Regression Analysis

Larry Caretto
Mechanical Engineering 309
Numerical Analysis of Engineering Systems
March 19, 2014

Outline

• Regression with two variables, x and y
• Confidence limits for regression results
  – Student’s t distribution
• MATLAB and Excel functions for regression
• Data transformations
• Multilinear regression
  – Use of LINEST for multilinear regression
• Fifth programming assignment

Regression Analysis

Fifth programming assignment

Typically have many more data points than constants in the regression

Regression vs. Interpolation

• Interpolation seeks to pass an approximation function through every data point
  – Useful when we trust the data absolutely
  – Usually used with accurate tabular data to determine intermediate points not in the table
  – Used in numerical analysis applications such as numerical integration
• Less important with computer calculations of complex engineering functions

Regression (fitting data)

• Regression is a statistical approach that seeks an approximate relationship among variables
• Looks for measure of error in result
• Used to determine trends in experimental data that have some uncertainty
• Not necessary (as in interpolation) for relationship to pass through data points
• Typically have many more data points than constants in the regression

Linear Regression

• Seeks approximate linear relationship among data set (x, y)
• Fit equation: \( \hat{y}_i = a + bx_i \)
• Notation \( \hat{y}_i \) indicates approximate value, which may be different from data \( y_i \)
• Equations for a and b based on minimizing sum of squares of differences between actual and approximate data

\[
S = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2 \rightarrow \text{min}
\]

Finding a and b

• To get the minimum take partial derivatives with respect to a and b and set them to zero

\[
\frac{\partial S}{\partial a} = \sum_{i=1}^{N} (y_i - a - bx_i) = 2 \sum_{i=1}^{N} y_i - 2a N - 2b \sum_{i=1}^{N} x_i = 0
\]

\[
\frac{\partial S}{\partial b} = \sum_{i=1}^{N} (y_i - a - bx_i)x_i = \sum_{i=1}^{N} x_i y_i - 2a \sum_{i=1}^{N} x_i - 2b \sum_{i=1}^{N} x_i^2 = 0
\]

• Solve the first equation for a and substitute it into the second equation

\[
a = -b \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N \sum_{i=1}^{N} x_i^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \quad \bar{x} \text{ and } \bar{y} \text{ are mean values}
\]
**Equations for a and b**

- Substitute equation for a into equation for b (both copied below) and solve for b

\[ a = \frac{\sum y_i - b \sum x_i}{N} \]

\[ b = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2} \]

- First solve for b then solve for a
  - Can set a = 0 to force line through origin
  - Can use equations with all sums or means

**Measures of Error**

- Standard Error, \( s_{y|x} \), and \( R^2 \)

\[ s_{y|x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N-2}} \]

\[ R^2 = 1 - \frac{(N-2)s_{y|x}^2}{\sum (y_i - \bar{y})^2} \]

- \( R^2 \) varies between 0 and 1 and is a measure of how well the regression equation explains the data variation
  - \( R^2 = 1 \) means the regression is perfect and \( R^2 = 0 \) means the regression is useless
  - The \( R^2 \) desired in a particular problem depends on experience with typical data for that problem

**Example Calculation**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x^2</th>
<th>y^2</th>
<th>xy</th>
<th>( y_i - \bar{y} )</th>
<th>( a )</th>
<th>( b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>100</td>
<td>10</td>
<td>7.4</td>
<td>-4.10</td>
<td>12.10</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>4</td>
<td>361</td>
<td>38</td>
<td>19.3</td>
<td>0.25</td>
<td>1.16</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>9</td>
<td>729</td>
<td>28</td>
<td>18.1</td>
<td>1.12</td>
<td>1.16</td>
</tr>
<tr>
<td>4</td>
<td>44</td>
<td>16</td>
<td>1936</td>
<td>38</td>
<td>43.7</td>
<td>0.09</td>
<td>1.16</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>25</td>
<td>3364</td>
<td>59</td>
<td>55.8</td>
<td>4.84</td>
<td>1.16</td>
</tr>
<tr>
<td>15</td>
<td>158</td>
<td>225</td>
<td>5995</td>
<td>180</td>
<td>595</td>
<td>33.18</td>
<td>1.16</td>
</tr>
</tbody>
</table>

\[ b = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2} = \frac{15(158) - 15(15)}{15(225) - (15)^2} = 12.100 \]

\[ a = \frac{\sum y_i - b \sum x_i}{N} = \frac{1.12 - 12.10(15)}{5} = -4.10 \]

\[ s_{y|x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{595 - 6.76}{4}} = 12.100 \]

\[ R^2 = 1 - \frac{(N-2)s_{y|x}^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{(5-2)(12.100)^2}{595} = 0.9779 \]

**Confidence Limits**

- What is the statistical uncertainty in calculated values of a or b?

  - Based on use of Student’s t distribution
    - Important statistical distribution for determining uncertainty in many applications
    - Distribution depends on random variable, t, and degrees of freedom, \( \nu \)
    - Define \( t_{\alpha, \nu} \) as the point in the distribution where the probability that \( t \geq t_{\alpha, \nu} \) is \( \alpha \)
    - Find \( t_{\alpha, \nu} \) from tables or computer functions

**Confidence Limits II**

- The confidence limits for the regression parameters a and b are given by the following equations
  - For using n data points and probability \( (1 - \alpha) \) that the limits are correct
    - Typically choose \( \alpha = .05 \) for 95% confidence

\[ b \pm t_{\alpha/2, \nu} s_{y|x} \sqrt{\sum \frac{(x_i - \bar{x})^2}{N-2}} \]

\[ a \pm t_{\alpha/2, \nu} s_{y|x} \sqrt{\frac{1}{N} \left( \sum (x_i - \bar{x})^2 \right)} \]

**Student’s t Distribution**

- \( \nu = 5 \)
- \( \nu = 30 \)
t-Distribution Functions

- Excel: t.inv.2t(α,ν) returns the value \( t_{α/2,ν} \) such that the probability that the absolute value of a random \( t \)-distributed variable, \( |t| \), is greater than \( t_{α/2,ν} \) is \( α \).
- MATLAB: The function tinv(1 − α/2, ν) returns the value \( t_{α/2,ν} \) such that the probability that a random \( t \)-distributed variable, \( t \), is greater than \( t_{α/2,ν} \) is \( 1 - \alpha/2 \).

\[ t \approx tinv(.975, 3) \]

Excel: TINV(0.05, 3)  
\[ \text{ans} = 3.1824 \]

Excel compatibility function is TINV

Critical \( t \)-Distribution Values and Equivalents

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>60%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
<th>99.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-sided</td>
<td>0.800</td>
<td>0.900</td>
<td>0.950</td>
<td>0.975</td>
<td>0.990</td>
<td>0.995</td>
<td>0.9999</td>
</tr>
<tr>
<td>2-sided</td>
<td>0.800</td>
<td>0.900</td>
<td>0.950</td>
<td>0.975</td>
<td>0.990</td>
<td>0.995</td>
<td>0.9999</td>
</tr>
<tr>
<td>MATLAB</td>
<td>0.800</td>
<td>0.900</td>
<td>0.950</td>
<td>0.975</td>
<td>0.990</td>
<td>0.995</td>
<td>0.9999</td>
</tr>
<tr>
<td>MathTable</td>
<td>0.800</td>
<td>0.900</td>
<td>0.950</td>
<td>0.975</td>
<td>0.990</td>
<td>0.995</td>
<td>0.9999</td>
</tr>
<tr>
<td>( n )</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>50</td>
<td>100</td>
<td>1500</td>
</tr>
<tr>
<td>( t )</td>
<td>1.645</td>
<td>2.447</td>
<td>2.660</td>
<td>2.796</td>
<td>3.085</td>
<td>3.839</td>
<td>4.303</td>
</tr>
</tbody>
</table>

95% probability that \( |t| \leq 2.131 \) and 97.5% probability that \( |t| \leq 2.131 \)

Back to Example

- Get confidence limits on \( a \) and \( b \)
  \[ b \pm t_{α/2,ν} \frac{s_y}{\sqrt{\sum (x_i - \bar{x})^2}} \]
  \[ a \pm t_{α/2,ν} \frac{\bar{y}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}} \]

- Pick \( α = 0.05 \)
  \[ |t|_{0.05,2,5} = 3.1824 \]
  \[ a = 10.82 \pm (3.1824)(1.05) = 10.82 \pm 3.34 \]
  \[ b = 0.845 \pm (3.1824)(3.65) = 0.845 \pm 3.30 \]

- Note that \( a = 0 \) is within the confidence limits
Regression Analysis

March 19, 2014

Prediction Confidence Limits

- What does \( \hat{y}_i = a + bx_i \) predict?
  - Average value of \( y^* \) for several observations of the same \( x^* \)

\[
\text{Average} \quad a + bx \pm \tilde{t}_{i,2,df} \left( s_{\hat{y}|x} \right) = \left( a + bx \right) \pm \tilde{t}_{i,2,df} \sqrt{ \frac{s^2_{\text{res}}}{N} + \frac{s^2_{\text{res}}}{N} \left( \frac{1}{N} - \frac{1}{N^2} \right) (x_i - \bar{x})^2 }.
\]
  - One observation

Uncertainty increases as we get further from the mean \( x \).

Excel LINEST Function

- Array function for linear regression
- Returns slope, intercept, standard errors for slope and intercept, \( R^2 \), \( s_{\hat{y}|x} \), and various regression statistics
- Select area of two column and five rows
- Enter formula =LINEST( \( y \) array, \( x \) array, zero choice, TRUE)
  - Set zero choice to FALSE to force regression through origin (default for omitted zero choice is TRUE)

LINEST Worksheet Results

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Slope</td>
<td>Intercept</td>
</tr>
<tr>
<td>2</td>
<td>Standard Error in Slope ( (SE_b) )</td>
<td>Standard Error in Intercept ( (SE_a) )</td>
</tr>
<tr>
<td>3</td>
<td>( R^2 )</td>
<td>( s_{\text{res}} )</td>
</tr>
<tr>
<td>4</td>
<td>F statistic</td>
<td>Degrees of Freedom (df)</td>
</tr>
<tr>
<td>5</td>
<td>Regression sum of squares*</td>
<td>Residual sum of squares*</td>
</tr>
</tbody>
</table>

**Rule of thumb:** The slope and intercept should be at least twice their standard errors to be significantly different from zero

\[
\text{ConfLimits} = \text{Coefficient} \pm \tilde{t}_{i,2,df} \left( \text{Standard Error} \right) \]

*Definitions on next slide

MATLAB Regression

\[
x = 1 \quad 2 \quad 3 \quad 4 \quad 5 \\
y = 10 \quad 19 \quad 27 \quad 44 \quad 58
\]

\[
>> p = polyfit(x,y,1) \\
p = 12.1000 \quad -4.7000 \quad \% \text{ slope intercept}
\]

\[
>> yfit=polyval(p,x) \\
yfit = 7.400 \quad 19.500 \quad 31.600 \quad 43.700 \quad 55.800
\]

\[
>> \text{yresid} = y-yfit; \\
>> \text{ssresid} = \sum(yresid.^2) \\
>> \text{sstotal} = (\text{length}(y)-1)*\text{var}(y) \\
\text{Rsqd} = 1-\frac{\text{ssresid}}{\text{sstotal}}
\]

\[
\text{Rsqd} = 0.9779
\]

\[
>> [r m b] = \text{regression}(x,y) \\
r = 0.9889 \quad m = 12.1000 \quad b = -4.700
\]

\[
\text{lineest sum of squares} = \text{ssresid} - \frac{\text{lineest sum of squares}}{\text{sstotal}}
\]

MATLAB 2nd-order polyfit

\[
>> x=1:5; \\
>> y = [10 19 27 44 58];
\]

\[
>> p=polyfit(x,y,2)
\]

\[
p = 1.3571 \quad 3.9571 \quad 4.8000
\]

\[
>> xx = 1:.1:5; \\
>> yfit = polyval(p,xx); \\
>> \text{plot (x,y,‘o’,xx,yfit)}
\]

Polyval evaluates a polynomial with coefficients \( p_i \) to \( p_{n+1} \) for input \( x \), which is a row matrix here

Data Transformations

- Linear Regression can be used for nonlinear problems if they can be transformed into linear ones
  - From data on \( k \) and \( T \), fit \( A \) and \( E \) in the following equation: \( k = Ae^{-E/RT} \) (R known)
  - Solution: \( \ln(k) = \ln(A) - E/RT \)
  - For \( y = a + bx \) with \( y = \ln(K) \) and \( x = 1/T \), a regression gives \( a = \ln(k) \) and \( b = -E/R \)
  - Several different transformations possible
Multilinear Regression

- Want to examine case where more than one variable affects an outcome
  - Example is emissions from diesel engine that depends on fuel properties
    - emissions = $b_0 + b_1(cetane) + b_2(\text{aromatics}) + b_3(\text{density})$
  - Use measured data on emissions, cetane, aromatics to find $b_0, b_1, b_2, b_3$
- Another example: CSUN_GPA = $b_0 + b_1(\text{HS_GPA}) + b_2(\text{SAT_MATH}) + b_3(\text{SAT_VERBAL})$

General Regression

- Use notation so we can write code for any number of predictive variables
- Call predictive variables $x_1, x_2, x_3, \ldots$
- Call response variable $y$
  - In previous emissions example, $x_1 = \text{cetane}, x_2 = \text{aromatics}, x_3 = \text{density}, \text{and } y = \text{emissions}$
  - For three variables the equation is $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

General Equation and Data

- In general we can have $K$ predictive variables, $x_1$ to $x_K$
- General model equation: $y = b_0 + \sum_{j=1}^{K} b_j x_j$
- How do we represent the data?
  - Each data set consists of one value of $y$ and one value for each of the $x_j$ variables
  - For data set $m$, we can call the value of $y_m$, and we can call the value of $x_j$ for data set $m$ as $x_{jm}$

Data Set with $K = 3$ and $N = 8$

- We use $K$ different variables ($x_1$ to $x_K$) to predict the value of another variable, $y$
- We have $N$ sets of data
  - Numbered from $m = 0$ to $m = N - 1$
  - Each data set has one value of $y$, called $y_m$
  - One value of each $x_j$, called $x_{jm}$
  - $x$ and $y$ data usually from file input
  - All data used to determine $b_0$ to $b_K$
- Derive equations for $b_0$ to $b_K$ by minimizing sum of squares of $y - \hat{y}_{\text{regression}}$

Summary of Data

- Each data set, $m$, has a value for $y$ and each $x_j$
- What are $y_4, x_{26}, x_{17}$?

How do we find $b_j$?

- The derived equations for the $b_j$ are solved in the following manner
- Define $x_{0m} = 1$ for all $m = 0, \ldots , N - 1$
  - Note that there is no $x_0$ in model
  - Setting $x_{0m} = 1$ used to simplify equations
- Values of $b_0, \ldots , b_K$ found by solving $K + 1$ simultaneous linear equations $A_0 b_0 + A_1 b_1 + A_2 b_2 + \ldots + A_K b_K = c_i$ ($i = 0, \ldots , K$)
  - Compute $A_j$ and $c_i$ from input data
  - Use Gaussian elimination to solve equations
Regression Analysis

March 19, 2014

ME 309 – Numerical Analysis of Engineering Systems

Equations to be Used

- Compute the $A_{ij}$ coefficients
  \[ A_{ij} = \sum_{m=0}^{N-1} x_{im} x_{jm} \]

- Compute the $c_i$ coefficients
  \[ c_i = \sum_{m=0}^{N-1} x_{im} y_m \]

- Use Gaussian elimination routine to solve for the $b_j$
  \[ \sum_{j=0}^{K} A_{ij} b_j = c_i \quad i = 0, \ldots, K \]

More Equations to be Used

- Compute estimated $y$ values
  \[ \hat{y}_m = b_0 + \sum_{j=1}^{K} b_j x_{jm} \]

- Compute the $R^2$ value
  \[ R^2 = 1 - \frac{\sum_{m=0}^{N-1} (y_m - \hat{y}_m)^2}{\sum_{m=0}^{N-1} y_m^2} - N \left( \bar{y} \right)^2 \]

Multilinear LINEST

- Excel LINEST can handle multilinear regressions
- To fit $K$ different independent variables
  - Select a range of $K+1$ columns and five rows and enter the following formula
    \[ =\text{LINEST}(y\text{_range, range\_for\_all\_x, zero\_choice, TRUE}) \]
  - Press control+shift+enter
  - Data for all $x$ variables must be in adjoining columns

Excel LINEST Function

- General formula is
  \[ =\text{LINEST}(y\text{Range}, x\text{Range}, zero?, TRUE) \]
- Select $K+1$ columns and 5 rows for formula
- In this example
  \[ =\text{LINEST}(A2:A9, B2:D9, , TRUE) \]
- Control+Shift+Enter
- Results on slide after next

LINEST Results

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope K</td>
<td>Slope K – 1</td>
<td>Slope K – 2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Standard Error in Slope ($\text{SE}_{K}$)</td>
<td>Standard Error in Slope ($\text{SE}_{K-1}$)</td>
<td>Standard Error ($\text{SE}_{K-2}$)</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$R^2$</td>
<td>$s_{\text{res}}$</td>
<td>#N/A</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>F statistic</td>
<td>Degrees of Freedom</td>
<td>#N/A</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Regression sum of squares</td>
<td>Residual sum of squares</td>
<td>#N/A</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Rule of thumb: The slope and intercept should be at least twice their standard errors to be significantly different from zero

Example Results

<table>
<thead>
<tr>
<th>$k_1$</th>
<th>slope</th>
<th>$k_2$</th>
<th>Slope</th>
<th>$k_3$</th>
<th>slope</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>-0.00509</td>
<td>0.00940</td>
<td>-0.5146</td>
<td>2.3966</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std err</td>
<td>0.00429</td>
<td>0.00448</td>
<td>0.4216</td>
<td>1.3503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>$s_{\text{res}}$</td>
<td>0.579</td>
<td>0.444</td>
<td>#N/A</td>
<td>#N/A</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>df</td>
<td>1.835</td>
<td>4</td>
<td>#N/A</td>
<td>#N/A</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>1.087</td>
<td>0.789</td>
<td>#N/A</td>
<td>#N/A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $F \mid \text{df}$ indicates $F$ statistic and degrees of freedom in two adjacent columns
- SS is sum of squares terms
- Bad fit because of large standard errors
Fifth Programming Assignment

- Use MATLAB tools to generate cubic spline and plot results
- Use MATLAB tools to generate Newton polynomial and plot results
- Write a VBA program for Newton polynomials
- Use LINEST for multilinear regression in Excel
- Do calculations with finite difference derivative expressions