DCP4300/URP6931. AI&BE

Lecture 04: Modeling Basics and Linear Regressions

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

Reminder: Use Zoom Video Recording

Announcement

1. Pset 1. Data description and visualization. Due on Oct 6th.
2. Office hour: Monday 2-3pm. I will announce it later.
3. Practicum AI materials.

Recap Lecture 03

Urban Data Description and Visualization

1. Data landscape: an example in Chicago
2. Two specific data sets: census and shapefiles.
3. Python lab 3.1: descriptive data analysis
4. Python lab 3.2: spatial data visualization
5. Python lab 3.3: census data collection and processing (skipped)

Lecture 04 - Outline

1	2	3	4
Summarizing Data (review)	What is a model?	Univariate Linear Regression	Python Lab 4 – Univariate Linear Regression

Disclaimer: To facilitate your learning, I have significantly simplified the mathematical foundations of modeling. Therefore, many statements are roughly correct, but lack the math rigor.

Sep 26: Lecture + Lab + one reading material
Oct 3: one reading material + review + survey

Part 1. Summarizing data

Two categorizations

- Categorization 1: nominal, ordinal, and cardinal numbers
- Categorization 2: discrete and continuous numbers.

Summarizing one variable

- Central tendency (mean, mode, and median)
- Variability (range, quartiles, variance, and standard deviation).

Summarizing two variables

- Covariance and correlation

1. Nominal Numbers

Nominal Numbers. For identification only; They cannot be used for ranking or algebraic operations. (e.g. driver license ID)

Nominal numbers usually take **discrete values**.

Example 1: city names

City List: Boston, Lima, Los Angeles, Minneapolis, Osaka, Cairo

Number representation: 0, 1, 2, 3, 4, 5

Example 2: travel mode choice

List: automobile, other modes

Number representation: 0, 1

How to summarize the central tendency of nominal numbers?

Mode: the most frequent result.

Example 1: city names

City List: Boston, Lima, Los Angeles, Lima, Minneapolis, Osaka, Cairo

Result: Lima.

2. Ordinal Numbers

Ordinal Numbers. For identification and ranking. But cannot be used for algebraic operations (e.g. addition/subtraction)

Ordinal numbers usually take **discrete values**.

Example 1: Housing quality rating

3, 2, 3, 4, 2, 3, 1

Ranking:

1, 2, 2, 3, 3, 3, 4

Addition:

1 + 2 = 3(?)

How to summarize the central tendency of ordinal numbers?

Median: the middle observation when everything is in order.

Example 1: Housing quality rating

3, 2, 3, 4, 2, 3, 1

Ranking: 1, 2, 2, 3, 3, 3, 4

The median is 3.

3. Cardinal Numbers

Cardinal Numbers. For identification, ranking, and algebraic operations (e.g. addition/subtraction)

Cardinal numbers can take **discrete or continuous values**.

Example 1: Life expectancy at birth

59, 59, 61, 62, 71, 71, 73

Example 2: Household average income per year

$50,101.50; $72,100.30; $101,220.00; $142,100.00;

How to summarize the central tendency of cardinal numbers?

Mean: what we normally call “the average” – it is the sum of the scores divided by the number of observations (N).

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

Example 1: Life expectancy at birth

59, 59, 61, 62, 71, 71, 73

The mean is: $\frac{59+62+71+59+73+71+61}{7}$ = 65.14

Python: df[“variable name”].mean()

Can we use the mean to summarize ordinal numbers?

Example 1: Housing quality rating

3, 2, 3, 4, 2, 3, 1

Re-order: 1, 2, 2, 3, 3, 3, 4

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

The mean is: $\frac{3+2+3+4+2+3+1}{7}$ = 2.57

What is wrong with this?

Can we use the mean to summarize nominal numbers?

Example 2: travel mode choice

List: automobile, other modes

Number representation: 0, 1

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

The mean is: $\frac{0+1}{2}$ = 0.50

What is wrong with this?

Means vs. medians: sensitivity to extreme cases

Net Worth (I made up the numbers)

Shenhao Wang: $10

Emre Tepe: $97

Ruth Steiner: $109

Zhong-Ren Peng: $121

Chimay Anumba: $200

Median: $109; Mean: $107

Means vs. medians: sensitivity to extreme cases

Net Worth (I made up the numbers)

Shenhao Wang: $10

Emre Tepe: $97

Ruth Steiner: $109

Zhong-Ren Peng: $121

Chimay Anumba: $200

Warren Buffett: $87,000,000,000

Median: $115; Mean: $14.5BN

Which one to use? Depending on the distribution and the issue at hand, you may care more about medians or means. Remember: for ordinal data, the median is really all that is meaningful.

Measures of variability: Range

The Range is simply the difference between the highest and the lowest value in the distribution.

Examples

- 67, 67, 97, 98, 99, 100, 101
- The range is: 101 - 67 = 34.
- 2.4 3.5 3.5 6.7 7.0 7.0 9.1 9.9 11.2
- The range is: 11.2 – 2.4 = 8.8

Measures of variability: Quartiles

Quartiles divide up the range into four segments with an equal number of scores in each segment.

Example: Identifying quartiles in a histogram.

Quartiles are the most common, but some people use quintiles, deciles, or percentiles. These are all called “quantiles.”

Measures of variability: Variance of a population

Variance: the average squared deviation of the scores from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Measures of variability: Standard deviation of a population

Standard deviation: simple the square root of the variance; more useful for most statistics.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Q: Why do we need the variability measure on top of the central tendency?

A: Fixed central tendency could be associated with **ANY** variability. e.g. household income

Information Richness

Discrete		Discrete		Discrete or continuous
Nominal numbers	<	Ordinal numbers	<	Cardinal numbers
Mode		Mode & Median		Mode & Median & Mean
Range/Support		Range and quantiles		Range, quantiles, variance, and standard deviation

Information Richness Indicates the Direction of Data Processing

Example: Income

[High, Low] ← [3, 2, 1] ← [100K, 50K, 30K]

Exercises

1. What are the ordinal, nominal, and cardinal variables in the census data?
2. What are the discrete and continuous variables in the census data?

Note: these categories don't have a clear-cut boundary…

	pop_total	sex_male	sex_female	households	household_size_avg	full_ct_fips	state_fips	county_fips	property_value_median	inc_median_household	travel_driving_ratio
0	2812.0	1383.0	1429.0	931.0	3.020408	12086000211	12	86	240400.0	53533.0	0.892929
1	4709.0	2272.0	2437.0	1668.0	2.823141	12086000212	12	86	179900.0	33958.0	0.879096
2	5005.0	2444.0	2561.0	1379.0	3.629442	12086000213	12	86	254900.0	40250.0	0.818343
3	6754.0	2934.0	3820.0	2238.0	3.017873	12086000214	12	86	147800.0	39962.0	0.825070
4	3021.0	1695.0	1326.0	1364.0	2.214809	12086000128	12	86	205900.0	63889.0	0.966851
...	...	...	...	...	...	...	...	...	...	...	...
4162	15742.0	7957.0	7785.0	5517.0	2.853362	12019031200	12	19	206600.0	76846.0	0.931240
4163	5723.0	2914.0	2809.0	2001.0	2.860070	12019030801	12	19	211200.0	72344.0	0.930214
4164	10342.0	4657.0	5685.0	3746.0	2.760812	12019030902	12	19	141700.0	65786.0	0.952123
4165	8960.0	4166.0	4794.0	3324.0	2.695548	12019030301	12	19	169800.0	59236.0	0.888060
4166	5083.0	2573.0	2510.0	1755.0	2.896296	12019031400	12	19	114400.0	46875.0	0.880243

4167 rows × 11 columns

Summarizing the relationship between two variables

Suppose X and Y are both continuous and cardinal variables (e.g. income and property values).

Covariance

$$Cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

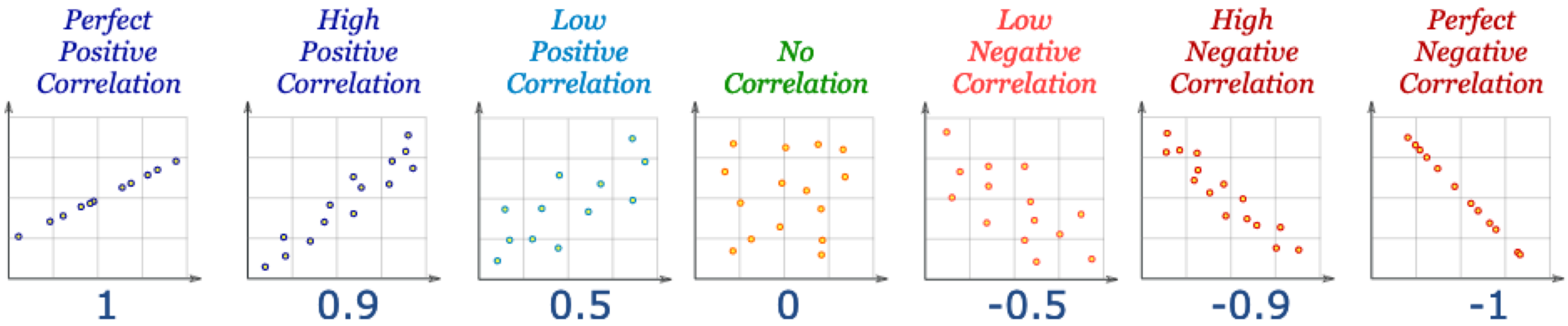
- x_i and y_i: individual observations in income and property values.
- $\bar{x}$ and $\bar{y}$: average values for x_i and y_i
- N: total number of observation.

Correlation

$$Cor_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- Correlation is a normalized covariance
- Correlation takes values between -1 and 1.
    - 1: perfect positive correlation
    - 0: no correlation
    - -1: perfect negative correlation

Visual intuition for correlation coefficients



Q: Which one looks like the relationship between property values and income?

Part 2. What is a (math) model?

And more importantly why does it matter?

Definition of a math model

Definition: a mathematical representation of some real-world phenomenon and process for analysis and prediction.

Most basic components: inputs and outputs.

Notations: Inputs: X; Outputs: Y

In modeling, both inputs and outputs are the **variables** from the spreadsheet.

Various terminology

- Inputs and outputs (CS)
- Independent and dependent variables (Stat/Econometrics)
- Exogeneous and endogenous variables (Stat/Econometrics)

Key Question in Modeling

How to choose the inputs and outputs in a model?

1. **Semi-causal effect test**. Typically the causes are the inputs and the effects are the outputs.
2. **Temporal scale test**. Typically inputs vary slowly with time, and outputs vary quickly with time.

Examples of typical inputs and outputs in urban analysis.

1. Travel behavior and income – which one is the input vs. output?
2. Car ownership and household size - which one is the input vs. output?
3. Rent and income – which one is the input vs. output?

However, sometimes the causal direction can be confusing…e.g. education & income.

Notes: Typically we use only one output variable but many input variables in a model. The only exception is univariate linear regression.

Categories of math models

Probabilistic nature: deterministic vs. stochastic

- **Deterministic model**: everything is predefined without uncertainty. e.g. classical physics.
- **Stochastic model**: examining the random variation by using probability distributions. e.g. prediction models.

Temporal dimension: static vs. dynamic

- **Static models**: describe the static structure of a system, which does not vary with the functions of the system. e.g. regression models.
- **Dynamic models**: describe the time-dependent aspects of a system with temporal changes. e.g. dynamic behavioral models.

Other categorization

- Fields: engineering, economic, or urban models.
- Knowledge: black or white box models.
- etc.

Why does a (math) model matter?

A model can formalize and generalize your intuition in an efficient manner.

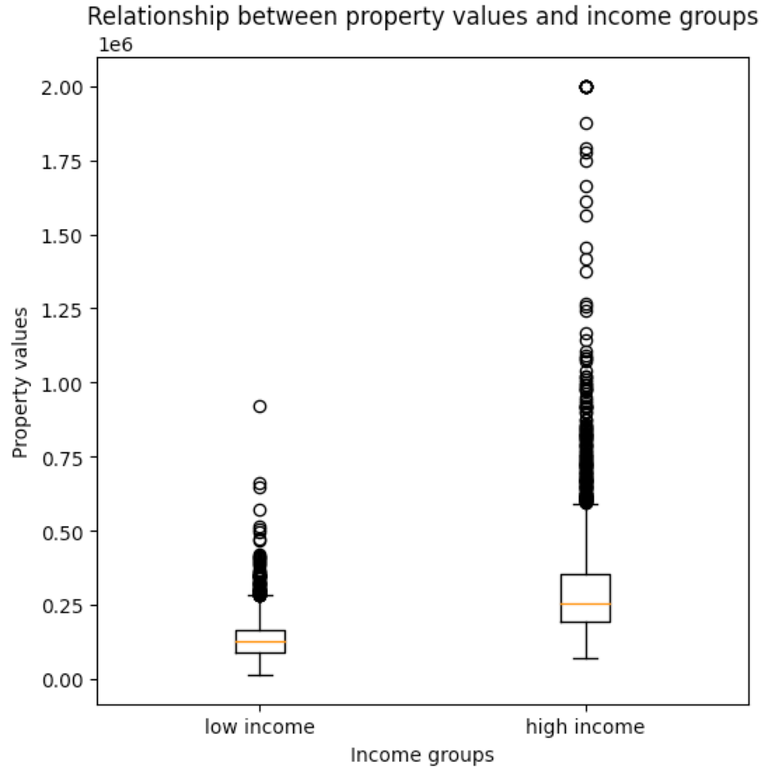
Hard to explain in an abstract matter, but let's see univariate linear regression.

Part 3. Univariate Linear Regression

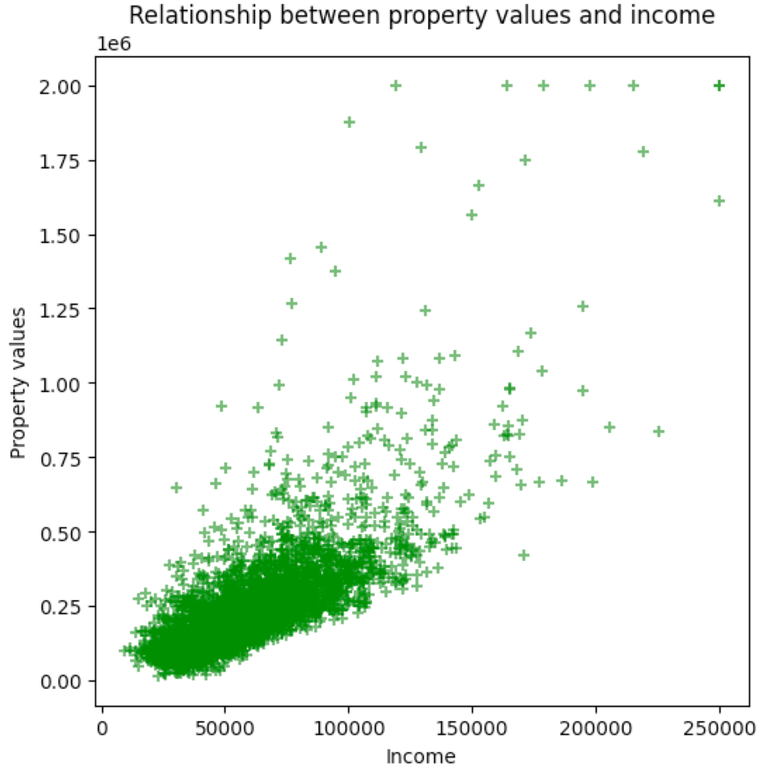
1. Univariate linear regression is a model (potentially the simplest one)
2. It formalizes and generalizes our intuition.
3. It efficiently summarize key information into two numbers

Visualizing the relationship between property values and income

Q1: What is the property value gap between the low-income and high-income groups? Is this gap significant?
Q2: How do property values vary marginally with $1,000 increase in income?



Y and discrete X



Y and continuous X

Univariate Linear Regression

The model 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 seek to summarize the relationship between two variables using a straight line.

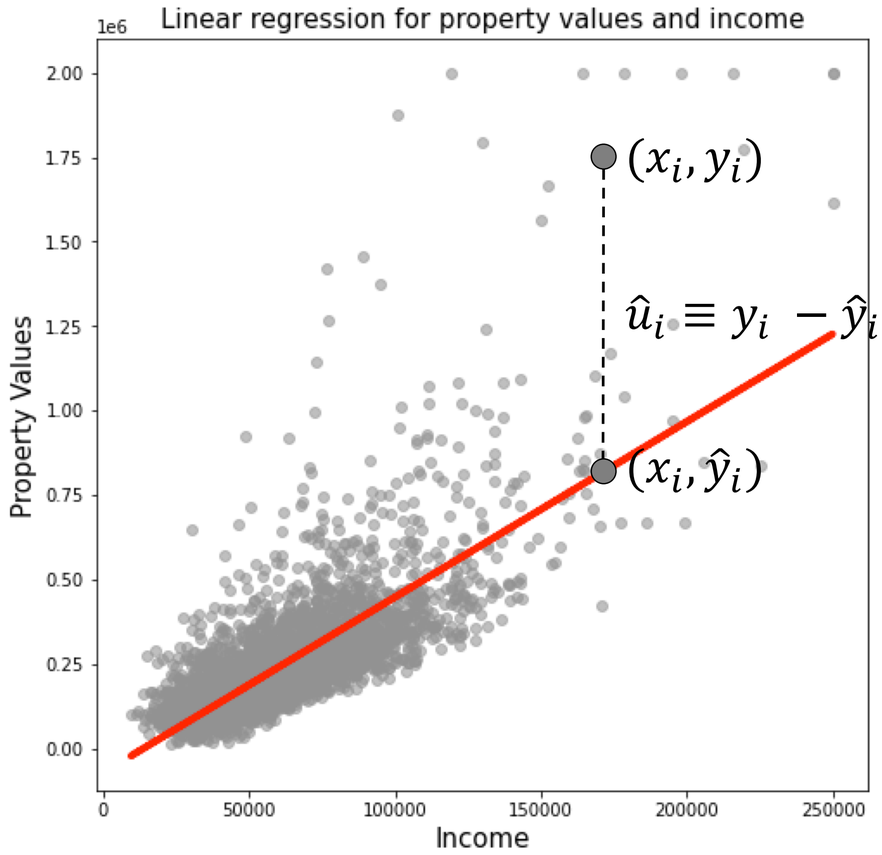
Motivating example: fitting property values and household income

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

A: intuition – you want to position a straight line as "close" as possible to all the points.

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: 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”

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

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

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

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

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

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?

Answer: at the entry level, Yes – the higher the better. In machine learning, the answer depends.

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.)

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

Univariate linear regression

A model can formalize and generalize your intuition in an efficient manner.

1. Formalize: mathematical modeling with a linear form with statistically meaningful coefficients.
2. Efficient: $\hat{\beta}_0$ and $\hat{\beta}_1$ capture the most essential information from a large number of Y & X observations.
3. Generalize: you can apply the univariate linear regression to ANY pair of valid input & output.

A model for any pair of inputs (continuous X) and outputs

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.

A model for any pair of inputs (discrete X) and outputs

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: rent; X: income (high, low).

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

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

However,

Valid challenges do exist

1. Property values also depend on household sizes
2. Property values also depend on education levels
3. Property values also depend on age
4. etc.

The omitting variable bias is quite significant in univariate linear regression!