Trip Generation

John K. Abraham

CE 7630

Trip Generation

- The study of trip generation attempts to identify and quantify the trip ends related to various urban activity
  - No description other trip characteristics such as direction, length or duration
  - Relating trip ends to land use and socioeconomic characteristics through regression analysis.
  - Relating trip ends to land area, floor area or other use measures such as employment through trip rates.
  - Classifying trip ends by characteristics of the analysis unit generally referred to as cross-classification analysis.
Trips

Centroid movement, Tij = 50
Vector

Spatial Interaction

Constructing an O/D Matrix

Spatial Interactions

O/D Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Tj</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>D</td>
<td>20</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Tj</td>
<td>20</td>
<td>0</td>
<td>28</td>
<td>40</td>
<td>50</td>
<td>39</td>
</tr>
</tbody>
</table>
Three trip purposes are used (at a minimum). Many other types are possible. Some large are models use 8 or more.

- Home Based Work (HBW), only trips from home to work or visa versa. Most urban area peak hour trips are this type.
- Home Based Other (HBO), any other kind of trip with one end at the home.
- Non Home Based (NHB), any trip that doesn't either come from or go to a home.

Trip number 1 was produced at home and attracted to the office -- therefore a home-based work trip.

Trip number 2 was produced at the office and attracted to the store -- therefore a non-home-based other trip.

Trip number 3 was produced at home and attracted to the store -- therefore a home-based other trip.
SEMCOG definitions

- Home-based work -- trips between a person's home and place of employment for the purpose of working.
- Home-based other -- trips between a person's home and any other destination for any other purpose.
- Non-home-based -- trips that have neither end at home, regardless of purpose. These may include truck and taxi trips.
- Internal-external trips -- trips with one end inside the study area and one outside the study area.
- Through trips -- trips that have neither end in the study area but pass through it.
- Truck and taxi trips -- if required, but often included as part of non-home-based trips.

Trip Generation

What trips will they make?

Forecast Data
- Employment (6)
- Population (3)
- Households (4)
- Single/Multi Family by Transportation Analysis Zone

Daily Person Trips
- Home-base Work
- Home-based Other
- Non-home-based
- College
- School
- Commercial Vehicles

Cross-classified Households
- persons x workers x income
- persons x college-age
- persons x student-age

Wayne State University
Residential Trip Generation

- Depends on socioeconomic and neighborhood characteristics rather than transportation system
  - Car ownership
  - Family size
  - # of people more than 5 yrs old
  - Length of residence
  - Family income
  - # of people above 16 yrs old
  - # of people who drive
  - Age of the head of household
  - Distance from CBD
  - Occupation of head of household
  - Type of house structure

Trip Productions

- Rate Method
- Cross Classification Method
- Regression Method
Rates Based on Activity

- Used for Traffic Impact Analysis or very detailed regional models
- Does not consider other characteristics such as household size, income and auto ownership
- This method used for non residential trip generation
  - By Land-use types: office, industry, commerce, shopping
    - Trips per 1,000 sq. ft.
    - Trips per employee
Cross Classification Method

- Used mainly for trip productions. Used for Regional Studies not Traffic Impact Analysis
- Advantages:
  - allows the transportation planner to get a good understanding of the importance of the variables of trip generation
  - There is no need for an assumption of linearity between independent and dependent variables.
- Disadvantages:
  - Requires very detailed data to construct and predict trip generation.
  - The independent variables selected for the study may not be independent.
Cross Classification

- Dependent variable cross-tabulated against 2 or more variables
  - such as trips per household Vs. income / auto ownership / income per household
- The variables are divided into distinct classes
Regression Method

- Regression is mostly used for attractions
- Graphical Solution to Trip Generation given values for variables.
- A linear or nonlinear relationship can be assumed.
- The dependent variable is the number of trips. (y-axis)
- The independent variables are assumed to be independent of each other although that is not always the case. If they are not independent of each other they are "colinear".
- The independent variables can be population, autos, number of dwelling units, etc.
Regression

- Regression Analysis provides the basis for predicting the values of a variable from values of one or more other variables
  - Linear – Assumes the relationship to be a straight line
- Correlation Analysis: Assesses the strength of the relationship between the variables

Scatter Diagrams

- Useful in examining relation between two variables
- Could be direct or inverse relations

<table>
<thead>
<tr>
<th>Zone #</th>
<th>Pop.</th>
<th>Households</th>
<th>Trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7212</td>
<td>2488</td>
<td>5126</td>
</tr>
<tr>
<td>2</td>
<td>4818</td>
<td>2188</td>
<td>3764</td>
</tr>
<tr>
<td>3</td>
<td>8789</td>
<td>2423</td>
<td>4030</td>
</tr>
<tr>
<td>4</td>
<td>5805</td>
<td>2141</td>
<td>3921</td>
</tr>
<tr>
<td>5</td>
<td>3054</td>
<td>1241</td>
<td>1644</td>
</tr>
<tr>
<td>6</td>
<td>1040</td>
<td>3451</td>
<td>4467</td>
</tr>
<tr>
<td>7</td>
<td>2735</td>
<td>1857</td>
<td>3407</td>
</tr>
<tr>
<td>8</td>
<td>1784</td>
<td>905</td>
<td>1143</td>
</tr>
<tr>
<td>9</td>
<td>4418</td>
<td>1695</td>
<td>3202</td>
</tr>
<tr>
<td>10</td>
<td>6557</td>
<td>1960</td>
<td>3900</td>
</tr>
</tbody>
</table>
Purposes of Regression / Correlation Analyses

- Provides estimates of values of dependent variable from values of independent variable
  - Provides estimates of mean value of Y for each value of X
- Obtain measures of error involved in using the regression line as a basis of estimation
- Measure degree of association between two variables / strength of relationship
Outline:

- Simple Linear Regression Analysis
- Multiple Linear Regression
- Understanding the Regression Outputs
- Validating a Regression Model

Scatter Diagram

- To answer the questions, we need to relate Y to the Xs
- Plot the observed Y vs X₁ (scatter chart) and then Add a Trend Line

Problem:
- There is no unique relationship between Y and X₁! - why?
City X has recently conducted a travel survey of its 10 TAZs. The collected data are summarized in the Table. Based on the given information,

- predict the total number of trips that will be produced by each zone in 10 years (assume zonal population in 10 years is known)?
- relate \( Y \) (total trips produced by a zone) to \( X_1 \) (zonal population)?
- expected total number of trips for a zone with a population of 5,000?
- How confident are you about your estimate?

### Linear Regression Analysis Example

![Observed Trips and Zonal Activity](Image)

<table>
<thead>
<tr>
<th>Zone #</th>
<th>Pop.</th>
<th>Households</th>
<th>Trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7212</td>
<td>2488</td>
<td>5126</td>
</tr>
<tr>
<td>2</td>
<td>4618</td>
<td>2188</td>
<td>3764</td>
</tr>
<tr>
<td>3</td>
<td>6789</td>
<td>2423</td>
<td>4030</td>
</tr>
<tr>
<td>4</td>
<td>5805</td>
<td>2141</td>
<td>3921</td>
</tr>
<tr>
<td>5</td>
<td>3054</td>
<td>1241</td>
<td>1644</td>
</tr>
<tr>
<td>6</td>
<td>10463</td>
<td>3451</td>
<td>4467</td>
</tr>
<tr>
<td>7</td>
<td>2735</td>
<td>1857</td>
<td>3407</td>
</tr>
<tr>
<td>8</td>
<td>1784</td>
<td>905</td>
<td>1143</td>
</tr>
<tr>
<td>9</td>
<td>4418</td>
<td>1695</td>
<td>3202</td>
</tr>
<tr>
<td>10</td>
<td>6657</td>
<td>1960</td>
<td>3900</td>
</tr>
</tbody>
</table>

### Simple Linear Regression

- **Problem:** consider two variables, \( X \) and \( Y \), and we have \( n \) paired observations on them (data): \((x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\), what is the relationship between \( Y \) and \( X \)?
- **Method:** Assumed \( Y \) and \( x \) are linearly related, that is
  \[
  Y = b_0 + b_1 X
  \]
  where: \( X \) - independent variable (a factor that is considered to have impact on \( Y \))
  \( Y \) - dependent variable
  \( b_0, b_1 \) - regression coefficients
- **Our Goal:**
  - Find the best value for \( b_0 \) and \( b_1 \)
  - Find out whether or not the identified relationship is good, i.e., validate the model
Regression Analysis

- Plot $Y$ vs. $X_1$ (scatter graph) and the straight line $Y = b_0 + b_1 X_1$
  with different intercept ($b_0$) and slope ($b_1$)

We shall find $b_0$ and $b_1$ so as to minimize the total estimation error ($E$):

$$E = \sum e_i^2 = \sum (y_i - (b_0 + b_1 x_i))^2$$

This is called the method of least squares.

Method of Least Squares

- What constitutes a goodness of fit
- Method of least squares is a curve fitting technique
  - Sum of deviations of $x$ from the regression line should be a minimum
  - Best fit line will have minimum value when compared to other possible lines
  - Passes through the point $(\bar{Y}, \bar{X})$ therefore total of positive and negative deviations is zero
Manual Calculation of Coefficients

- $b_0$ and $b_1$ can be determined mathematically as follows:

\[ Y = b_0 + b_1 X \]

\[ e_i = (y - b_0 + b_1 x_i) \]

\[ \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \]

\[ \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0 \]

\[ \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 \]

\[ \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i y_i \]

\[ b_0 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \]

\[ b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \]
Estimating the Regression Coefficients

- Data: \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)

- From data we can calculate the following statistics

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

\[
S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad S_{YX} = \frac{S_{XY}^2}{n - 2}
\]

- \(b_0\) and \(b_1\) can then be determined by

\[
b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{S_{XY}}{S_{XX}}
\]
Regression Coefficients, \( b_0, b_1, \ldots \)

Example 1: \( Y = 1546.2 + 0.343 \ X_1 \)

- \( b_0 = 1546.2 \) interpreted as: amount of trips would be produced if there is no population in a zone (\( X_1 = 0 \))?!)
- \( b_1 = 0.343 \) interpreted as: one additional person in a zone would produce additional 0.343 trips!
Multiple Regression

- The total number of trips produced by a zone should also be related to the households of that zone (in addition to population), can we relate $Y$ to both $X_1$ (zonal population) and $X_2$ (zonal households)?

In general:

- Data: $(y_1, x_{11}, x_{12}, ...), (y_2, x_{21}, x_{22}, ...), ..., (y_n, x_{n1}, x_{n2}, ...)$
- Assumed Relationship

\[ Y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_r x_r \]

- Problem: how to find coefficients $b_0, b_1, b_2, ...$

The concept to solve the problem is the same as that for the simple linear regression, - method of least squares.

---

Multiple Regression Using Excel

**SUMMARY OUTPUT**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.891</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.794</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.735</td>
</tr>
<tr>
<td>Standard Error</td>
<td>627.121</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>491.361</td>
<td>637.458</td>
<td>0.771</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.804</td>
<td>0.575</td>
<td>1.398</td>
</tr>
<tr>
<td>X Variable 2</td>
<td>0.552</td>
<td>0.398</td>
<td>1.385</td>
</tr>
</tbody>
</table>

Resulting Equation: $Y = 491.361 + 0.804X_1 + 0.552X_2$
Standard Error of Y Given X, $S_{Y|X}$

- Just like standard deviation for mean, for regression it is standard error or standard deviation of the residuals.
- Simply measures the scatter of the observed values of Y around corresponding calculated values.

$$S_{Y|X} = \sqrt{\frac{S_{YY} - S_{XY}^2 / S_{XY}}{n-2}} = \sqrt{\frac{\sum (\hat{y} - y)^2}{n-2}}$$

- $n-2 = \text{degrees of freedom} = \text{normally } n-k \text{ where } k \text{ is the number of constants in the regression equation. In case of a straight line, } k=2$
- If standard error is 0 it means the regression line is a perfect fit.

Standard error (deviation) of regression coefficient of $b_1$

$$\text{Standard error of coefficient estimate } b_1 = \frac{S_{Y|X}}{\sqrt{S_{XX}}}$$
## Regression Output from Excel

### SUMMARY OUTPUT

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.7801</td>
</tr>
<tr>
<td>R Square</td>
<td>0.6085</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.5595</td>
</tr>
<tr>
<td>Standard Error</td>
<td>808.12</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

### ANOVA

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8119632.892</td>
<td>8119633</td>
<td>12.43318</td>
<td>0.007777</td>
</tr>
<tr>
<td>8</td>
<td>5224491.608</td>
<td>653061.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>13344124.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Coefficients

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1546.2</td>
<td>600.034</td>
<td>2.576876</td>
<td>0.032775</td>
<td>162.5315</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.3435</td>
<td>0.097406453</td>
<td>3.526072</td>
<td>0.007777</td>
<td>0.118842</td>
</tr>
</tbody>
</table>

*This is the $S_{1|X}$ value!*

*This is the “Standard error (deviation) of regression coefficient of $b_1$” value!*

### Confidence Interval and Hypothesis Testing for $b_1$

**Test the hypothesis $H_0: \beta_1 = 0$.** If this hypothesis is *not* rejected, then $X_1$ is *not* a significant factor influencing $Y$ and should not be included in the function.

\[
|t| < t_{\alpha/2, n-k-1} \text{, reject } H_0 \text{ at } (1-\alpha)100\% \text{ confidence level}
\]

where $t_{\alpha/2, n-k-1}$ is a value of the (student) $t$-distribution with $n-k-1$ degrees of freedom ($k =$ number of independent variables)
**Regression Output from Excel**

**SUMMARY OUTPUT**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.7801</td>
</tr>
<tr>
<td>R Square</td>
<td>0.6085</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.5595</td>
</tr>
<tr>
<td>Standard Error</td>
<td>808.12</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>SS</td>
<td>MS</td>
</tr>
<tr>
<td>Regression</td>
<td>1</td>
<td>8119632.892</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>5224491.608</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>13344124.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1546.2</td>
<td>600.0341482</td>
<td>2.576876</td>
<td>0.032775</td>
<td>162.5315</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.3435</td>
<td>0.097406453</td>
<td>3.526072</td>
<td>0.007777</td>
<td>0.118842</td>
</tr>
</tbody>
</table>

This is the 95% confidence interval of b1.

---

**Coefficient of Determination, R²**

- R² is a measure of overall quality of the regression, specifically, it is the percentage of total variation in Y that is accounted by the sample regression line.

\[
R^2 = \frac{S_{XY}^2}{S_{XX} \cdot S_{YY}} = \frac{\text{regressionSS}}{\text{totalSS}} = 1 - \frac{\text{errorSS}}{\text{totalSS}}
\]

- A high value of R² means that most of the variation we observed in Y (y_i) can be attributed to their corresponding X values.
- There is however NO standard on how high a R² value is “good” enough - it depends on the application! In transportation planning, a R² score over 0.3 is often acceptable!
Regression Output from Excel

This is the $R^2$ value!
Example (continued):

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>8122236.20</td>
<td>8122236.20</td>
<td>12.43</td>
<td>0.05</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>5226522.20</td>
<td>653315.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>13348758.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Validate Regression Model

(pitfalls in regression analysis!)

- Important Assumptions Used in Linear Regression Analysis
  - Y is *linearly* related to x
  - For any given x, Y is *Normally* distributed
  - Residual Error has *constant* standard deviation (for all x)
  - Independent variables are NOT *linearly* related (multiple regression only!)

- How to find out whether or not these conditions are met?
Check for Linearity

- **Objective:** is $Y$ really linearly related to $x$
- **Method:** check visually from scatter plot

Check for Normality

- **Objective:** is the residual normally distributed ($e_i$)?
- **Method:** plot a histogram of the residuals
  
  \[ e_i = y_i - (b_0 + b_1 x_i) \]
Check for Multicolinearity

Objective: is independent variable $X_i$ linearly correlated to independent variable $X_j$?

Method: check scatter plot of $X_i$ vs. $X_j$

No evidence of multicolinearity

Evidence of multicolinearity

Check for Multicolinearity - cont.

- **Sample Correlation Coefficient ($r$):** is a measure of the degree of linear-relationship between two variables. For variable, $X_1$ and $X_2$, $r$ is defined as

  - $-1 \leq r \leq 1$:
    - if $r = +1$, $X_1$ and $X_2$ have a perfect positive correlation
    - if $r = -1$, $X_1$ and $X_2$ have a perfect negative correlation
    - if $r = 0$, $X_1$ and $X_2$ have no linear relationship
    - if $|r| > 0.4$ for two independent variables, it may run into the multicolinearity problem if both variables are included in a regression equation.

- Correlation Coefficient can also be directly obtained using Excel function
Objective: is the standard deviation of the residual (e) constant across all X values?

Method: check scatter plot of the residuals VS. each X variable

---

**Regression Output from Excel**

**SUMMARY OUTPUT**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>8119632.892</td>
<td>8119633</td>
<td>12.43318</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>5224491.608</td>
<td>653061.5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>13344124.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1546.2</td>
<td>600.0341482</td>
<td>2.576876</td>
<td>0.032775</td>
<td>162.5315</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.3435</td>
<td>0.097406453</td>
<td>3.526072</td>
<td>0.007777</td>
<td>0.11842</td>
</tr>
</tbody>
</table>