DCP4300/URP6931. AI&BE

Lecture 05: Urban Application 1 and Multivariate 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

Outline

1. Recap and extension of lecture 04
2. Urban application 1: Gasoline consumption and cities
3. Multivariate linear regression
4. Statistical analysis in urban practices (not theory)
5. Python Lab 05: Multivariate linear regressions

Part 1. Recap and Extension of Lecture 04

1. Data summary
2. Univariate linear regression

Summarizing Data

Q: Why does data summary matter?

A: Without data summary, you cannot generate succinct insights from the data.

	pop_total	sex_male_ratio	households	household_size_avg	full_ct_fips	state_fips	property_value_median	inc_median_household	travel_driving_ratio
0	2812.00	0.49	931.00	3.02	12086000211	12	240400.00	53533.00	0.89
1	4709.00	0.48	1668.00	2.82	12086000212	12	179900.00	33958.00	0.88
2	5005.00	0.49	1379.00	3.63	12086000213	12	254900.00	40250.00	0.82
3	6754.00	0.43	2238.00	3.02	12086000214	12	147800.00	39962.00	0.83
4	3021.00	0.56	1364.00	2.21	12086000128	12	205900.00	63889.00	0.97
...	...	...	...	...	...	...	...	...	...
4162	15742.00	0.51	5517.00	2.85	12019031200	12	206600.00	76846.00	0.93
4163	5723.00	0.51	2001.00	2.86	12019030801	12	211200.00	72344.00	0.93
4164	10342.00	0.45	3746.00	2.76	12019030902	12	141700.00	65786.00	0.95
4165	8960.00	0.46	3324.00	2.70	12019030301	12	169800.00	59236.00	0.89
4166	5083.00	0.51	1755.00	2.90	12019031400	12	114400.00	46875.00	0.88

4167 rows × 9 columns

Summarizing Data

- A simple syntax: **dataframe.describe()** – See today's lab.
- With data description, you can generate succinct insights from the data by checking the mean, standard deviations, and quartiles. Example: ratio of males, average household incomes, etc.
- With data description, you can also observe the unreasonable patterns or mistakes.

- **Q: What can you observe as problematic in the data set from this data summary table?**
  - Check the mean vs. standard deviation vs. quantiles
  - Check the two categorizations of numbers.

	pop_total	sex_male_ratio	households	household_size_avg	full_ct_fips	state_fips	property_value_median	inc_median_household	travel_driving_ratio
count	4167.00	4167.00	4167.00	4167.00	4167.00	4167.00	4167.00	4167.00	4167.00
mean	5015.99	0.49	1856.57	8.26	12071695284.50	12.00	237906.52	59617.55	0.87
std	2984.86	0.05	1035.62	153.71	36186548.43	0.00	180782.60	26232.84	0.09
min	11.00	0.00	1.00	1.27	12001000200.00	12.00	15800.00	9463.00	0.00
25%	3132.50	0.46	1229.00	2.28	12039020101.50	12.00	134250.00	41647.00	0.84
50%	4519.00	0.49	1699.00	2.64	12086001402.00	12.00	195200.00	54140.00	0.89
75%	6226.00	0.51	2272.50	3.03	12099007759.50	12.00	284450.00	72027.50	0.93
max	39928.00	1.00	21209.00	6876.00	12133970303.00	12.00	2000001.00	250001.00	1.00

From correlation to **correlation matrix**

Correlation takes values between -1 and 1.

- 1: perfect positive correlation
- 0: no correlation
- -1: perfect negative correlation

Example: the correlation between property values and income?

Correlation matrix

- Describe the correlations between two sets of variables
- Examining one row/column in this matrix. E.g. property value
- An example in today's reading material: gasoline consumption correlating with other variables

	pop_total	sex_male_ratio	households	household_size_avg	property_value_median	inc_median_household	travel_driving_ratio
pop_total	1.00	-0.00	0.92	-0.00	-0.05	0.09	0.18
sex_male_ratio	-0.00	1.00	-0.09	0.13	0.01	0.03	-0.13
households	0.92	-0.09	1.00	-0.06	-0.03	0.07	0.14
household_size_avg	-0.00	0.13	-0.06	1.00	-0.03	-0.04	-0.25
property_value_median	-0.05	0.01	-0.03	-0.03	1.00	0.75	-0.34
inc_median_household	0.09	0.03	0.07	-0.04	0.75	1.00	-0.09
travel_driving_ratio	0.18	-0.13	0.14	-0.25	-0.34	-0.09	1.00

Univariate Linear Regression between Y and X

Univariate Linear Regression

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

- Y: Outputs (dependent variable)
- X: Inputs (independent variable)
- β_0, β_1: Intercept and slope.

Q: What is the reasoning behind choosing inputs and outputs?

Interpretation

- $\hat{\beta}_1$ measures how much Y changes when we change X by one unit. This is a normalized correlation coefficient.
- $\hat{\beta}_0$ measures the intercept of Y value when X = 0

Evaluation

- R^2: percent total variation in Y that is explained by X.

Why modeling?

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

Urban Application 1

Part 2. Gasoline Consumption and Cities — A Comparison of U.S. Cities with a Global Survey

Peter W. G. Newman & Jeffrey R. Kenworthy
1989, Journal of the American Planning Association

Leading Question: How did the authors use the skills in the AI&BE course to answer their question?

Reading Questions

1. What are the major debates the authors seek to resolve in this study?
2. What does the data set look like? Particularly what are the rows and columns?
3. What are the major modeling tools in this study?
4. What are the inputs and outputs? How many inputs and outputs are there? Why do these variables constitute a relatively valid input-output relationship?
5. What are the coding skills you could use to conduct this study?

Research Question

Evaluate how physical planning can conserve transportation energy in urban areas

Sub-Question 1: how physical planning policies can conserve transportation energy in U.S. urban areas?

Sub-Question 2: how physical planning policies can conserve transportation energy in global urban areas?

Sub-Question 3: Comparing socioeconomic backgrounds, economic factors, versus physical planning factors for gasoline consumption.

Q1: What are the major debates the authors seek to resolve in this study?

A: Sub-Question 3. In fact, it is a "soul-searching" question towards all the designers and planners.

Literature Review

Reference Group 1

Framing

Average gasoline consumption in U.S. cities was nearly twice as high as in Australian cities, four times higher than in European cities and ten times higher than in Asian cities.

Research Gap

Allowing for variations in gasoline price, income, and vehicle efficiency explains only half of these differences**.**

Literature Review

Reference Group 2

Framing

- Policies to reduce oil consumption in the United States have focused on only stationary uses (e.g., industry and home heating) and vehicle fuel efficiency.
- Policy studies on gasoline consumption derived mainly from simulation studies.

Research Gap

1. Studies rarely focus on cities, and often suggest that only minimal energy savings would result from greater use of transit and land use changes.

2. Studies tend to have a limited data base because urban energy statistics are generally unavailable.

Research Contexts

32 worldwide cities

1. North American cities: Houston, Phoenix, Detroit, Denver, Los Angeles, San Francisco, Boston, Washington DC, Chicago, New York, Toronto
2. European cities: Hamburg, Frankfurt, London, West Berlin, Vienna, Paris, Stockholm, Zurich, Munich, Amsterdam, Copenhagen, Brussels, Moscow
3. Asian cities: Tokyo, Hongkong, Singapore
4. Australian cities: Perth, Sydney, Brisbane, Adelaide, Melbourne

Year: 1960s, 1970s, and 1980s

Dataset

Collected over a five-year period primarily by **visiting each city, sometimes twice**, with considerable follow-up correspondence.

Some specifics

1. Gasoline consumption was verified by using vehicle miles of travel (VMTs) for each city and national vehicle fuel efficiency data adjusted for average speed.
2. The U.S. Bureau of the Census "urbanized areas" are used to define overall densities.

Modeling approaches

Key modeling formula

$$\textbf{Correlation Coefficient (r)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

X and Y are the values of the two variables, and $\bar{x}$ and $\bar{y}$ are their respective means.

Inputs (independent variables)

1. Population and job densities
2. Central city strength (e.g. jobs in central cities)
3. Modal split (e.g., transit passenger miles)
4. Automobile provision (e.g., road supplies, parking space)
5. Other

Outputs (dependent variables)

1. gasoline consumption per capita

Q2: What does the data set look like? Particularly what are the rows and columns?

Q3: What are the major modeling tools in this study?

Q4: How many inputs and outputs are there? Why do these variables constitute a relatively valid input-output relationship?

Empirical Findings 1: the United States sample

Land use planning parameters

1. The density of population and jobs in the US. cities is negatively correlated with gasoline use.

2. Significant correlations between gasoline use and both the number and the proportion of jobs in the city center.

3. The metropolitan journey-to-work trip length does not reveal significant relationship to gasoline use.

Conclusion: Urban planning can limit energy consumption in the US context.

Q: What are the coding skills you could use to draw Table 1?

Table 1. Gasoline use and land use variables in U.S. cities (1980)

		Density (persons or jobs/acre)						Central city strength		Other	
		Total		Inner		Outer					
City	Gasoline gallons per capita	Population density	Job density	Population density	Job density	Population density	Job density	Jobs in city center	Proportion of jobs in city center (%)	Proportion of population in inner city (%)	Averag to-worl trip leng (miles)
Houston	567	3.6	2.0	8.5	10.5	3.2	1.2	173,540	12	17	9.3
Phoenix	532	3.2	1.6	7.7	9.7	3.2	1.6	25,920	4	4	8.1
Detroit	503	5.7	2.4	19.4	8.1	4.5	2.0	110,700	7	30	8.7
Denver	483	4.9	3.2	7.7	7.1	4.1	2.0	100,330	12	31	6.8
Los Angeles	445	8.1	4.5	12.1	5.8	6.9	3.6	152,920	5	31	9.3
San Francisco	422	6.1	3.2	23.9	19.4	5.3	2.0	273,160	17	21	7.5
Boston	413	4.9	2.4	18.2	13.3	4.1	1.6	218,210	16	24	6.2
Washington	390	5.3	3.2	17.8	15.4	4.5	2.4	268,700	16	21	8.7
Chicago	367	7.3	3.2	21.9	10.5	4.5	2.0	388,280	12	42	8.1
New York	335	8.1	3.6	43.3	21.5	5.3	2.4	1,930,000	23	40	10.6
Average	446	5.7	2.8	18.2	12.1	4.5	2.0	364,180	12	27	8.1
Correlation with gasoline		−.7390	−.6734	−.7803	−.5923	−.5006	−.4399	−.6367	−.6908	−.6451	−.073
Significance		0.007	0.016	0.004	0.035	0.069	0.101	0.023	0.013	0.021	0.419

Empirical Findings 1: the United States sample

Transportation planning factors

1. The ratio of non-automobile modes is negatively correlated with gasoline use, particularly in the journey-to-work modal.

2. The availability of roads and central city parking are highly and positively correlated with gasoline consumption.

Conclusion: It limits the energy consumption by developing the non-automobile (e.g. transit) systems.

Q: What are the coding skills you could use to draw Table 3?

Table 3. Modal split and automobile provision factors in U.S. cities (1980)

	Modal split						Automobile provision					
	Transit passenger miles		Total passenger miles on transit[b] (%)	Journey to work[c] (%)						Average speed of transit (mph)[g]		
City	Total[a]	Rail		Car	Transit	Walk-bike	Road supply[d] (feet per capita)	Parking spaces[e] (per 1,000 CBD workers)	Average speed of traffic[f] (mph)	Bus	Train	Street car/light rail
Houston	80	—	0.8	94	3	3	34.8	370	32	14	—	—
Phoenix	41	—	0.5	95	2	3	34.1	1033	26	14	—	—
Detroit	70	2	0.8	93	4	3	19.0	473	27	13	26	—
Denver	135	—	1.8	88	7	5	30.8	498	28	13	—	—
Los Angeles	239	—	2.7	88	8	4	14.8	524	28	13	—	—
San Francisco	575	232	6.6	78	17	6	16.1	145	29	14	27	9
Boston	322	217	4.0	74	16	10	17.1	322	24	11	24	12
Washington	383	142	5.0	81	14	5	16.7	264	24	11	25	—
Chicago	603	402	8.0	76	18	6	16.4	91	26	11	29	—
New York	798	623	14.1	64	28	8	15.4	75	22	9	22	—
Average	324	162	4.4	83	12	5	21.6	380	27	12	26	11
Correlation with gasoline	−.9062	−.8455	−.8761	−.9234	−.9281	−.7529	+.8217	+.7140	+.7781	+.8632	+.2607	—
Significance	0.000	0.001	0.000	0.000	0.000	0.006	0.002	0.010	0.004	0.001	0.309	—

a. All transit rides including private buses.
b. Private car passenger miles from VMT and average auto occupancies.
c. U.S. census data.
d. All road types including local roads.
e. On-street and off-street parking.
f. Derived from each city's traffic model.
g. Bus includes trolley buses. Train includes all separated heavy rail systems.

Empirical Findings 2: the global sample

Land use planning factors

1. The density of population and jobs is strongly and negatively correlated with gasoline use.

2. There is a negative correlation between gasoline use and the proportion of jobs in the city center, but not for the absolute number of jobs.

3. The proportion of population living in the inner city is negatively correlated with gasoline use in the global sample.

4. The global sample shows a significant positive correlation of journey-to-work trip lengths with gasoline use.

Conclusion: Urban planning can limit energy consumption in the global context.

Table 5. Gasoline use and land use variables in global cities (1980)

		Density (persons or jobs/acre)						Central city strength		Other	
		Total		Inner		Outer					
City	Gasoline gallons per capita	Population density	Job density	Population density	Job density	Population density	Job density	Jobs in city center	Proportion of jobs in city center (%)	Proportion of population in inner city (%)	Average to-work trip length (miles)
U.S. cities	446	5.7	2.8	18.2	12.1	4.5	2.0	364,180	12	27	8.1
Australian cities	227	5.7	2.4	9.7	10.9	5.3	1.6	107,736	16	17	7.5
Toronto	265	16.2	8.1	23.1	15.4	13.8	5.8	142,645	13	36	8.1
European cities	101	21.9	12.6	36.8	32.0	17.4	8.1	265,505	19	41	5.0
Asian cities	42	64.8	28.8	18.8	119.8	46.6	17.8	785,225	19	32	2.5
Moscow	3	56.2	—	—	—	—	—	—	—	—	—
Correlation with gasoline[a]		−.5778	−.6571	−.3917	−.4846	−.5751	−.5912	−.1026	−.5027	−.4577	+.7620
Significance		0.000	0.000	0.0150	0.003	0.000	0.000	0.291	0.002	0.005	0.000

a. Correlations are on all 32 separate cities in the sample.

Empirical Findings 2: The global sample

Transportation planning factors

1. Gasoline use is related to the use of transit (especially rail) and the amount of provision for the automobile.

2. The average traffic speed is strongly correlated with gasoline use per capita and is positive.

Conclusion: It limits the energy consumption by developing the non-automobile (e.g. transit) systems.

Table 6. A comparison of modal split and automobile provision factors in global cities (1980)[a]

	Modal split						Automobile provision					
	Transit passenger miles			Journey to work (%)						Average speed of transit (mph)		
City	Total	Rail	Total passenger miles on transit (%)	Car	Public transportation	Walk-bike	Road supply (ft/capita)	Parking spaces (per 1,000 CBD workers)	Average speed of traffic (mph)	Bus	Train	Street car/light rail
U.S. cities	324	162	4.4	83	12	5	21.6	380	27	12	26	11
Australian cities	532	306	7.5	76	19	5	28.5	327	27	13	24	14
Toronto	1,227	673	16.7	63	31	6	8.9	198	—	12	21	10
European cities	1,112	801	24.8	44	35	21	6.9	211	19	12	27	11
Asian cities	1,900	1,112	64.1	15	60	25	3.3	67	15	9	22	8
Moscow	>2,647	>1,886	>95	2	74	24	1.3	—	28	13	>26	11
Correlation with gasoline[b]	−.7191	−.5484	−.7530	+.8733	−.8201	−.7301	+.6918	+.5754	+.7034	+.2734	−.0481	+.2602
Significance	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.072	0.410	0.184

a. All variables defined in Table 3 (rail means train and streetcar/light rail).
b. Correlations with gasoline use per capita are for all 32 separate cities in the sample.

A famous plot: visualization of energy consumption vs. population density

Q: What are the coding skills you could use to draw this graph?

Q: What is the model here?

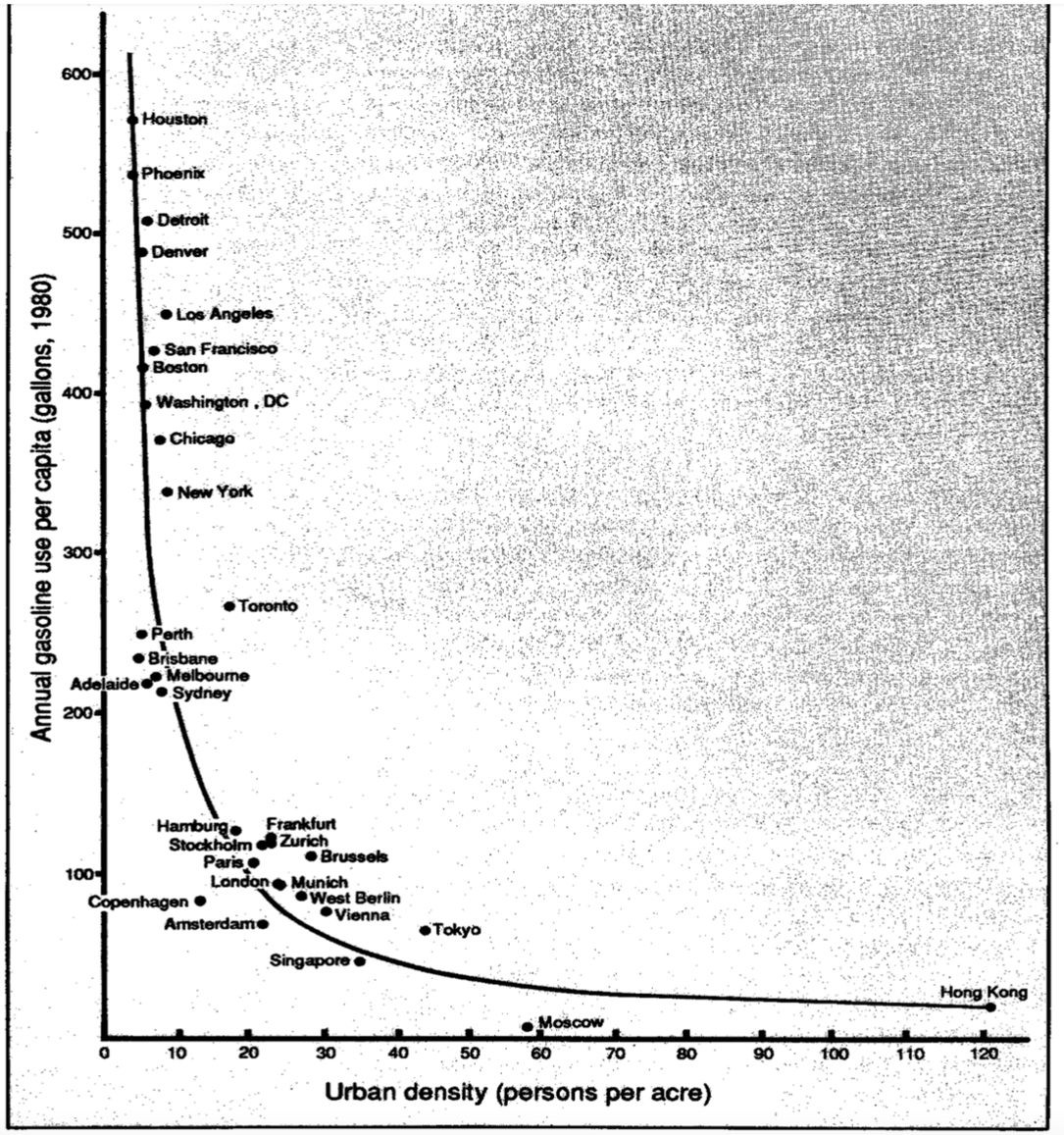


Figure 1. Gasoline use per capita versus population density (1980)

Address the research gaps

1. Use physical planning factors to explain average gasoline consumption

Table 1. Gasoline use and land use variables in U.S. cities (1980)

City	Gasoline gallons per capita	Density (persons or jobs/acre)						Central city strength		Other	
		Total		Inner		Outer					
		Population density	Job density	Population density	Job density	Population density	Job density	Jobs in city center	Proportion of jobs in city center (%)	Proportion of population in inner city (%)	Averag to-worl trip leng (miles)
Houston	567	3.6	2.0	8.5	10.5	3.2	1.2	173,540	12	17	9.3
Phoenix	532	3.2	1.6	7.7	9.7	3.2	1.6	25,920	4	4	8.1
Detroit	503	5.7	2.4	19.4	8.1	4.5	2.0	110,700	7	30	8.7
Denver	483	4.9	3.2	7.7	7.1	4.1	2.0	100,330	12	31	6.8
Los Angeles	445	8.1	4.5	12.1	5.8	6.9	3.6	152,920	5	31	9.3
San Francisco	422	6.1	3.2	23.9	19.4	5.3	2.0	273,160	17	21	7.5
Boston	413	4.9	2.4	18.2	13.3	4.1	1.6	218,210	16	24	6.2
Washington	390	5.3	3.2	17.8	15.4	4.5	2.4	268,700	16	21	8.7
Chicago	367	7.3	3.2	21.9	10.5	4.5	2.0	388,280	12	42	8.1
New York	335	8.1	3.6	43.3	21.5	5.3	2.4	1,930,000	23	40	10.6
Average	446	5.7	2.8	18.2	12.1	4.5	2.0	364,180	12	27	8.1
Correlation with gasoline		−.7390	−.6734	−.7803	−.5923	−.5006	−.4399	−.6367	−.6908	−.6451	−.073
Significance		0.007	0.016	0.004	0.035	0.069	0.101	0.023	0.013	0.021	0.419

Address the research gaps

2. Collected over a five-year period primarily by visiting each city with considerable follow-up correspondence.

Comparative studies of cities around the world are rare, mainly because of the difficulty of collecting data. Most transportation, energy, and planning data are collected on a state or national basis, although each city generally has that data in disparate physical planning and transportation agencies. The data in this study were collected over a five-year period primarily by visiting each city, sometimes twice, with considerable follow-up correspondence. All data were verified by other sources, e.g., gasoline consumption was verified by using vehicle miles of travel (VMTs) for each city and national vehicle fuel efficiency data adjusted for average speed. To achieve comparable land use data, we defined urban areas to exclude all rural land, such as farms, forests, undeveloped land, and large bodies of water. Except in the case of New York, the U.S. city referred to is the standard metropolitan statistical area (SMSA) with the previously mentioned rural land uses excluded, i.e., the U.S. Bureau of the Census "urbanized areas" are used to define overall densities. "Central area" refers to the old, highly built-up central business district defined by each city on the basis of census tracts or traffic planning zones. "Inner area" refers to the pre-World War II urban area; in practical terms, for six of the cities, it is the original city lying at the heart of the urban region, e.g., the city of Detroit. The exceptions are Boston, where inner area is defined as Suffolk County; Washington, D.C., where the inner area is the District of Columbia; Phoenix, where the city authority defined

Major conclusion 1

1. **There are a variety of potential policies to save fuel:**

(1) Increasing urban density;

(2) Strengthening the city center;

(3) Extending the proportion of city that has inner-area land use;

(4) Providing a good transit option;

(5) Restraining the provision of automobile infrastructure.

Major conclusion 2

2. The potential of Re-urbanization is quite considerable.

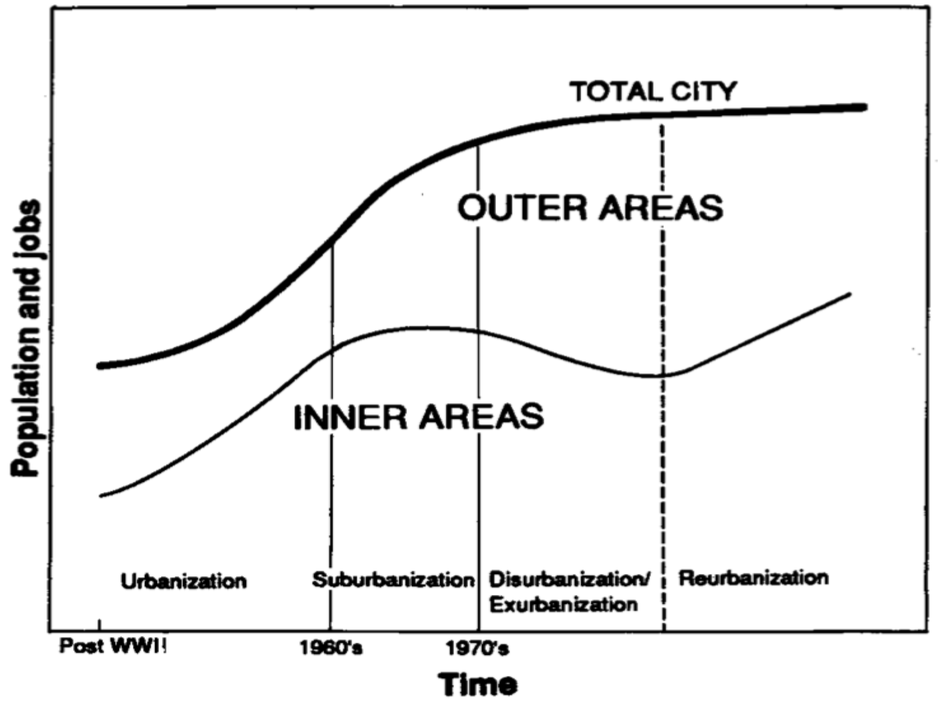


Figure 2. Reurbanization as a fourth phase in urban development

Major conclusion 3

3. A reorientation of transportation priorities in U.S. cities include:

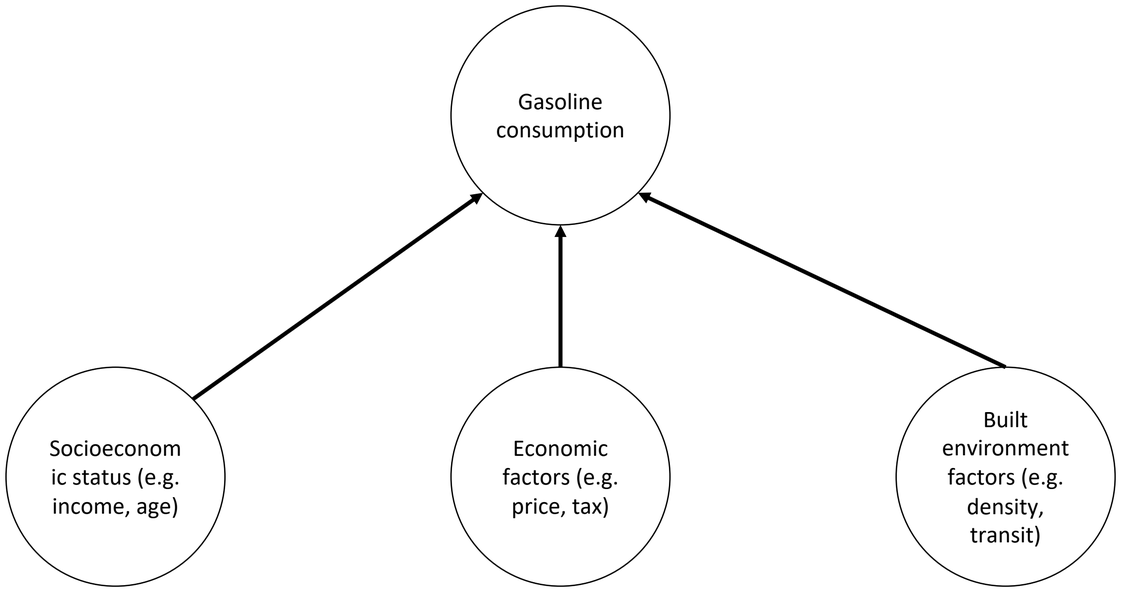
(1) Upgraded and extended transit.

(2) Increased pedestrianization and bicyclization.

(3) Planned congestion: placing a limit on private vehicle movement and adjusting priorities to give advantage to other transportation modes

Back to the Reading Questions: How are the lectures related to the article?

1. What are the major debates the authors seek to resolve in this study?
2. What are the major modeling tools in this study?
3. What does the data set look like? Particularly what are the rows and columns?
4. What are the inputs and outputs? How many inputs and outputs are there? Why are these variables constitute a relatively valid input-output relationship?
5. What are the coding skills you could use from this to conduct this study?

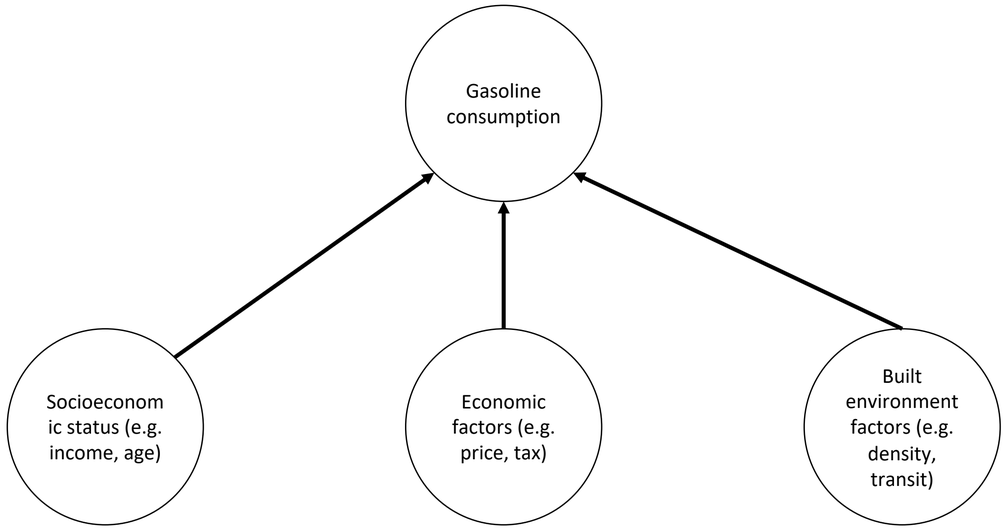


Highlight

You have learnt all the coding and modeling skills within two lectures (lec03-04) to answer the "soul-searching" question in this article.

However, there are limitations to this study.

- **The correlation coefficient** only compares effects of X's one by one. But it cannot capture the relationship of one effect and many causes (one Y and simultaneously many X's).
- **The correlation coefficient** is a weaker version of the $\hat{\beta}_1$ in a univariate linear regression. Therefore, all the critiques to the univariate linear regression are valid to this study.

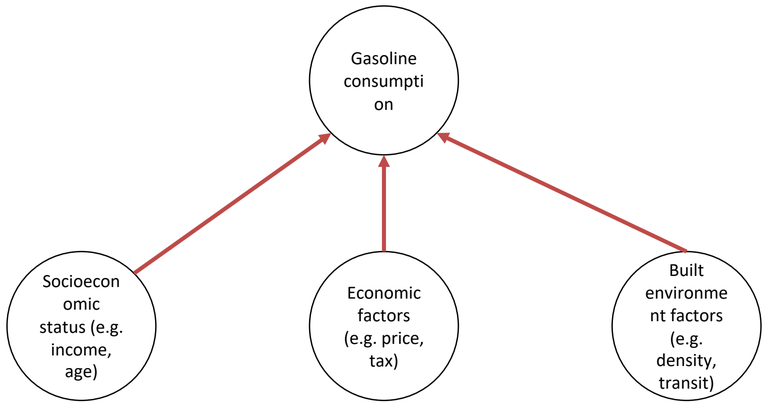


Solution: multivariate linear regression

Paper: Travel demand and the 3Ds: Density, diversity, and design

Part 3. Multivariate Linear Regression

a single output + many inputs



Why do we want more than one input?

1. Summarize more information from regressions
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. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$)

Q: Are the formula in points 4-6 LINEAR?

A: Yes. Linear regressions indicate linear-in-parameter regressions. Therefore, even linear regression can capture complex nonlinear relationship between Y and X.

Example: Property values in Florida

Property value depends on the household income.

However, property value could be caused by other factors: (1) age, (2) household size, (3) education levels, (4) built environment, and many others

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

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 $\widehat{\beta}_1$?

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.)

One caveat

Hard to compare the magnitudes of $\hat{\beta}_1$ and $\hat{\beta}_2$

- $\hat{\beta}_1$: how property values vary with income
- $\hat{\beta}_2$: how property values vary with education levels

However, income ($) and education (years) have different units. It is hard to compare the effects of income and education levels by checking $\hat{\beta}_1$ and $\hat{\beta}_2$

Part 5. Python Lab 05