Residuals & Non-Linear Models

Dr. Bashar Al-Shboul

Residuals

- A residual value is the difference between an actual observed *y* value and the corresponding predicted *y* value, *y'.*
- Residuals are just errors.
    - Residual= error= (observed- predicted) = *(y –y')*

Example

- A least-squares regression line was fitted to the weights (in pounds) versus age (in months) of a group of many young children. The equation of the line is *y* == 16.6 + *0.65t,* where *y* is the predicted weight and tis the age of the child. A 20-month old child in this group has an actual weight of 25 pounds. What is the residual weight, in pounds, for this child?
- Y= 25
- Y'=29.6
- Resedual = 25 – 29.6 = -4.6

- The plot of the residual values against the x values can tell us a lot about our LSRL model.

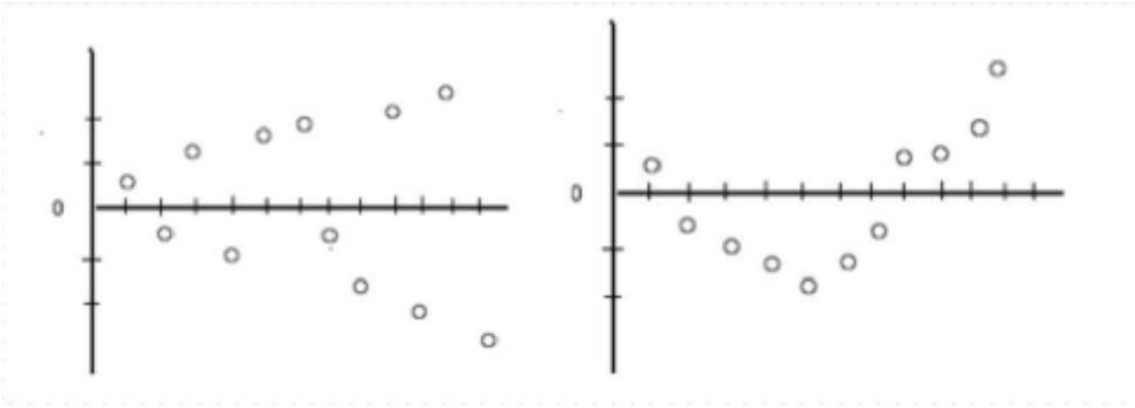
- Plots of residuals may display patterns that would give some idea about the appropriateness of the model.

- If the functional form of the regression model is incorrect, the residual plots constructed by using the model will often display a pattern.

- The pattern can then be used to propose a more appropriate model.

- When a residual plot shows no pattern, it indicates that the proposed model is a reasonable fit to a set of data.

Example Plots



- Since the residuals show how far the data falls from the LSRL, examining the values of the residuals will help us to gauge how well the LSRL describes the data. The sum of the residuals is always 0 so the plot will always be centered around the x-ax1s.

- An **outlier** is a value that is well separated from the rest of the data set. An outlier will have a large absolute residual value.

- An observation that causes the values of the slope and the intercebt in the line of best fit to be considerably different from what they would be if the observation were removed from the data set is said to be **influential**.

Example

Johnny keeps track of his best swimming times for the 50 meter freestyle from each summer swim team season. Here is his data:

Age (Years)	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Time (Secs)	34.8	34.2	32.9	29.1	28.4	22.4	25.2	24.9

Investigate the LSRL model

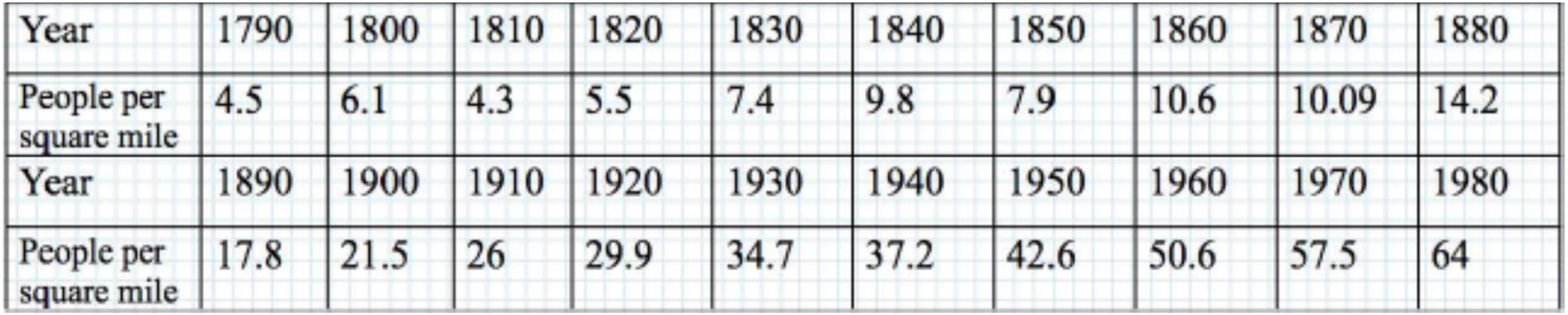
Non-Linear Models

- Many times a scatter-plot reveals a curved pattern instead of a linear pattern.

- We can transform the data by changing the scale of the measurement that was used when the data was collected.

- In order to find a good model we may need to transform our *x* value or our *y* value or both.

- Suppose this is the population per square mile in different years in the same county. Is the LSRL a good fit? If not, transform the data to find a better fit. (Steps follow)



Year	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880
People per square mile	4.5	6.1	4.3	5.5	7.4	9.8	7.9	10.6	10.09	14.2
Year	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
People per square mile	17.8	21.5	26	29.9	34.7	37.2	42.6	50.6	57.5	64

Steps:

- Draw data with the linear regression line
- Draw the data with residuals
- Modify the data by taking the natural log of y, then draw the data after correction.
- Solve for y, then draw x with the new regression non-linear model

In MATLAB

- x = 1790:10:1980
- y=[4.5, 6.1, 4.3, 5.5, 7.4, 9.8, 7.9, 10.6, 10.09, 14.2, 17.8, 21.5, 26, 29.9, 34.7, 37.2, 42.6, 50.6, 57.5, 64]
- scatter(x,y)
- Sx = std(x)
- Sy = std(y)
- r=corrcoef(x,y)
- a = r(2,1)*Sy/Sx
- b = 4.5 - a*1790

- newY = a * x + b
- hold on, plot(x, newY)
- newNewY = log(y)
- figure, scatter(x, newNewY)
- figure, scatter(x , y)
- newNewNewY = exp (a * x + b)
- hold on, plot(x, newNewNewY)