CMPT 155: Computer Applications for Life Sciences

Lecture 9: Regression

Ivan E. Perez

March 4, 2022

Presentation Outline

Homework And Administrative

- Midterm 2 is on April 13th.
- You may bring a 1 page cheat sheet print or handwritten.
- Homework 5 is due ???

Trendline (Regression Line)

- Data analysis is about revealing patterns and facts about data sets.
- Linear regression is a technique describes outputs y as linearly dependent on inputs x.

Trendline Example

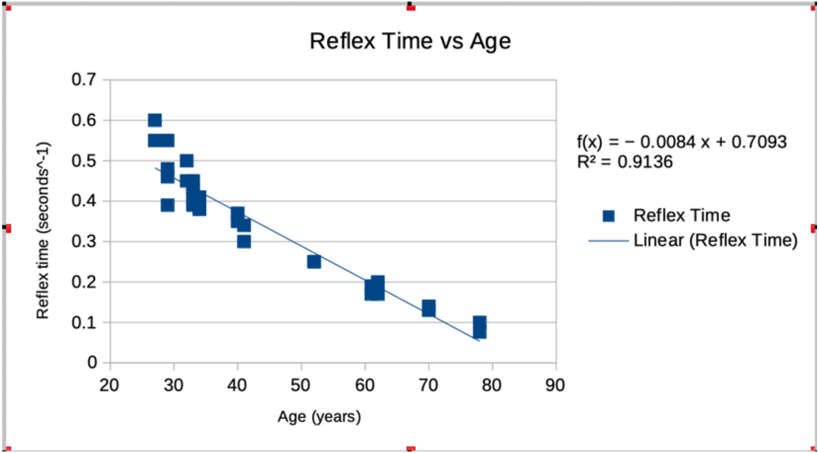

Figure: Example Linear Regression of Age against Reflex Time

Adding a Trendline

A Trendline can be added to XY (Scatter) graphs.
Lets try plotting *CancerStudy.xlsx* and adding a Trendline:

1. Download *CancerStudy.xlsx* from moodle.
2. Select cells `B2:C18`
3. Navigate to the XY Scatter Plot, and create scatter plot.
4. be sure to include a chart title and axis titles.
5. To add a trendline we can:
    - 'Chart Tools' → 'Add Chart Element' → 'Linear'.
    - Select the Datapoints → right(Ctrl)-click → 'Add Trendline'

Types of Trendlines

Type	Excel Option	General Form
Linear	Linear	$y = mx + b$
Exponential	Exponential	$y = ae^{bx}$
Quadratic	Polynomial deg=2	$y = ax^2 + bx + c$
Cubic	Polynomial deg=3	$y = ax^3 + bx^2 + cx + d$

When computing predictions in excel, we can write trendline equations in the following fashion.

Type	Excel Form
Linear	`m * X + b`
Exponential	`(a)*EXP((b)*x)`
Quadratic	`(a)*(X^2) + (b)*X + c`
Cubic	`(a)*(X^3) +(b)*(X^2) + (c)*X` $+$ `d`

Where `X` is a *cell reference* to an x-value in the dataset and `a,b,c,d, m` are real numbers that can be written as a decimal or in scientific notation $1.234 * 10^{-4}$, (i.e., `1.234 * (10^-4)`)

Formatting Trendlines

Trendlines can be formatted through the 'Format Trendline' pane, which can be accessed by:

1. Selecting Data by left clicking a point on the XY scatter graph.
2. Right(Ctrl) clicking the points
3. Selecting the 'Format Trendline' menu option.

Trendlines modifitcations include:

- Equation and Displayed Statistics.
  - The equation of the trendline
  - whether to display the equation
  - whether to display correlation squared, R^2.
- Thickness and color
  - Thickness, color, effects, labelling
- Extrapolation
  - The range of the trendline is within the dataset by default
  - can be extraoplated by increasing the Forward/Backward periods option

Analyzing Trendlines

After specifying a model we must be able to specify, analyze and measure models. Common functions for Analyzing a *linear* trendline include:

- slope using SLOPE()
- y- intercept using INTERCEPT()
- Correlation, R,using CORRELL()

Non Linear Trendlines can be analyzed using *Sum of Squared Residuals*,SSR.

SLOPE()

SLOPE() - computes the slope of a linear regression line for a collection of x and y values.

- inputs
  - `known_ys` : selection
    array/selection of known y values
  - `known_xs : selection`
    array/selection of known x values
- outputs
  - computed slope : numeric
    estimated slope for a linear regression line for the given data.

INTERCEPT()

Compute the y intercept of a linear regression line for a collection of x and y values.

- inputs
  - `known_ys` : selection
    array/selection of known y values
  - `known_xs` : selection
    array/selection of known x values
- outputs
  - computed intercept : numeric
    estimated **intercept** for a linear regression line for the given data.

Correlation Coefficeint (R)

Measures the correlations between x's and y's.

- inputs :
  - `known_x` : selection
    array/selection of known x's.
  - `known_y` : selection
    array/selection of known y's.
- outputs :
  - correlation : numeric
    correlation coeffecient; between -1 and 1.

Interpretting Correlation Coeffecients

R value	Qualitative description
$-1 \leq R \leq -0.7$	Very Strong Negative Correlation
$-0.7 < R \leq -0.4$	Strong Negative Correlation
$-0.40 < R \leq -0.3$	Moderate Negative relationship
$-0.30 < R \leq -0.2$	weak positive relationship
$-0.2 < R < 0$	no or negligiabel relationship
$R = 0$	No relationship
$0 < R \leq 0.2$	No or gelible relationship
$0.2 < R \leq 0.3$	weak positive correlation
$0.3 < R \leq 0.4$	moderate positive correlation
$0.4 < R \leq 0.7$	Strong positive correlation
$0.7 < R \leq 1$	Very Strong Positive Correlation.

Want to learn more for a correlation coeffecients? Check out CorrelationCoefficient - StatisticsHowTo

Which Trendline to choose?

Things to Consider:

- What kind of relationship do you expect between your datapoints?
- What are the limitations of this dataset?
    - Do you expect to collect data outside this range?
    - Is this survey of physical data?
- How will this trendline be used in your later analysis?
- How do you measure best fit and trendline performance?

Which Trendline to choose?

Common measures of trendline fit are

- Pearsons R^2.
  - is equal to squared correlation
  - can only be used with *Linear* regression lines
- Sum of Squared Residuals *SSR*.
  - Takes a sum of the *square* of the residuals (i.e., difference between the actual data and estimated function value).
  - the *smaller* the *better* the trendline fit.

Prediction

- Trendlines can be used to to predict values you don't have.
  - In Sample Predicition (interpolation) : Using the *trendline* to predict values that fall within the range of sample data that was used to create the trendline.
  - Out of Sample Predition(extrapolation) : Using the *trendline* to predict values that fall out of the range of sample data that was used to create the trendline.
- for Extrapolation, you can visualize these predictions by adding *forward* and *backward* periods to the trendline in the *Format Trendline* panel.

Prediciton

- Lets try predicting the mortality in regions by:
  - Interpolation : average annual temperatures between 30 and 50 degrees, in 1 degree increments.
  - Extrapolating : average annual temperatures between 55 and 65 degrees, in 1 degree increments.
- Save the spreadsheet for future reference.
- Follow the same format when working through the homework.
- See *'CancerStudyStolution(Complete).xlsx'*

Exercise 1: Lung Cancer Prevelance

1. Download *Cigarettes.xlsx*
2. Find
    - m : the slope of the linear regression line
    - b : the intercept of the linear regression line
    - r : the correlation coefficient.
3. Create a Scatter plot of the data and:
    - Add the linear regression line to the chart
    - Add exponential, qudratic, and cubic regression curves.
4. In the excel document answer the following questions in a text box.
    - Which is the best model?
    - What is your prediction if a region's average number of cigarettes per person is 3500? How about 4000?

Exercise : Solution

1. Create a scatter plot of Ciggarettes vs Lung Cancer deaths by:
   1. selecting Cells A2:B16.
   2. Going to "Insert" → "X Y (Scatter)".
2. Add a linear regression line by selecting the data, right-(Crtl) clicking the data points and selecting "Add Trendline"
3. In Cell B18 compute the slope of linear regression line, by writing:
   - `=SLOPE(B3:B16, A3:A16)`
4. In Cell B19 compute the y-intercept of linear regression line, by writing:
   - `=INTERCEPT(B3:B16, A3:A16)`
5. In Cell B20 compute the correlation between the x and y values by writing:
   - `=CORRELL(B3:B16, A3:A16)`

Exercise : Solution (continued)

1. Compute the linear estimates for all x's using the m and b computed earlier.
   - In Cell D3 write: `=$B$18*A3 + $B$19`
   - Use autofill to apply the formula for all x's.
2. Find the equations for the non-linear regression curves by modifying the trendline:
   1. Select the trendline and right-(Ctrl) click the trendline.and click "Format Trendline"
   2. In the "Format Trendline" menu check the box for "Display Equation on chart" and change the selection for trendline to be:
      - Linear for Linear
      - Exponential for Expon
      - Polynomial with deg=2 for Quad
      - Polynomial with deg=3 for Cubic

Exercise : Solution (continued)

1. In cells E3, F3, and G3 compute the estimates given regression curve equations found previously. In cells:
   - E3 write: `=9.1389*EXP(0.0003*A3)`
   - F3 write: `=-3*(10∧-6)*(A3∧2) + 0.0205*A3 -14.218`
   - G3 write: `=-3*(10∧-9)*(A3∧3) + 2*(10∧-5)*(A3∧2) -0.0507*A3 + 43.723`
2. Use autofill to fill in estimates for the three non-linear models.
3. Compute the residuals between the *estimates* and *observed* values by taking their difference. In cells:
   - I3 write : `=B3-D3`
   - J3 write : `=B3-E3`
   - K3 write : `=B3-F3`
   - L3 write : `=B3-G3`
4. Use autofill to compute residuals for all pairs of y observations and estimates for each model.

Exercise : Solution (continued)

1. Compute the Sum of Sqaured Squared Residuals by using SUMSQ().
   - In Cell I17 write : `=SUMSQ(I3:I16)`
   - Autofill from I17 to L17.
2. Compare the computed sum of squared residuals
   - The model with the best fit is the one with the *smallest* Sum of Squared Residuals
3. Compute out of sample Estimates using the cubic model by
   1. writing down the x values, 3500, 4000 in cells A24, and A25 respectively.
   2. Copy over the equation text from cell L3 and paste into Cell B24.
   3. Edit the equation in B24 such that it is passing in the x values from A24.
   4. Once you have an estimate in B24, use autofill to get the estimate in B25.

Exercise : Solution (continued)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Row data			Estimates					Residuals				
2	Cigarettes per person	Lung Cancer		Linear	Expon	Quad	Cubic		Linear	Expon	Quad	Cubic	
3	2860	22.07		21.08	21.55	19.87	23.65		-0.99	-0.52	-2.20	1.58	
4	2010	13.58		16.65	16.70	14.87	14.62		3.07	3.12	1.29	1.04	
5	2791	22.8		20.72	21.11	19.63	22.93		-2.08	-1.69	-3.17	0.13	
6	2618	20.3		19.82	20.04	18.89	21.03		-0.48	-0.26	-1.41	0.73	
7	2212	16.59		17.71	17.75	16.45	16.55		1.12	1.16	-0.14	-0.04	
8	2184	16.84		17.56	17.60	16.24	16.26		0.72	0.76	-0.60	-0.58	
9	2344	17.71		18.39	18.46	17.35	17.96		0.68	0.75	-0.36	0.25	
10	2692	22.04		20.20	20.49	19.23	21.85		-1.84	-1.55	-2.81	-0.19	
11	2206	14.2		17.67	17.71	16.41	16.49		3.47	3.51	2.21	2.29	
12	2914	25.02		21.36	21.91	20.04	24.19		-3.66	-3.11	-4.98	-0.83	
13	3034	25.88		21.98	22.71	20.36	25.31		-3.90	-3.17	-5.52	-0.57	
14	4240	23.03		28.26	32.61	18.77	23.04		5.23	9.58	-4.26	0.01	
15	1400	12.01		13.48	13.91	8.60	11.92		1.47	1.90	-3.41	-0.09	
16	2257	20.74		17.94	17.99	16.77	17.02		-2.80	-2.75	-3.97	-3.72	
17								SUMSQ	98.609	152.812	132.536	24.6352	
18													
19	m=	0.00520226											
20	b=	6.19762832									Cubic is the best model.		
21	r=	0.7758662	Strong Linear Relationship										
22													
23	Prediction (using the cubic model)												
24	3500	28.00											
25	4000	26.27											

Further Reading

Computer Applications for Life Sciences Chapter 2, p. 15-20