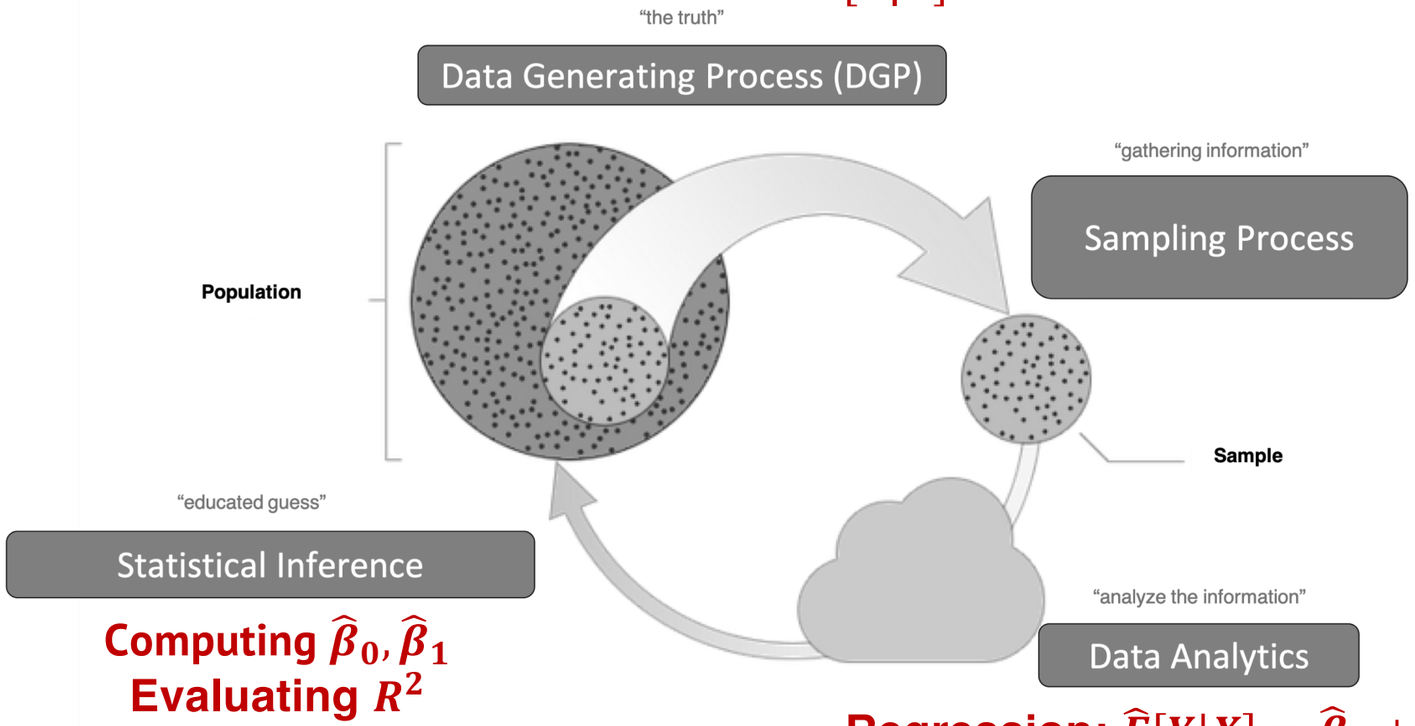
URP 6931. Introduction to Urban Analytics

Lecture 04: Multivariate linear regression

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

Review Lecture 03

Regression: recovering $E[Y|X]$
Q: Intuition about $E[Y|X]$?



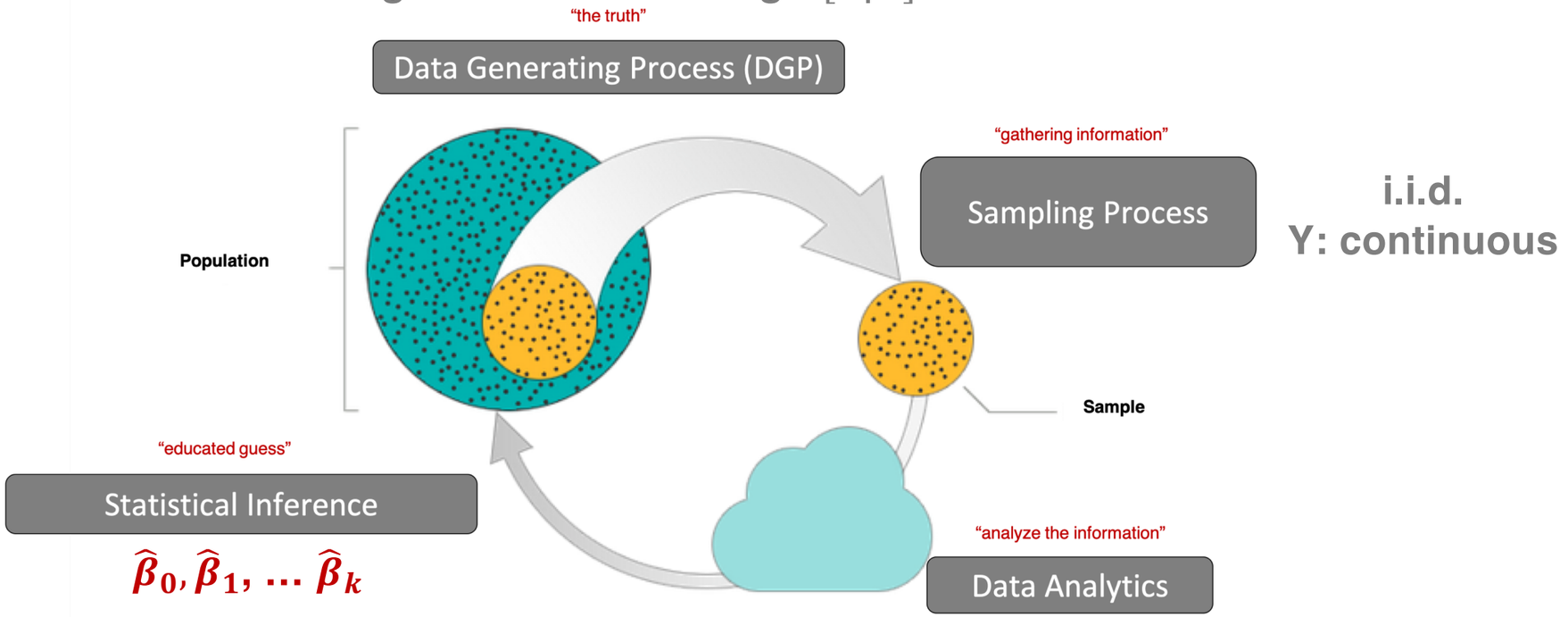
i.i.d.
Y: continuous

Computing $\hat{\beta}_0, \hat{\beta}_1$
Evaluating R^2

Regression: $\hat{E}[Y|X] = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$

Preview Lecture 04

Regression: recovering $E[Y|X]$



i.i.d.
Y: continuous

Regression: $\hat{E}[Y|X] = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$

Lecture 04. multivariate linear regression

1	2	3	4	5
Regression with multiple independent variables	Regression assumptions and their consequences	Statistical properties of the OLS estimators in theory (not required)	Example in practice	One-page idea note

Part 1. Regression with multiple independent variables

Why do we want more than one predictor?

1. Summarize more information for descriptive inference
2. Improve the fit and predictive power of our model
3. Control for confounding factors for causal inference
4. Analyze more complex non-linearities (e.g. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$)
5. Incorporate more information (e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$)
6. Describe interactive effects (e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$)

Terminology - estimators

- $E[Y|X_1, X_2, \ldots X_k]$ is the conditional mean function of interest (DGP)
- $\hat{E}[Y|X_1, X_2, \ldots X_k]$ is the estimator (Model)

In linear regressions, this estimator is a function with the following geometric intuition:

- A line in cases with a single X_1
- A plane in cases with two independent variables (X_1 and X_2)
- A hyperplane in cases with more than two variables

We start with two cases

- Regression with one continuous (X_1) and one dummy variable (X_2).
- Regression with two continuous variables (X_1 and X_2)

Florida census data as the leading example

Variables

- Y: property value
- X_1: household income
- X_2: ratio of people with higher education

Old question: Does income (X_1) predict or explain the level of property values (Y)?

New question: Does income (X_1) predict or explain the level of property values (Y), once we "control" for education effects?

What is the meaning of "controlling for another variable"?

Property value ~ Income

We previously looked at the regression of property values on income.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

Estimation results

- $\hat{\beta}_0 = -\$71{,}700$
- $\hat{\beta}_1 = 5.19$

Question: What is the interpretation of $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$?

Interpretation: one unit increase in household income is associated with 5.19 increase in the property value.

But we might consider another factor: education.

- Higher education leads to more investment interests in properties.

Adding a variable: property value ~ income + education

How to add education?

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

This is a simple example of predicting the property values using both income and education with a linear functional form.

Notice that we write x_{ji} where:

- $j = 1, \ldots, J$ is the index for the explanatory variables.
- $i = 1, \ldots, N$ is the index for observations.

Sometimes I omit i to simplify the notation

Suppose x_{2i} is binary variable {0, 1},

How to interpret the model?

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

- When $x_2 = 0$, the model becomes

$$\hat{y}_i(x_{2i} = 0) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

- When $x_2 = 1$, the model becomes

$$\hat{y}_i(x_{2i} = 1) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2$$

Therefore,

- $\hat{\beta}_2$ measures the difference of property values between high and low education groups.
- Meanwhile, $\hat{\beta}_1$ measures the slope of y regarding x_1, which is the same for both education groups.
- Essentially, we are fitting two lines with the same slope but different intercepts.

What is the new result?

Property values ~ income and education (binary)

Education (binary): high vs. low education groups

Estimation results

- $\hat{\beta}_0 = -\$59{,}760$
- $\hat{\beta}_1 = 4.70$
- $\hat{\beta}_2 = \$44{,}930$

Interpretation

- $\hat{\beta}_0$ is the intercept for the low education group's property value at income = 0.
- $\hat{\beta}_1$ is the slope for both high and low education groups.
- $\hat{\beta}_2$ is the vertical distance between the two lines.

Question: the value of $\hat{\beta}_1$ changes in this regression, so should we trust the old or the new value?

Suppose x_{2i} is a continuous variable

We interpret $\hat{\beta}_1$ and $\hat{\beta}_2$ as partial effects

$$\frac{\partial \hat{y}_i}{\partial x_{1i}} = \frac{\partial \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}}{\partial x_{1i}} = \hat{\beta}_1$$

$$\frac{\partial \hat{y}_i}{\partial x_{2i}} = \frac{\partial \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}}{\partial x_{2i}} = \hat{\beta}_2$$

$\hat{\beta}_1$ measures the slope of y_i regarding x_{1i}, which is the same for any x_{2i}

$\hat{\beta}_2$ measures the slope of y_i regarding x_{2i}, which is the same for any x_{1i}

What is the new result?

Property values ~ income and education (continuous)

Education (continuous): ratio of people with higher education in a census tract

Estimation results

- $\hat{\beta}_0 = -\$81{,}870$
- $\hat{\beta}_1 = 4.08$
- $\hat{\beta}_2 = \$279{,}200$

Interpretation

- $\hat{\beta}_1$ is the slope of property value regarding household income for any education level
- $\hat{\beta}_2$ is the slope of property value regarding education for any income level
- "controlling for another variable" – e.g. $\hat{\beta}_1$: holding x_2 constant (at any value), one unit increase in x_1 is associated by $\hat{\beta}_1$ unit increase in y.

From two variables to multiple variables

In practice, people typically use more than two variables:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ ... + \hat{\beta}_k x_{ki}$$

How to interpret the model? The partial effect and the intuition about the slope controlling for other variables still hold true.

Using x_{ji} as an example,

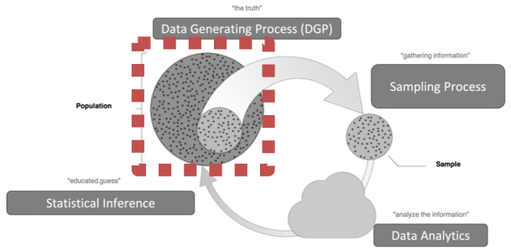
$$\frac{\partial \hat{y}_i}{\partial x_{1i}} = \frac{\partial \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ ... + \hat{\beta}_k x_{ki}}{\partial x_{ji}} = \hat{\beta}_j$$

- One unit increase in x_{ji} is associated with $\hat{\beta}_j$ unit increase in $\hat{y}_i$, controlling for all the other variables.
- e.g. one unit increase in income is associated with $\hat{\beta}_1$ unit increase in property values, regardless of the other control variables (education, gender, age, etc.)

Part 2. Regression assumptions and their consequences

- Three layers of consequences

OLS Terminology (last lecture)



The population model (DGP) is given by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ \ldots \beta_k x_{ki} + u$$

- y_i : Dependent variable.
- $x_{1i}, x_{2i}, \ldots, x_{ki}$: Independent variable.
- $\beta_0, \beta_1, \ldots, \beta_k$: 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.

What are the assumptions?

Assumption 1. **No omitted variables** - zero conditional mean $E[u|x_{1i}, \dots x_{ki}] = 0$.

a.k.a. no confounding factor.

Assumption 2. **Linearity in Parameters -** the population model is linear in parameters and correctly specified.

a.k.a. correct specification assumption.

Question: Is this a linear regression $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2$?

Assumption 3. **Independent and identically distributed random sampling** – the observed data represent a random sample from the population described by the model.

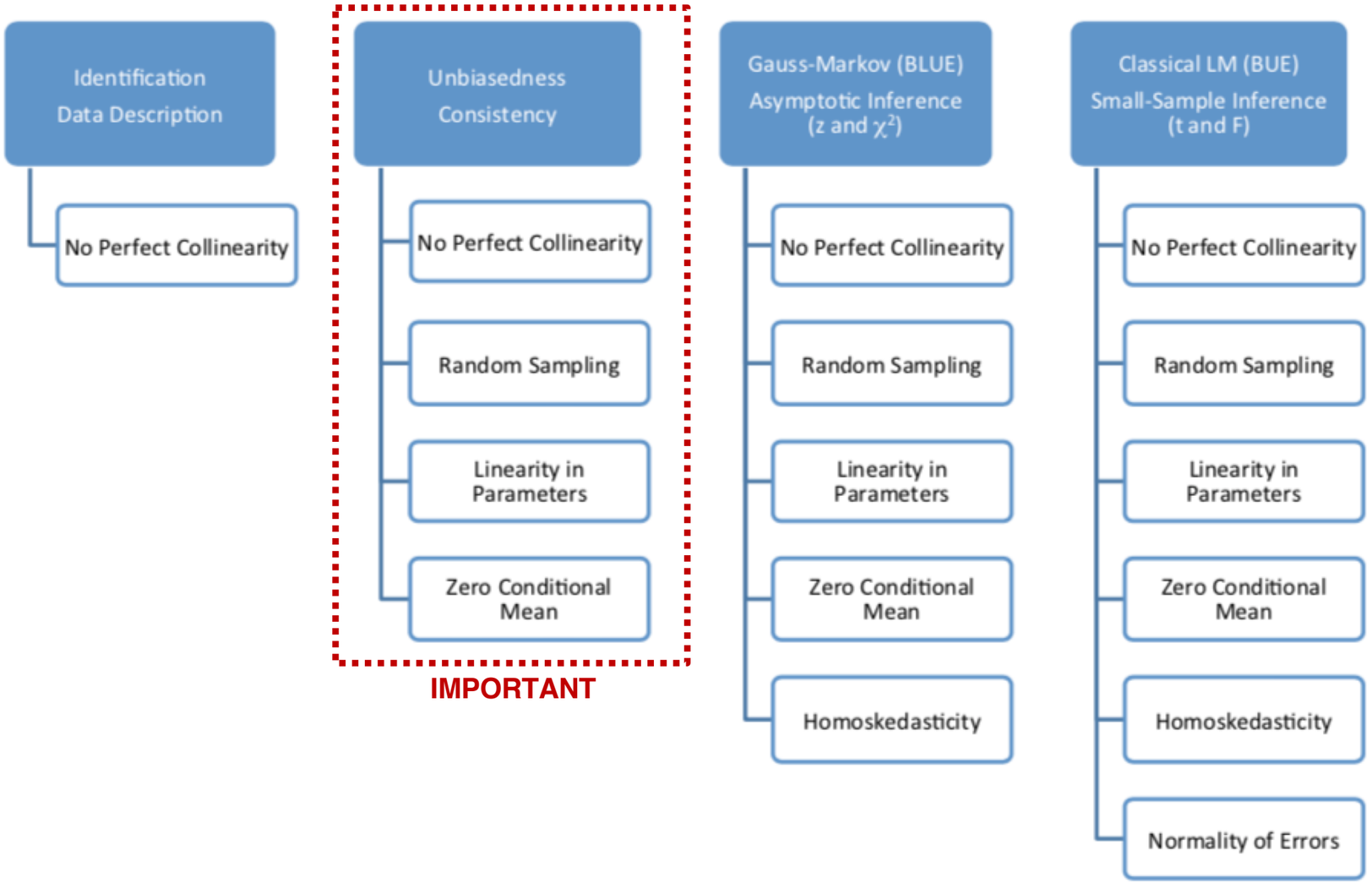
Assumption 4. No perfect collinearity. No predictor is a linear combination of other predictors.

Assumption 5. Homoskedasticity: The error term u has the same variance conditioning on all values of the explanatory variables.

Assumption 6. Normality: The error term is independent of the explanatory variables and normally distributed

The importance of the assumptions is ranked in order.

Layer 1. Consequences of satisfying the assumptions



Layer 2. Consequences of violating the assumptions in theory

Assumption 1. No omitted variables

– Biased and inconsistent estimates

An example of biased and inconsistent estimates:

- **Truth (DGP)**: one unit income increase leads to two unit increase in property value.
- **Estimate**: one unit income increase leads to five unit increase in property value.

Translation: the estimate could be **ANYTHING**.

Simple example for Assumption 1

Case 1: simple DGP & richer model.

DGP: $\mathrm{y_i} = \beta_0 + \beta_1 x_{1i} + u_i$

Model: $\mathrm{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i$

Question: What are the consequences?

Case 2: rich DGP & simple model - violating A1

DGP: $\mathrm{y_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$

Model: $\mathrm{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{u}_i$

Question: What are the consequences?

Answer: you will obtain a biased $\hat{\beta}_1$ (i.e., any value is possible), which combines the true β_1 and the β_2 working through the correlation between x_{1i} and x_{2i}.

Consequence in practice: People always like to use nearly all possible explanatory variables in the data set, thus avoiding the violation of A1.

Layer 2. Consequences of violating the assumptions in theory

Assumption 2. Linearity in Parameters

– Biased and inconsistent estimates

An example of biased and inconsistent estimates:

- **Truth**: one unit income increase leads to two unit increase in property value.
- **Estimate**: one unit income increase leads to five unit increase in property value.

Translation: the estimate could be **ANYTHING**.

Simple example for Assumption 2

Case 1: simple DGP & richer model.

DGP: $\mathrm{y_i} = \beta_0 + \beta_1 x_{1i} + u_i$

Model: $\mathrm{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{1i}^2 + \hat{u}_i$

Question: What are the consequences?

Case 2: rich DGP & simple model - violating A2

DGP: $\mathrm{y_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + u_i$

Model: $\mathrm{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{u}_i$

Question: What are the consequences?

Answer: you will obtain a biased $\hat{\beta}_1$ (i.e., any value is possible), which combines the true linear effect β_1 and the β_2 working through the correlation between x_{1i} and x_{1i}^2.

Consequence in practice: People always like to specify a richer model, thus avoiding the violation of A2.

A1 vs. A2. A2 is less a problem because you can change the model to address A2.

Layer 2. Consequences of violating the assumptions in theory

Assumption 3. **Independent and identically distributed random sampling**

– Potentially biased and inconsistent estimates

Assumption 4. No perfect collinearity – estimation cannot be completed.

Assumption 5. Homoskedasticity - not critical (in large sample).

Assumption 6. Normality – not critical (in large sample).

Layer 3. How to satisfy the assumptions in practice

Assumption 1. **No omitted variables**

- We should collect as many variables as possible.
- However, critiques are often **silent** about it.

Assumption 2. **Linearity in parameters**

- We should design relatively rich models.
- **Question**: Are linear models naive models?
- **Answer**: Linear-in-parameter models are extremely powerful models (e.g. power series as "universal approximator")

Assumption 3. **Independent and identically distributed random sampling**

- Carefully design the sampling process.
- Advanced methods can partially mitigate this problem.

Assumption 4. No perfect collinearity

- Data preprocessing: simply remove the perfectly correlated column.
- **Question**: How about imperfect multicollinearity of independent variables (e.g. correlation = 0.8)?

Assumption 5. Homoskedasticity - not critical.

Assumption 6. Normality – not critical.

Regression as causal inference

Question: Can linear regression also be used for causal inference? e.g. one unit increase in income CAUSES the increase in property value by five unit?

Answer: Yes and No. (Four-layers argument)

1. No is a naïve answer – an old cliché is: “correlation is not causality”.

2. Yes is a qualified answer after scientific training.

$\hat{\beta}_1$ could be interpreted as causal effect of X on Y under two unrealistic assumptions:

- A1. No confounding factor influences both X and Y.
- A2. The linear-in-parameter function is correct.

3. No is a safe answer, because it is challenging to satisfy both assumptions in the framework of linear regressions.

4. Provide a constructive answer, e.g. DID, regression discontinuity, etc.

Suggestion in critiques: provide at least some points 2-4 in your critiques.

Part 3. Statistical properties of the OLS estimators in theory (not required)

1. The properties are critical to know in practice, so I cannot avoid them.
2. You should take statistical classes to fully understand this part.

Assumptions underlying statistical analysis

Assumption 1. No omitted variables

Assumption 2. Linearity in Parameters

Assumption 3. Independent and identically distributed random sampling

Assumption 4. No perfect collinearity

Assumption 5. Homoskedasticity

Assumption 6. Normality

Notes

1. When the six (particularly the first two) assumptions hold, we have statistical properties for formal analysis (Section 3 in today's lecture)
2. But as I suggested, it is tricky to challenge A1-A6. Hence people are often cheap in practice (Section 4 in today's lecture)

Statistical properties

Theorem

Under Assumptions 1-6, the (k + 1)*1 vector of OLS estimators $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, follows a multivariate normal distribution with mean β and variance-covariance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

- It indicates that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Intuitively, it is still a normalized correlation between x and y.
- Each element of $\widehat{\boldsymbol{\beta}}$ (i.e. $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$) is normally distributed, and $\hat{\beta}$ is an unbiased estimator of β as $\mathrm{E}[\hat{\beta}] = \beta$.
- Variance and covariances are given by $V[\hat{\beta}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- An unbiased estimator for the error variance σ^2 is given by

$$\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}} = \frac{\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}}{\boldsymbol{n} - (\boldsymbol{k} + \boldsymbol{1})}$$

Statistical properties

Theorem

Under Assumptions 1-6, the (k + 1)*1 vector of OLS estimators $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, follows a multivariate normal distribution with mean β and variance-covariance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

The variance-covariance matrix of the OLS estimators is:

$$V[\hat{\beta}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} =$$

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\cdots$	$\widehat{\beta}_k$
$\widehat{\beta}_0$	$V[\widehat{\beta}_0]$	$\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\cdots$	$\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_k]$
$\widehat{\beta}_1$	$\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$V[\widehat{\beta}_1]$	$\mathrm{Cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\cdots$	$\mathrm{Cov}[\widehat{\beta}_1, \widehat{\beta}_k]$
$\widehat{\beta}_2$	$\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\mathrm{Cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$V[\widehat{\beta}_2]$	$\cdots$	$\mathrm{Cov}[\widehat{\beta}_2, \widehat{\beta}_k]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\widehat{\beta}_k$	$\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_k]$	$\mathrm{Cov}[\widehat{\beta}_k, \widehat{\beta}_1]$	$\mathrm{Cov}[\widehat{\beta}_k, \widehat{\beta}_2]$	$\cdots$	$V[\widehat{\beta}_k]$

t-value for multivariate linear regression

Theorem

Given Assumptions I-V, the OLS estimator $\hat{\beta}_j$ is asymptotically normally distributed:

$$T = \frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]} \sim N(0,1)$$

Where

$$SE[\hat{\beta}_j] = \sqrt{\hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}}$$

What matters here?

- We can obtain a distribution of $\hat{\beta}_j$ for statistical test.
- We can test the $\hat{\beta}_j$ one by one.
- We can use Python (or R, Stata, etc.) to compute this distribution.

Using the t-value as a test statistic

Null hypothesis (H_0): $\beta_j = c$. (typically c = 0).

- Compute the t–value as $T = \frac{\widehat{\beta}_j - c}{SE[\widehat{\beta}_j]}$
- Compare the t–value to the critical value $t_{\alpha/2}$ (typically 2) for the α level test (typically $\alpha = 0.05$; then confidence level is 0.95), which under the null hypothesis satisfies

$$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = 1 - \alpha$$

- Decide whether the realized value of T is unusually large given the known distribution of the test statistic.
- Finally, either declare that we reject H_0 or not, or report the p-value.
- P-value. It measures the probability of observing a t-value at least as extreme as one we observe assuming that H_0 is true.

Part 4. Example in practice

1. Process of enriching the model. (consequences of the assumptions)
2. Process of making statistical analysis

state-of-the-practice in urban applications

Statement: “With 95% level of confidence, we observe a statistically significant relationship between y and x.”

Find the large t-value ($t > 2$)

Find the **small p-value ($p < 0.05$)**

First Regression: property value ~ income

OLS Regression Results

Dep. Variable:	property_value_median	R-squared:	0.568
Model:	OLS	Adj. R-squared:	0.568
Method:	Least Squares	F-statistic:	5473.
Date:	Thu, 26 Jan 2023	Prob (F-statistic):	0.00
Time:	01:55:16	Log-Likelihood:	-54606.
No. Observations:	4167	AIC:	1.092e+05
Df Residuals:	4165	BIC:	1.092e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.17e+04	4572.103	-15.681	0.000	-8.07e+04	-6.27e+04
inc_median_household	5.1931	0.070	73.980	0.000	5.056	5.331

Omnibus:	3603.056	Durbin-Watson:	1.284
Prob(Omnibus):	0.000	Jarque-Bera (JB):	172406.098
Skew:	3.901	Prob(JB):	0.00
Kurtosis:	33.530	Cond. No.	1.62e+05

Process

1. Check R Square.
2. Check coefficients
3. Check t and p values.
4. Interpret coefficients.
5. Enrich the model

Second Regression: property value ~ income + households

OLS Regression Results

Dep. Variable:	property_value_median	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.574
Method:	Least Squares	F-statistic:	2812.
Date:	Thu, 26 Jan 2023	Prob (F-statistic):	0.00
Time:	02:51:33	Log-Likelihood:	-54573.
No. Observations:	4167	AIC:	1.092e+05
Df Residuals:	4164	BIC:	1.092e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.751e+04	5430.518	-8.749	0.000	-5.82e+04	-3.69e+04
inc_median_household	5.2339	0.070	74.942	0.000	5.097	5.371
households	-14.3360	1.769	-8.104	0.000	-17.804	-10.868

Omnibus:	3524.983	Durbin-Watson:	1.302
Prob(Omnibus):	0.000	Jarque-Bera (JB):	160647.295
Skew:	3.783	Prob(JB):	0.00
Kurtosis:	32.462	Cond. No.	1.94e+05

Process

1. Check R Square.
2. Check coefficients
3. Check t and p values.
4. Interpret coefficients.
5. Enrich the model

Third Regression: property value ~ income + others

OLS Regression Results

Dep. Variable:	property_value_median	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.680
Method:	Least Squares	F-statistic:	680.8
Date:	Thu, 26 Jan 2023	Prob (F-statistic):	0.00
Time:	02:51:35	Log-Likelihood:	-53976.
No. Observations:	4167	AIC:	1.080e+05
Df Residuals:	4153	BIC:	1.081e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.625e+05	3.46e+04	4.690	0.000	9.46e+04	2.3e+05
inc_median_household	4.4778	0.109	41.158	0.000	4.265	4.691
households	-5.1170	1.593	-3.212	0.001	-8.240	-1.994
travel_driving_ratio	-3.369e+05	3.47e+04	-9.701	0.000	-4.05e+05	-2.69e+05
travel_pt_ratio	4.119e+05	6.72e+04	6.131	0.000	2.8e+05	5.44e+05
travel_taxi_ratio	7.753e+05	2.2e+05	3.518	0.000	3.43e+05	1.21e+06
travel_work_home_ratio	-3.548e+04	4.94e+04	-0.718	0.473	-1.32e+05	6.14e+04
edu_higher_edu_ratio	2.182e+05	2.58e+04	8.468	0.000	1.68e+05	2.69e+05
edu_dummy	-1.753e+04	6021.500	-2.911	0.004	-2.93e+04	-5720.300
household_size_avg	-38.4908	11.568	-3.327	0.001	-61.170	-15.812
vacancy_ratio	1.651e+05	1.43e+04	11.542	0.000	1.37e+05	1.93e+05
rent_median	22.6483	5.355	4.229	0.000	12.149	33.147
race_white_ratio	1.157e+04	9415.728	1.229	0.219	-6892.050	3e+04
race_asian_ratio	-5.866e+05	5.96e+04	-9.842	0.000	-7.03e+05	-4.7e+05

Omnibus:	3427.814	Durbin-Watson:	1.537
Prob(Omnibus):	0.000	Jarque-Bera (JB):	170907.189
Skew:	3.578	Prob(JB):	0.00
Kurtosis:	33.547	Cond. No.	9.08e+06

Process

1. Check R Square.
2. Check coefficients
3. Check t and p values.
4. Interpret coefficients.
5. Enrich the model

Fourth Regression property value ~ income + $\mathbf{income}^2$ + others

OLS Regression Results

Dep. Variable:	property_value_median	R-squared:	0.700
Model:	OLS	Adj. R-squared:	0.699
Method:	Least Squares	F-statistic:	693.5
Date:	Thu, 26 Jan 2023	Prob (F-statistic):	0.00
Time:	02:51:38	Log-Likelihood:	-53842.
No. Observations:	4167	AIC:	1.077e+05
Df Residuals:	4152	BIC:	1.078e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.014e+05	3.36e+04	5.987	0.000	1.35e+05	2.67e+05
inc_median_household	0.8502	0.243	3.502	0.000	0.374	1.326
inc_median_household_squared	1.938e-05	1.17e-06	16.587	0.000	1.71e-05	2.17e-05
households	-3.1154	1.548	-2.013	0.044	-6.149	-0.081
travel_driving_ratio	-2.747e+05	3.38e+04	-8.117	0.000	-3.41e+05	-2.08e+05
travel_pt_ratio	3.035e+05	6.54e+04	4.641	0.000	1.75e+05	4.32e+05
travel_taxi_ratio	7.704e+05	2.13e+05	3.609	0.000	3.52e+05	1.19e+06
travel_work_home_ratio	-1.482e+04	4.79e+04	-0.310	0.757	-1.09e+05	7.9e+04
edu_higher_edu_ratio	2.564e+05	2.51e+04	10.228	0.000	2.07e+05	3.06e+05
edu_dummy	-7530.3750	5863.156	-1.284	0.199	-1.9e+04	3964.551
household_size_avg	-37.0006	11.204	-3.302	0.001	-58.967	-15.034
vacancy_ratio	1.539e+05	1.39e+04	11.094	0.000	1.27e+05	1.81e+05
rent_median	31.0684	5.212	5.961	0.000	20.851	41.286
race_white_ratio	2.978e+04	9185.455	3.242	0.001	1.18e+04	4.78e+04
race_asian_ratio	-5.482e+05	5.78e+04	-9.488	0.000	-6.61e+05	-4.35e+05

Omnibus:	3275.733	Durbin-Watson:	1.554
Prob(Omnibus):	0.000	Jarque-Bera (JB):	171668.909
Skew:	3.305	Prob(JB):	0.00
Kurtosis:	33.742	Cond. No.	8.63e+11

Process

1. Check R Square.
2. Check coefficients
3. Check t and p values.
4. Interpret coefficients.
5. Write a report.

e.g. “With 95% level of confidence, we observe a statistically significant relationship between medium property value and income, after controlling for all the other variables. The controlling variables include travel, household size, and racial composition. This relationship is quadratic, etc.”

Part 5. One-page idea note

Deadline: **Feb 14** (not Feb 7) – I will update the syllabus and post the assignment today/tomorrow.

Team: 2~3 persons

Page limit: 1 page (excluding references).

Five sections (5 pts)

1. Research question and significance.
2. Background literature (> 5 references)
3. Data (preliminary data access or access plan)
4. Analytical approach (ideally focusing on only one modeling paradigm)
5. Anticipated results.