Regression Analysis

Larry Caretto
Mechanical Engineering 309

Numerical Analysis of Engineering Systems

March 19, 2014

California State University Northridge

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

California State University Northridge 2

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

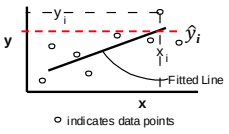
California State University Northridge 3

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

California State University Northridge 4

Linear Regression



- Seeks approximate linear relationship among data set (x_i, y_i)
- 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

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

California State University Northridge 5

Finding a and b

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

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

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

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

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad a = \frac{\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} x_i}{N} = \bar{x} - b\bar{y}$$

$\bar{x}$ and $\bar{y}$ are mean values

California State University Northridge 6

Equations for a and b

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

$$a = \frac{\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} x_i}{N}$$

$$2\sum_{i=1}^{N} x_i y_i - 2a\sum_{i=1}^{N} x_i - 2b\sum_{i=1}^{N} x_i^2 = 0$$

$$b = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\sum_{i=1}^{N} x_i y_i - N(\bar{x})\bar{y}}{\sum_{i=1}^{N} x_i^2 - N(\bar{x})^2}$$

- 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

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}} \qquad R^2 = 1 - \frac{(N-2)s_{y|x}^2}{\sum_{i=1}^{N} y_i^2 - N(\bar{y})^2}$$

Numerator in R^2 called residual sum of squares; denominator called total sum of squares

- 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

x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	10	1	100	10	7.4	6.76
2	19	4	361	38	19.5	0.25
3	27	9	729	81	31.6	21.16
4	44	16	1936	176	43.7	0.09
5	58	25	3364	290	55.8	4.84
15	**158**	**55**	**6490**	**595**	**sums**	**33.10**

$$b = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{5(595) - (15)(158)}{5(55) - (15)^2} = 12.100$$

$$a = \frac{\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} x_i}{5} = \frac{158 - 12.1(15)}{5} = -4.710$$

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{33.10}{5-2}} = 3.3216$$

$$R^2 = 1 - \frac{(N-2)s_{y|x}^2}{\sum_{i=1}^{N} y_i^2 - N(\bar{y})^2} = 1 - \frac{(5-2)3.3216^2}{6490 - 5(158/5)^2} = 0.9779$$

Confidence Limits

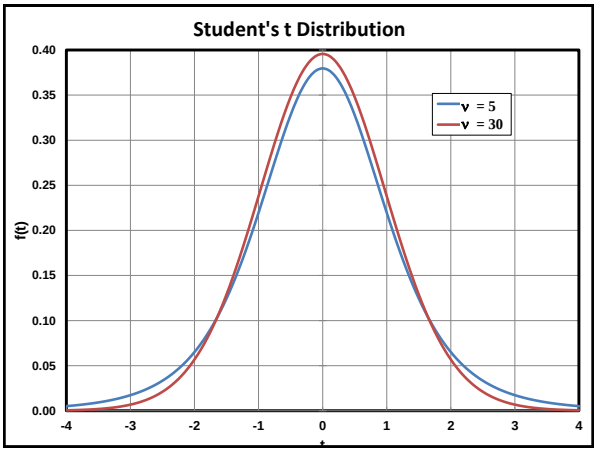
- What is the statistical uncertainty in calculated values of a or b?
- Based on used of Student's t distribution
  - Important statistical distribution for determining uncertainty in many applications
  - Distribution depends on random variable, t, and degrees of freedom, ν
  - Define $t_{\alpha,\nu}$ as the point in the distribution where the probability that $t \geq t_{\alpha,\nu}$ is α
  - Find $t_{\alpha,\nu}$ from tables or computer functions

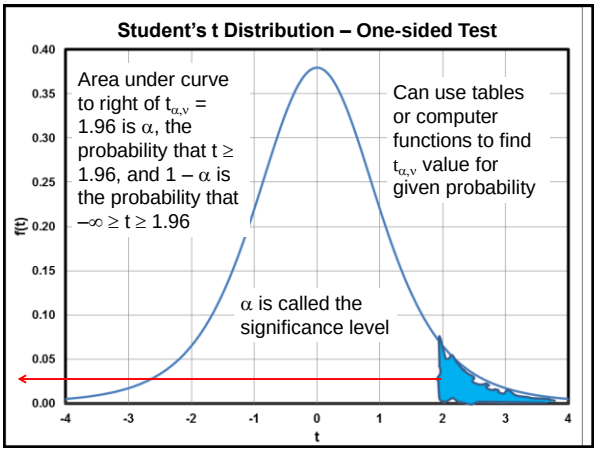
Confidence Limits II

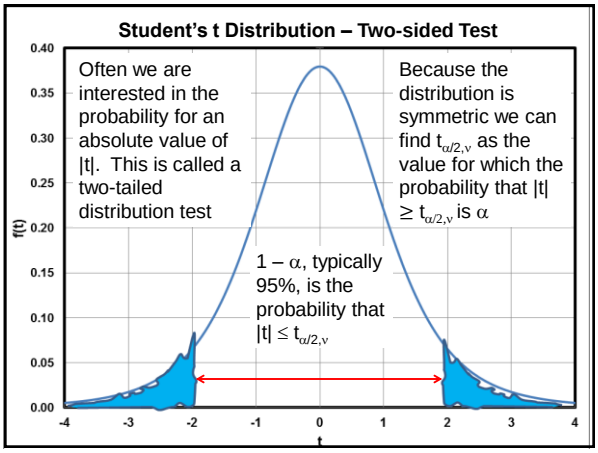
- The confidence limits for the regression parameters a and b are given by the following equations
  - For using n data points and probability $(1 - \alpha)$ that the limits are correct
    - Typically choose $\alpha = .05$ for 95% confidence

$$b \pm t_{\alpha/2,n-2} \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2}} \qquad a \pm t_{\alpha/2,n-2} s_{y|x} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

Standard errors for a and b







t-Distribution Functions

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

```
>> tinv(.975,3)
ans =  3.1824
```

Excel: TINV(0.05, 3) = 3.1824

California State University Northridge

Excel compatibility function is TINV

15

Student's t-Distribution Table

Critical t Distribution Values and Equivalents							
1-sided	80%	90%	95%	97.5%	99%	99.5%	99.9%
2-sided	60%	80%	90%	95%	98%	99%	99.8%
MATLAB	0.800	0.900	0.950	0.975	0.990	0.995	0.999
Excel	0.40	0.20	0.10	0.05	0.02	0.01	0.002
MathTable	0.2	0.1	0.05	0.025	0.01	0.005	0.001
ν = 1	1.376	3.078	6.314	12.71	31.82	63.66	318.3
ν = 5	0.920	1.476	2.015	2.571	3.365	4.032	5.893
ν = 10	0.879	1.372	1.812	2.228	2.764	3.169	4.144
ν = 15	0.866	1.341	1.753	2.131	2.602	2.947	3.733
ν = 30	0.854	1.310	1.697	2.042	2.457	2.750	3.385
ν = 50	0.849	1.299	1.676	2.009	2.403	2.678	3.261
ν = 100	0.845	1.290	1.660	1.984	2.364	2.626	3.174
ν = 1000	0.842	1.282	1.646	1.962	2.330	2.581	3.098

95% probability that |t| ≤ 2.131 and 97.5% probability that t ≤ 2.131

California State University Northridge

16

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm

Back to Example

- Get confidence limits on a and b

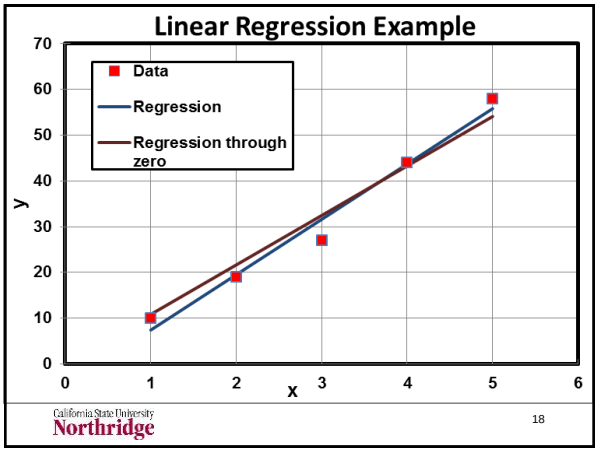
$$b \pm t_{\alpha/2,n-2}\frac{s_{y|x}}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x}_i)^2}} \qquad \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x}_i)^2}} = \frac{3.321646}{\sqrt{10}} = 1.050397$$

$$a \pm t_{\alpha/2,n-2}s_{y|x}\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}} \qquad s_{y|x}\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}} = (3.321646)\sqrt{\frac{1}{5}+\frac{(15/5)^2}{10}} = 3.6538$$

- Pick α = 0.05, $t_{0.05/2,5-2}$ = 3.1824
  - b = 10.82 ± (3.1824)(1.05) = 10.82±3.34
  - a = –4.7 ± (3.1824)(3.65) = –4.7±11.6
    - Note that a = 0 is within the confidence limits

California State University Northridge

17



California State University Northridge

18

Prediction Confidence Limits

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

Average $\overline{a+bx^*} \pm t_{\alpha/2,n-2} s_{y|x} \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}$

One observation $a+bx^* \pm t_{\alpha/2,n-2} s_{y|x} \sqrt{1+\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}$

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, $s_{y|x}$, 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

	A	B
1	Slope	Intercept
2	Standard Error in Slope(SE_b)	Standard Error in Intercept (SE_a)
3	R^2	$s_{y\|x}$
4	F statistic	Degrees of Freedom(df)
5	Regression sum of squares*	Residual sum of squares*

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

ConfLimits = Coefficient ± $t_{\alpha/2,df}$(Standard Error)

*Definitions on next slide

MATLAB Regression

```
x =  1     2     3     4     5
y = 10    19    27    44    58
>> p = polyfit(x,y,1)
p =  12.1000   -4.7000  % slope intercept
>> yfit=polyval(p,x)
yfit = 7.400 19.500 31.600 43.700 55.800
>> yresid = y - yfit;
>> ssresid= sum(yresid.^2);
>> sstotal=(length(y)-1)*var(y)
>> Rsqd=1-ssresid/sstotal
Rsqd = 0.9779
>> [r m b] = regression(x,y)
r =    0.9889   m =  12.1000 b =  -4.700
```

ssresid is residual sum of squares from linest

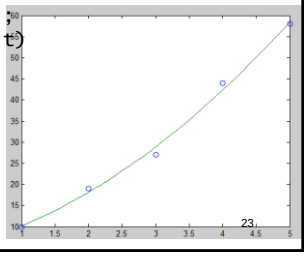
linest regression sum of squares = sstotal - ssresid

MATLAB 2nd-order polyfit

```
>> x=1:5;
>> y= {10 19   27   44   58];
>> p=polyfit(x,y,2)
p =  1.3571  3.9571  4.8000
>> xx = 1:.1:5;
>> yfit = polyval(p,xx);
>> plot (x,y,'o',xx,yfit)
```

Polyval evaluates a polynomial with coefficients p_1 to p_{n+1} for input x, which is a row matrix here

Polyfit returns coefficients p(k) such that p(x) = $p_1x^n + p_2x^{n-1} + \cdots + p_{n+1}$; here p(x) = $1.3571x^2$ + 3.9751x + 4.8000



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 that affects an outcome
  - Example is emissions from diesel engine that depends on fuel properties
    - emissions = $b_0 + b_1$(cetane) + b_2(aromatics) + b_3(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_1HS_GPA + b_2SAT_MATH + b_3SAT_VERBAL

California State University Northridge 25

General Regression

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

California State University Northridge 26

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, y_m, and we can call the value of x_j for data set m x_{jm}

California State University Northridge 27

Data Set with K = 3 and N = 8

m	y_0 y	x_1	x_2	x_3
0	2.55	3.00	440	500
1	1.95	3.47	350	400
2	1.89	3.14	440	540
3	2.24	3.46	350	370
4	2.31	3.59	450	480
5	1.74	1.75	200	320
6	1.87	3.03	310	470
7	0.83	3.18	290	400

x_{12} x_{35}

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

California State University Northridge 28

Summary of Data

- 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, and 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 $y - y_{regression}$

California State University Northridge 29

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, ... , N – 1
  - Note that there is no x_0 in model
  - Setting $x_{0m} = 1$ used to simplify equations
- Values of b_0, ... , b_K found by solving K + 1 simultaneous linear equations

$A_{i0} b_0 + A_{i1} b_1 + A_{i2} b_2 + \ldots + A_{iK} b_K = c_i \quad (i = 0, \ldots, K)$

  - Compute A_{ij} and c_i from input data
  - Use Gaussian elimination solve equations

California State University Northridge 30

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* $\sum_{j=0}^{K} A_{ij} b_j = c_i \quad i = 0, \ldots, K$

California State University Northridge

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*

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

Will need these equations for programming assignment five

California State University Northridge

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

  =LINEST(y_range, range_for_all_x, zero_choice, TRUE)
  - Press control+shift+enter
  - Data for all x variables must be in adjoining columns

California State University Northridge

Excel LINEST Function

	A	B	C	D
1	y	x_1	x_2	x_3
2	2.55	3.00	440	500
3	1.95	3.47	350	400
4	1.89	3.14	440	540
5	2.24	3.46	350	370
6	2.31	3.59	450	480
7	1.74	1.75	200	320
8	1.87	3.03	310	470
9	0.83	3.18	290	400

Bold red letters and numbers are worksheet rows and columns

- General formula is =LINEST(yRange, xRange, zero?, TRUE)
- Select K+1 columns and 5 rows for formula
- In this example

  =LINEST(A2:A9, B2:D9,, TRUE)
- Control+Shift+Enter
- Results on slide after next

California State University Northridge

LINEST Results

	A	B	C
1	Slope K	Slope K – 1	Slope K – 2
2	Standard Error in Slope K(SE_K)	Standard Error in Slope (SE_{K-1})	Standard Error (SE_{K-2})
3	R^2	$s_{y\|x}$	#N/A
4	F statistic	Degrees of Freedom	#N/A
5	Regression sum of squares	Residual sum of squares	#N/A

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

California State University Northridge

Example Reslts

	x_3 slope	x_2 Slope	x_1 Slope	Intercept
Value	-0.00509	0.00940	-0.5146	2.3966
Std err	0.00429	0.00448	0.4216	1.3503
R^2 | $s_{y\|x}$	0.579	0.444	#N/A	#N/A
F | df	1.835	4	#N/A	#N/A
SS	1.087	0.789	#N/A	#N/A

- F | 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

California State University Northridge

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

California State University Northridge

37