Extrapolation: Concepts and Techniques

Pam Perlich
Urban Planning 5/6020
University of Utah
Learning Objectives

1. Extrapolation
   - Concepts, assumptions, limitations
   - Alternative functional forms – linear and nonlinear

2. Simple linear regression
   - Computational basis
   - Alternative techniques in Excel

3. Calculating “forecasted” values from fitted functions in Excel
Part 1: Extrapolation
Uncertainty (forecasting error) increases with
- Longer forecast horizon
- Smaller areas

Extrapolation techniques have a higher probability of success in
- Short time horizons
- Large areas
Extrapolation Technique

- Fit function to a set of observations and extend this pattern into the future
- Use the function that
  - Is the “function of best fit”
    - (Least squares or regression)
  - Approximates our best understanding of future conditions
    - Incorporate growth constraints or known conditions
Assumptions

- Use of aggregate data, generally across time (population, employment, etc.)
- Future movement of the data series is determined by past patterns embedded in the series
- The essential information about the future of the data series is contained in the history of the series
- Past trends will continue into the future
Advantages / Benefits

- Computational simplicity
- Transparent methodology
- Ease of application
- May work for
  - Large areas
  - Short time horizons
  - Slow grow areas
Disadvantages / Risks

- Does not account for underlying causes / structural conditions
  - Example: Cohorts are invisible
- Ignores structural / systemic context
- Current trend often do not continue
- Excludes any external considerations
Alternative Functional Forms (Klosterman)

- Linear – constant increments of change
- Geometric – constant rate of change
- Parabolic – Accelerating growth rate
- Modified Exponential – growth limit
- Gompertz – growth limit
- Logistic – growth limit

Explore these in spreadsheets:
http://home.business.utah.edu/bebrpssp/URPL5020/Trend/
Klosterman’s Technique

- Klosterman
  - Transforms curves into lines
  - Performs linear regression
- Some functions are not available in Excel (e.g., Gompertz, modified exponential, logistic)
- His approach can be applied in Excel so that these alternative functions are available.
  - Transform data according to his technique
  - Fit a trend line using the Excel function
  - Reverse the transformation to compute forecasted values.
Linear Function

$Y_C = a + bX$

where

$a$ is the intercept and $b$ is the slope

Constant increments of growth
Geometric Function

Geometric Function: \( Y_c = 10(1.1^x) \)

\[ Y_c = ab^x \]

where
- \( a \) is the intercept
- \( b \) is the growth rate plus one

Constant rate of growth
Parabolic Function

\[ Y_C = a + bX + cX^2 \]

where

- \( a \) is the Y intercept and \( b \) is the slope
- **Constantly changing slope**
- If \( b > 0 \) \( \Rightarrow \) Growth is accelerating
Modified Exponential Function:  \( Y_C = 200 - 100 \times (0.8)^X \)

where

- \( a \) is \( c \) minus the \( Y \) intercept
- \( b \) is the ratio of successive growth increments (constant)
- \( c \) is the **asymptotic** value

\[ Y_C = c + ab^X \]
Gompertz Function

\[ Y_C = 100 \times (0.9)^{0.8^X} \]

If \( \ln(a) < 0 \) with \( 0 < b < 1 \) \( \Rightarrow \) \( C \) is the upper limit

Ratio of the logarithms of successive observations is constant

\[ Y_C = ca^{b^X} \]
Logistic Function

\[ Y_c = \frac{1}{10 + (0.5)^x} \]

- If \( b \) is between 0 and 1 and \( a < 0 \)
- Then
  1. Curve takes the “S” shape and
  2. \( 1/c \) is the asymptotic value (upper limit) and
  3. 0 is the lower limit
Excel Tool to Fit S-Curve to Data

- Developed by Stephen R. Lawrence of University of Colorado
  - [http://leeds-faculty.colorado.edu/Lawrence/Tools/SCurve/scurve.xls](http://leeds-faculty.colorado.edu/Lawrence/Tools/SCurve/scurve.xls)
- Algorithm for fitting a logistic function to a set of data.
Extrapolation Method - Summary

- Simple technique that may be the most appropriate for
  - Tight time constraints
  - Slow changing conditions
  - Short time frames
  - General trend identification

- Judgment must be exercised or results may potentially be absurd
Part 2: Simple Linear Regression
The regression (least squares) technique is used to:

- Establish a trend in time series data
- Extend this pattern into the future

Given the existence of a time trend, fitting equations enables us to identify the mathematical function that best captures the relationship.
Meanings of Coefficients

- $R^2$ is the regression coefficient
- $0 \leq R^2 \leq 1$
  - 0 $\rightarrow$ No relationship
  - 1 $\rightarrow$ Perfect fit
- $R$ : correlation coefficient
  - Square root of $R^2$ and signed according to the direction of the relationship
    - $-1 \leq R \leq 1$
      - 1 $\rightarrow$ Perfect fit, positive relationship
      - -1 $\rightarrow$ Perfect fit, inverse relationship
      - 0 $\rightarrow$ No relationship
Fit a Function to the Data

Relationship between Educational Attainment and Median 1999 Household Income: Utah

\[ y = 594.16x + 27179 \]
\[ R^2 = 0.6031 \]

Fit a Function to the Data

Deviation = Observed Minus Fitted

2+ Races
Least Squares Method

- Find the line that minimizes the squared differences between the observed dependent variable and the calculated dependent variable.
- \( y = ax + b \) is the linear function
- \((x', y') \) is the observed pair.
- Minimize the sum of all \((y' - y)^2\)
Solve these Simultaneous Equations

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

\[ a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \]
Example: Scatter Plot

\[ y = 0.5147x + 1.2794 \]

\[ R^2 = 0.9577 \]

From: Gottfried, page 106
Matrix Algebra Solution: Covered in the Next Section of Course

In Matrix Form

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>a</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>5</td>
<td></td>
<td>27.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>479</td>
<td>41</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>=</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>299</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Inverse

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>-0.0574</th>
<th>0.0070</th>
<th></th>
<th>27.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>0.6709</td>
<td>-0.0574</td>
<td>*</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.514706</td>
<td>1.279412</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From: Gottfried, page 106
Goodness of Fit Measure

Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} [y_i - f(x_i)]^2$$
r-squared

\[ r^2 = 1 - \frac{SSE}{SST} \]

Where:

\[ SST = \sum_{i=1}^{n} [y_i - \bar{y}]^2 \]
Values of $r^2$

- $0 \leq r^2 \leq 1$
- As $r^2$ approaches 1, the fit is better
- As $r^2$ approaches 0, the fit is worse
Calculating $r^2$ in Excel

<table>
<thead>
<tr>
<th>I</th>
<th>x</th>
<th>y</th>
<th>f(x)</th>
<th>y-f(x)</th>
<th>Error</th>
<th>Error^2</th>
<th>y-(AveY)</th>
<th>(y-(AveY))^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2.00</td>
<td>2.308824</td>
<td>-0.308824</td>
<td>0.095372</td>
<td>-3.50</td>
<td>12.25</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3.50</td>
<td>3.338235</td>
<td>0.161765</td>
<td>0.026168</td>
<td>-2.00</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>4.50</td>
<td>4.882353</td>
<td>-0.382353</td>
<td>0.146194</td>
<td>-1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>8.00</td>
<td>6.941176</td>
<td>1.058824</td>
<td>1.121107</td>
<td>2.50</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>9.50</td>
<td>10.029412</td>
<td>-0.529412</td>
<td>0.280277</td>
<td>4.00</td>
<td>16.00</td>
<td></td>
</tr>
</tbody>
</table>

Average of Y: 5.5

SSE: 1.67

SST: 39.50

$R^2 = 1 - (\text{SSE/\text{SST}})$

0.957743857 is the $R$ squared

From: Gottfried, page 106

Spreadsheet online: [http://home.business.utah.edu/bebrpsp/URPL5020/Matrix/LinearRegression.xls](http://home.business.utah.edu/bebrpsp/URPL5020/Matrix/LinearRegression.xls)
“Add a Trend Line” in Excel

- Plot the data in an x-y scatter
- Right click the series on the graph
- Select “Add a Trend Line”
- Select “Linear” from the Type tab
- Select “Display Equation” and “Display r squared” from the “Options” tab
- Note: for other applications you may select “Forecast” as well
Using the Analysis Tool Pack

- Enter the data into the worksheet
- From “Tools” menu, select “Data Analysis/Regression Tool.
- Complete the required selections.
Results from Analysis Tool Pack

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>SUMMARY OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.5</td>
<td></td>
</tr>
</tbody>
</table>

**Regression Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.978643887</td>
</tr>
<tr>
<td>R Square</td>
<td>0.957743857</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.943658476</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.745903847</td>
</tr>
<tr>
<td>Observations</td>
<td>5</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Source</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>37.83088235</td>
<td>37.83088</td>
<td>67.99559</td>
<td>0.003734402</td>
</tr>
<tr>
<td>Residual</td>
<td>3</td>
<td>1.669117647</td>
<td>0.556373</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>39.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
<th>Lower 95.0%</th>
<th>Upper 95.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.279411765</td>
<td>0.610944132</td>
<td>2.094155</td>
<td>0.127272</td>
<td>-0.664886955</td>
<td>3.223710484</td>
<td>-0.664886955</td>
<td>3.223710484</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.514705882</td>
<td>0.062419278</td>
<td>8.245944</td>
<td>0.003734</td>
<td>0.316059694</td>
<td>0.71335207</td>
<td>0.316059694</td>
<td>0.71335207</td>
</tr>
</tbody>
</table>
Simple Linear Regression - Summary

- Simple linear regression fits a line to a set of x,y coordinates.
- This procedure minimizes squared errors.
- $r^2$ is a measure of goodness of fit
  - The better the fit, the closer $r^2$ is to 1
- There are multiple ways to compute linear regression in Excel.
Part 3: Calculating “Forecasted” Values from Fitted Functions in Excel

http://home.business.utah.edu/bebrsp/URPL5020/Trend/CalcExtrap.xls

Worksheet examples
Forecasted Values in Excel

- Select data series
- Right click
- Add a trendline
- Select Type
  - Linear
- Go to “Options” tab
Forecasted Values in Excel

Options Menu

- Forecast period
- Display equation
- Display R-squared
- Note: You can customize trend label
- Click “OK”
State of Utah Population

Tip: Select the formula label, then format, and increase the number of digits to the largest possible. This will result in a more precise computation.

\[ y = 29481.450370525x + 316996.533799534 \]

\[ R^2 = 0.962332124 \]
State of Utah Population

\[ y = 29,481.45x + 31,6996.53 \]

\[ R^2 = 0.962 \]

Tip: For final presentation purposes, reduce the number of digits displayed in the equation.
Calculating Forecasted Values

- Calculate the forecasted population for the year 2020.
- Equation:
  - \( y = 29481.450370525x + 316996.533799534 \)
    - \( Y = \) population
    - \( X = \) time marker – substitute the year (?)
  \[
  29481.450370525 \times (2020) + 316996.533799534 \\
  = 59,869,526
  \]
  This is much too high.
Calculating Forecasted Values

- Data series is 1940 through 2020.
- If the actual year does not work, create an index for each year, incrementing by 1.
  - 1940 = 1, 1941 = 2, etc.
  - 2020 = 81

$$29481.450370525 \times (81) + 316996.533799534 = 2,704,994$$
- This is the correct formula.
- You can determine how your version of excel interprets the “x” in your equations by experimenting.
- See example spreadsheet. CalcExtrap.xls
Residuals: Forecast Minus Actual

![Graph showing residuals over time]

-350,000
-300,000
-250,000
-200,000
-150,000
-100,000
-50,000
0
50,000
100,000
150,000
200,000

Residuals: Forecast Minus Actual

- Linear regression on residuals collapses to the x axis.
- The sum of the residuals is zero.
Ratio Methods

- Smith, Tayman, Swanson – Chapter 8
- Smaller region (city) is contained in larger region (county or state)
- Projection of larger region $\rightarrow$ projection of smaller region
- Depends upon a preexisting forecast / projection of the larger region
- These will be used in the economic models section of the course.
Types of Ratio Methods

- Constant share: small area maintains same growth rate as larger area
- Shift share: trend in small area’s share of region is extended into the future
  - Observed differential growth rates are maintained
- Share-of-growth: small area’s share of larger region’s growth is maintained.
Extrapolation - Summary

- Use with care.
  - Just because a function “fits” (high $r^2$) does not mean that the extrapolation is reasonable.
  - Make your assumptions explicit
  - Generally there are growth limits at some point

- Explore various approaches.
  - Create your own functions in excel based on your knowledge of the area (growth limits, etc.)
  - Use Excel to fit a trend line and extrapolate it into the future

- Calculation of the forecast value when using Excel may require the construction of an index.
  - Use the reported equation and substitute either the year or index number into the formula for “x”.
  - If you create an index, the beginning value should be 1.