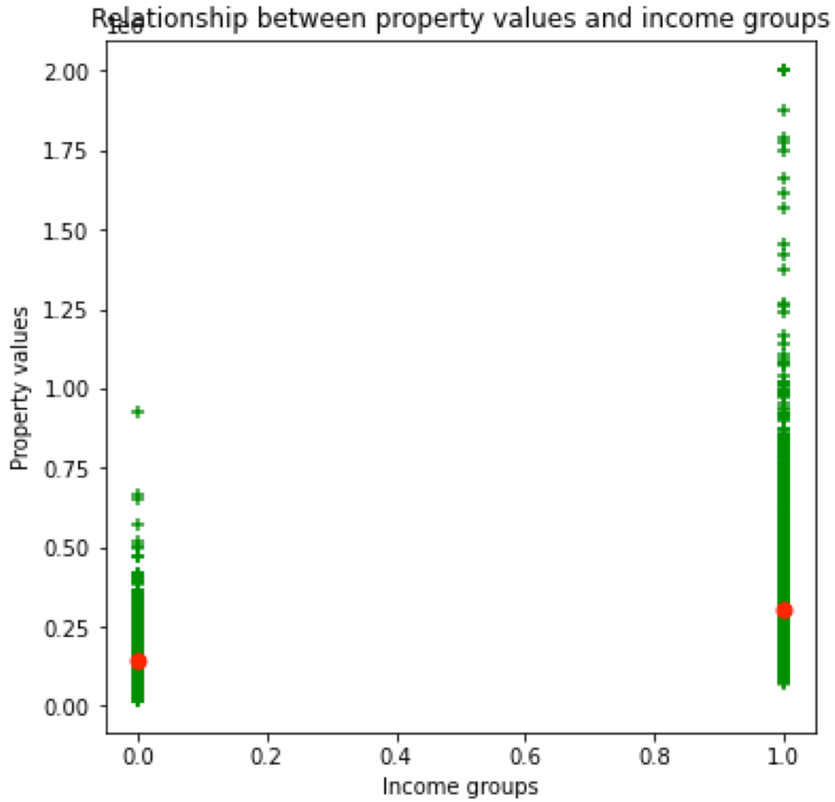
- Regression quantifies how an outcome variable Y varies as a function of one or more predictor variables X
- Key idea: conditioning on X. e.g. characterizing P(YIX), E[YIX], etc.
- In linear regression, we seek to recover the conditional mean of Y given X:

$$\boldsymbol{E[Y|X=x]}$$

- Prof (skipped): $E[Y|X=x] = \underset{f(x)}{\operatorname{argmin}} E[\big(Y - g(X)\big)^2]$
- Examples
  - X: income, Y: property values;
  - X: education, Y: income.
  - X: education, income; Y: automobile ownership
  - X: ∅, Y: property values.

What is the intuition about conditional mean function $E[Y|X = x]$

- X: continuous household income or discrete income groups
- Y: property values
- Why do we focus on conditional mean? It is the most intuitive and useful one.



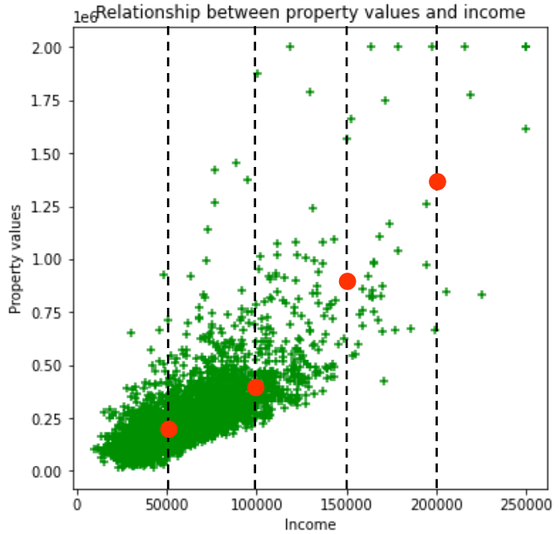
Imposing a linear assumption

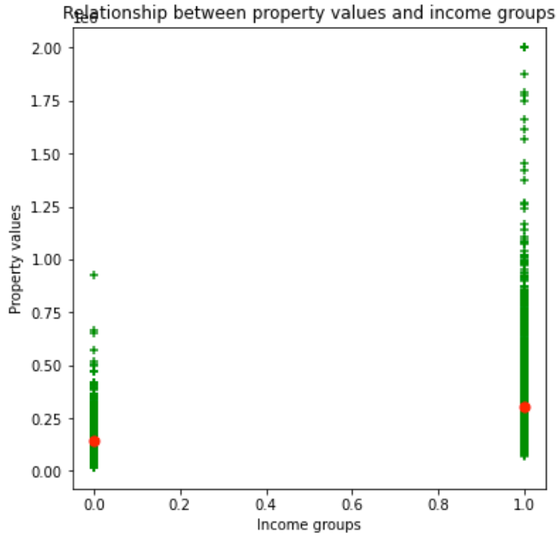
Notes

- What I just described is a non-parametric approach, which is challenging for continuous independent variables.
- A common starting point is a parametric approach with a linear assumption

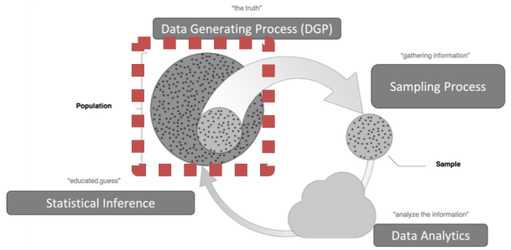
$$E[Y|X = x] = f(x)$$

- Assume $f(x)$ is linear. e.g. $f(x) = \beta_0 + \beta_1 x_1$. Use this simple linear function to approximate the nonlinear relationship.
- Why? We want to summarize the whole picture into one single number.





OLS Terminology



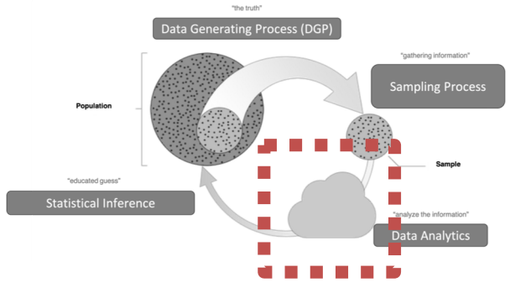
The population model (DGP) is given by:

$$Y = \beta_0 + \beta_1 X + u$$

- Y: Dependent variable.
- X: Independent variable.
- β_0, β_1: Intercept and slope.
- u: error term or disturbance term.

Note: We make very strong assumptions on DGP, the consequences of which will be discussed in the next week.

OLS Terminology



The estimated model is:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u}$$

- $\hat{\beta}_0, \hat{\beta}_1$: estimated intercept and slope
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. Fitted values or predictions.
- $\hat{u}$: Residuals – variation in Y left unexplained by X. It can be viewed as an estimate of u.

Note:

1. We are discussing the model assumptions. It can be translated as "we roughly know the truth" (for today's lecture).
2. But DGP and model could be different. e.g. rich DGP but simple model; simple DGP but rich model.

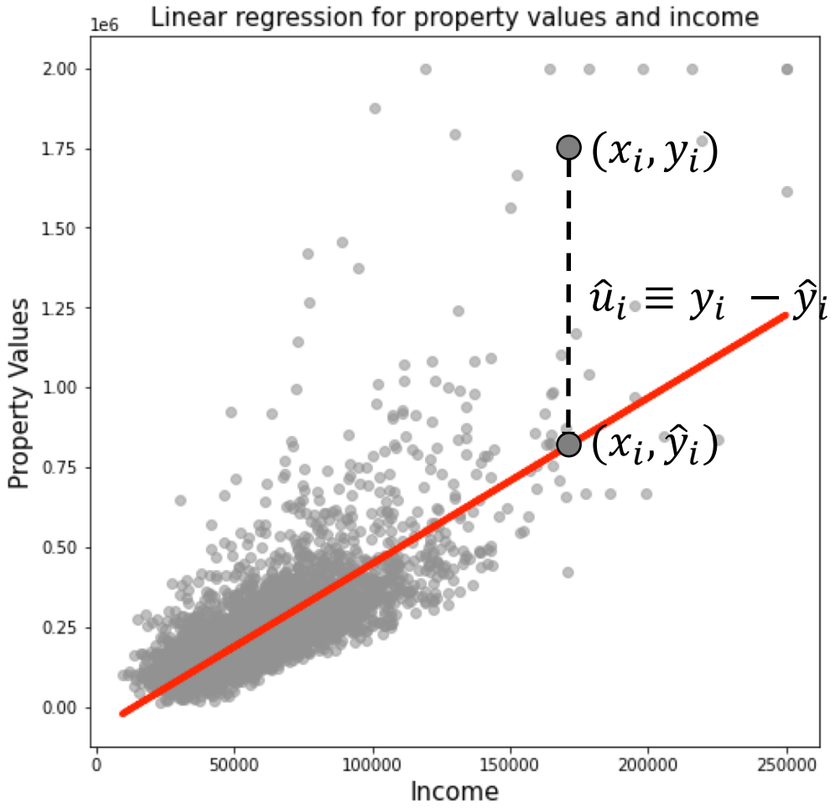
Part 2. Ordinary least square (OLS) for univariate linear regression

Motivating example: fitting property values and household income

Q: How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?

Motivating example: fitting property values and household income

A: We will minimize the sum of squared residuals



Ordinary Least Square (OLS)

Ordinary Least Squares (OLS) regression picks the following:

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{N} \hat{u}_i^2 = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - x_i\beta_1)^2$$

Why least squares, i.e., minimize the sum of squared differences?

- Easy to derive and analytically investigate.
- Optimal under certain assumptions.

OLS Example 1: only Y.

Let's understand how to calculate least squares analytically.

Consider the simplest case: minimizing the least squares for a single variable Y

N sample observations: $y_1, y_2, y_3, \ldots, y_N$.

Find $\hat{\mu}$ that minimizes the sum of squared residuals (SSR)

$$\underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \mu)^2$$

OLS Example 1: only Y.

$$S(\mu) = \sum_{i=1}^{N} (y_i - \mu)^2 = \sum_{i=1}^{N} \left(y_i^2 - 2y_i\mu + \mu^2\right)$$

Setting its derivative to zero

$$\frac{\partial S(\mu)}{\partial \mu} = \sum_{i=1}^{N} (-2y_i + 2\mu) = 0$$

Solving

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Remember that regression targets to recover $E[Y|X]$

When $X = \emptyset$, we recover $E[Y|\emptyset]$

In other words, you could treat OLS as an operational definition for computing the sample average.

Q: Do you have any guess what we recover if we use $\underset{\mu}{\text{argmin}} \sum_{i=1}^{N} |y_i - \mu|$?

A: Median. We can recover the median by minimizing mean absolute error.

OLS Example 2: Y and X.

Now we look at two variables, Y and X

N pairs of sample observations: $\{y_1, x_1\}, \{y_2, x_2\}, \ldots, \{y_n, x_n\}$

We need to find the $\{\hat{\beta}_0, \hat{\beta}_1\}$ to minimize the following objective function:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - x_i\beta_1)^2$$

How do we derive the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1?

- Analytically, compute the derivatives and set them to zero.
- Computationally, just click a button in Python.

OLS Example 2: Y and X.

First order condition:

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{N} (-2y_i + 2\beta_0 + 2x_i\beta_1) = 0$$

and

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{N} \left(-2y_i x_i + 2\beta_0 x_i + 2x_i^2\beta_1\right) = 0$$

Two equations and two unknown variables.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

OLS Example 2: Y and X.

Intuition of $\hat{\beta}_1$: "normalized" correlation between X and Y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\text{Sample Covariance between X and Y}}{\text{Sample Variance of X}}$$

The slope coefficient is the (partial) derivative of the regression function with respect to X:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u}$$
$$\frac{\partial Y(\hat{\beta}_0, \hat{\beta}_1)}{\partial X} = \hat{\beta}_1$$

$\hat{\beta}_1$ = how much Y changes when we change X by one unit

Intuition of $\hat{\beta}_0$: intercept – Y value when X = 0

OLS Example 2: Y and X. - What if X is a nominal value {0, 1}?

Example: X – {high income, low income}. Y: Property values

High income: 1; Low income: 0. – a.k.a. “dummy variable”

Question: Is it a nominal, ordinal, or cardinal number?

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

X = 0: $\hat{Y}(X = 0) = \hat{\beta}_0$

X = 1: $\hat{Y}(X = 1) = \hat{\beta}_0 + \hat{\beta}_1$

$\hat{\beta}_1 = \hat{Y}(X = 1) - \hat{Y}(X = 0)$

Intuition of $\hat{\beta}_1$: Difference of property values between the high and low income groups.

Intuition of $\hat{\beta}_0$: Average property value of the low income groups

Statistics is formalized intuition (continuous X)

Y: income; X: education (years).

$\hat{\beta}_1 <=> 0$?

People have higher income when they have more education.

Y: automobiles; X: income (dollars).

$\hat{\beta}_1 <=> 0$?

People buy more automobiles when they have higher income.

Statistics is formalized intuition (discrete X)

Y: income; X: race (majority, minority).

$\hat{\beta}_1 <=> 0$?

The majority groups have $\hat{\beta}_1$ more income than the minority groups. (Typical equity argument)

Y: automobiles; X: income (high, low).

$\hat{\beta}_1 <=> 0$?

The high-income groups have $\hat{\beta}_1$ more automobiles than the low-income groups.

Part 3. Evaluation, interpretation, and prediction

1. Model evaluation with R^2

Definitions

- $\sum_{i=1}^{N}(y_i - \bar{y})^2 = SST$ (Total Sum of Squares)
- $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = SSE$ (Explained Sum of Squares)
- $\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 = SSR$ (Residual Sum of Squares)

We could show that: SST = SSE + SSR

$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST}$$

1. Model evaluation with R^2

Since

$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST},$$

$$\frac{SSE}{SST} = 1 - \frac{SSR}{SST} \equiv R^2$$

R^2 Interpretation: percent total variation in Y that is explained by X. It is a very common performance metric to measure the power of the model.

Properties:

$$0 \leq R^2 \leq 1$$

- If $R^2 = 1$, all points are on a straight line (perfect fit).
- If $R^2 = 0$, no correlation between Y and X.

Question: Is it true that larger R^2 always indicates a better model?

2. Interpreting the model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

If X is continuous,

$\hat{\beta}_1$ measures how much Y changes when we change X by one unit

$\hat{\beta}_0$ measures the intercept of Y value when X = 0

If X is a binary discrete variable {0, 1},

$\hat{\beta}_1$ measures the average difference of Y between the two groups

$\hat{\beta}_0$ measures the Y value of group 0

3. Regression as prediction

Regression can be used for prediction.

1. New observation

$$Y_{new} = \hat{\beta}_0 + \hat{\beta}_1 X_{new}$$

2. delta disturbance (e.g. increase income, price, etc.)

$$Y_{new} = \hat{\beta}_0 + \hat{\beta}_1 (X + \delta X)$$

Part 4. Back to the general diagram

Introducing a general data analytical process in practice (across domains)

A general process across modeling paradigms

Univariate Linear Regression

1. Establish the goal (DGP)
   e.g. recovering $E[Y|X]$
2. Make modeling assumptions (e.g. i.i.d.)
   e.g. $E[Y|X] = \beta_0 + \beta_1 x_i$
3. Estimate the model by minimizing an objective
   e.g. $\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - x_i\beta_1)^2$
4. Examine the performance
   e.g. R^2
5. Use the model (interpretation, prediction, etc.)
   e.g. $\hat{\beta}_0, \hat{\beta}_1$

Machine Learning

1. Establish the goal (DGP)
   e.g. generalizable $E[Y|X]$
2. Make modeling assumptions (e.g. i.i.d.)
   e.g. Neural network $E[Y|X] = f(x_i; w)$
3. Train the model by minimizing an objective
   e.g. $\underset{w}{\operatorname{argmin}}\ empirical\ loss$
4. Examine the performance
   e.g. R^2, MSE, cross-entropy, accuracy
5. Use the model (interpretation, prediction, etc.)
   e.g. $\hat{f}(x_i; w)$