Lecture-1: Introduction to Econometrics
**Definition**

- Econometrics may be defined as
  - the science in which the tools of economic theory, mathematics and statistical inference is applied to the analysis of behavioral and observed phenomena. (Source: Ragnar Frisch 1933, first issue of Econometrica)*

- The unification of the three elements makes Econometrics powerful
  - Economic theory
  - Mathematics
  - Statistics

* Number of other definitions of econometrics also exist
Why study Econometrics?

- Need to use non-experimental, or observational, data to make inferences

- Important to be able to apply economic theory to real world data to understand causality, interdependency.

- Behavior is hard to model and predict

- Only a model estimate can be obtained with some degree of certainty (confidence)
Why study Econometrics?

- An empirical analysis uses data to test a theory or to estimate a relationship.

- A formal economic model can be tested.

- Theory may be ambiguous as to the effect of some policy change – can use econometrics to evaluate the program.
  - Example-1: Capacity expansion on induced demand
  - Example-2: Congestion pricing on travel behavior
  - Example-3: Household or destination choice because of changes in transportation infrastructure.
Types of data

- Cross-sectional data (a)
- Time series data (b)
- Pooled cross-sectional data (a+b)
- Panel data
Types of Data – Cross Sectional

- Cross-sectional data is a random sample

- Each observation is a new individual, firm, etc. with information at the same point in time (no temporal variation)
  - Example-1: census survey (every 10 years)
  - Example-2: National household travel survey (every five years)
  - Example-3: Commodity flow survey (every five years)

- If the data is not a random sample, we have a sample-selection problem
Example: cross-sectional data

<table>
<thead>
<tr>
<th>Zone</th>
<th>Peak Hr Trips Attracted (Y)</th>
<th>Total (Xₙ)</th>
<th>Manufacturing (X₂)</th>
<th>Retail &amp; Services (X₃)</th>
<th>Other (X₄)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9,500</td>
<td>9,220</td>
<td>6,600</td>
<td>2,500</td>
<td>120</td>
</tr>
<tr>
<td>2</td>
<td>2,200</td>
<td>2,045</td>
<td>125</td>
<td>1,905</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>330</td>
<td>574</td>
<td>228</td>
<td>87</td>
<td>259</td>
</tr>
<tr>
<td>4</td>
<td>153</td>
<td>127</td>
<td>0</td>
<td>127</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3,960</td>
<td>3,850</td>
<td>2,750</td>
<td>800</td>
<td>300</td>
</tr>
<tr>
<td>6</td>
<td>1,200</td>
<td>995</td>
<td>105</td>
<td>805</td>
<td>85</td>
</tr>
<tr>
<td>7</td>
<td>240</td>
<td>223</td>
<td>165</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>36</td>
<td>6</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2,100</td>
<td>2,250</td>
<td>1,560</td>
<td>515</td>
<td>175</td>
</tr>
<tr>
<td>10</td>
<td>230</td>
<td>209</td>
<td>36</td>
<td>173</td>
<td>0</td>
</tr>
</tbody>
</table>
Heterogeneity

*Size* and *scale* effect must be taken into consideration otherwise model building will be erroneous

Not to mix apples with oranges
Cross-sectional data (cautions) - 2
Types of Data – Time Series

- Time series data has a separate observation for each time period – e.g. annual traffic volume on a corridor, census observations over multiple decades

- Time periods to consider
  - Daily, Weekly, Monthly, Quarterly, Annually, Quinquennially (every five years), Decennially (every 10 years)

- Since not a random sample, different problems to consider

- Trends and seasonality will be important

- Stationary issue
  - Loosely speaking a time series is stationary if its mean and standard deviation does not vary systematically over time
Stationary time series-example

- Stationarity
- Seasonality
- Trend effects
- We will explore more in time series lectures
Types of Data-pooled data

- Data elements consist of both cross-sectional and time series features
- Random samples over time are different (the same observations in sample may not remain same)
- Useful for before-after studies
A panel dataset consists of a time series for each cross-sectional member of the dataset.

Can pool random cross sections and treat similar to a normal cross section. Will just need to account for time differences.

Can follow the same random individual observations over time – known as panel data or longitudinal data.

Also known as longitudinal and micropanel data.
## Panel data - example

<table>
<thead>
<tr>
<th>Zone</th>
<th>Peak Hr Trips Attracted(Y)</th>
<th>Total Employment (X₁)</th>
<th>Manufacturing (X₂)</th>
<th>Retail &amp; Services (X₃)</th>
<th>Other (X₄)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9500</td>
<td>9220</td>
<td>6600</td>
<td>2500</td>
<td>120</td>
</tr>
<tr>
<td>1</td>
<td>5649</td>
<td>4996</td>
<td>3679</td>
<td>666</td>
<td>1288</td>
</tr>
<tr>
<td>2</td>
<td>2200</td>
<td>2045</td>
<td>125</td>
<td>1905</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>1839</td>
<td>4239</td>
<td>4822</td>
<td>1595</td>
<td>930</td>
</tr>
<tr>
<td>3</td>
<td>330</td>
<td>574</td>
<td>228</td>
<td>87</td>
<td>959</td>
</tr>
<tr>
<td>3</td>
<td>7057</td>
<td>2326</td>
<td>1238</td>
<td>487</td>
<td>628</td>
</tr>
<tr>
<td>4</td>
<td>153</td>
<td>127</td>
<td>0</td>
<td>127</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8754</td>
<td>2128</td>
<td>778</td>
<td>487</td>
<td>300</td>
</tr>
<tr>
<td>5</td>
<td>3960</td>
<td>3850</td>
<td>2750</td>
<td>800</td>
<td>2361</td>
</tr>
<tr>
<td>5</td>
<td>1408</td>
<td>5656</td>
<td>4126</td>
<td>1929</td>
<td>85</td>
</tr>
<tr>
<td>6</td>
<td>1200</td>
<td>995</td>
<td>105</td>
<td>805</td>
<td>1539</td>
</tr>
<tr>
<td>6</td>
<td>2328</td>
<td>5209</td>
<td>1178</td>
<td>1062</td>
<td>1539</td>
</tr>
<tr>
<td>7</td>
<td>240</td>
<td>223</td>
<td>165</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1095</td>
<td>4646</td>
<td>1858</td>
<td>848</td>
<td>1342</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>36</td>
<td>6</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>9315</td>
<td>2433</td>
<td>1826</td>
<td>1092</td>
<td>2733</td>
</tr>
<tr>
<td>9</td>
<td>2100</td>
<td>2250</td>
<td>1560</td>
<td>515</td>
<td>175</td>
</tr>
<tr>
<td>9</td>
<td>5642</td>
<td>3662</td>
<td>1671</td>
<td>1700</td>
<td>1146</td>
</tr>
<tr>
<td>10</td>
<td>6738</td>
<td>9493</td>
<td>3490</td>
<td>1701</td>
<td>2285</td>
</tr>
<tr>
<td>10</td>
<td>230</td>
<td>209</td>
<td>36</td>
<td>173</td>
<td>0</td>
</tr>
</tbody>
</table>
Measurement scale of variables

- Ratio scale
- Interval scale
- Ordinal scale
- Nominal scale
Ratio scale

- Measurements made with ratio scale can be added, subtracted, multiplied and divided.
- For a variable $X$ taking two values, $X_1$ and $X_2$, the ratio $X_1/X_2$ and $X_1-X_2$ are meaningful quantities.
- Also there is a natural ordering (ascending and descending) of the values along scale ($X_1 \leq X_2$).
- Most of the variables belong to this category.
- Example: Height, weight, speed, distance, etc.
Interval scale

- In interval scale the difference is meaningful but it does not satisfy the ratio property.
- Example: A 50 mph speed limit is higher than 40 mph but lower than 60 mph. The difference is meaningful (equal intervals)
- Interval scales normally have a minimum and maximum point.
A variable is ordinal if it only satisfies the natural ordering.

Ordinal scales classify subjects and rank them in terms of how they possess characteristics of interest.

Example
- Grading system
  - A, B, C grades
- Income class
  - Upper, middle, and lower

Ordering exists but not the differences and ratios.
Nominal scale

- Variables in this category does not have any of the features of ratio scale
  - Ratio
  - Difference
  - Natural ordering

- Variables such as
  - Gender (male, female);
  - Facility type (freeway, arterial)

- Also called as categorical variables
One of the important features of econometric analysis is causality. What is the causal effect of one variable (education) over another (income)? Ceteris paribus means “with all other (relevant) factors being equal” what is the causal effect of education over income.
The Question of Causality

- Simply establishing a relationship between variables is rarely sufficient
- Want to know the effect to be considered causal
- If we’ve truly controlled for enough other variables, then the estimated ceteris paribus effect can often be considered to be causal
- Can be difficult to establish causality
Definition of the simple linear regression model

\[ y = \beta_0 + \beta_1 x + u \]

- **Intercept**: \( \beta_0 \)
- **Slope parameter**: \( \beta_1 \)
- **Dependent variable, explained variable, response variable**: \( y \)
- **Independent variable, explanatory variable, regressor**: \( x \)
- **Error term, disturbance, unobservables**: \( u \)

Explains variable \( y \) in terms of variable \( x \)
Simple Regression Model (2)

- Fit as good as possible a regression line through the data points:

For example, the i-th data point \((x_i, y_i)\)
Simple Regression Model (3)

- What does \textbf{as good as possible} mean?
- Regression residuals

\[ \hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \]

- Minimize sum of squared regression residuals (commonly referred as ordinary least squares)

\[ \min \sum_{i=1}^{n} \hat{u}_i^2 \rightarrow \hat{\beta}_0, \hat{\beta}_1 \]
Simple Regression Model (4)

- Number of trips attracted as a function of total employment

<table>
<thead>
<tr>
<th>Zone</th>
<th>Peak Hr Trips Attracted (Y)</th>
<th>Total (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9,500</td>
<td>9,220</td>
</tr>
<tr>
<td>2</td>
<td>2,200</td>
<td>2,045</td>
</tr>
<tr>
<td>3</td>
<td>330</td>
<td>574</td>
</tr>
<tr>
<td>4</td>
<td>153</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>3,960</td>
<td>3,850</td>
</tr>
<tr>
<td>6</td>
<td>1,200</td>
<td>995</td>
</tr>
<tr>
<td>7</td>
<td>240</td>
<td>223</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>2,100</td>
<td>2,250</td>
</tr>
<tr>
<td>10</td>
<td>230</td>
<td>209</td>
</tr>
</tbody>
</table>

Y = a + bX
Other considerations

- How good is the model we just developed?
- How much confidence we have in this model?
- Is total employment a good predictor of number of trips attracted?
- How significant is employment?
- Do we need an intercept?
- What are the limitations of OLS?
Next class

- Two variable regression model properties
- Efficiency
- Consistency
- Limitations
- Extensions to three and more variables