Info 1301
Linear Regression (cont.)

14 November 2016

Prof. Michael Paul

Prof. William Aspray

Least Squares Regression

- We just eyeballed the linear regression in the past.
- Now we want to give a rigorous approach for finding the line.
- Want the line to fit the data as well as possible.
- This means reducing the residuals as much as possible
- Could minimize the sum of the absolute values of the residuals.
- More commonly minimize the sum of the squares of the residuals.
  - Common, emphasizes the problems with individual large residuals, easier to calculate

Root mean square error

- RMSE = $\mathbf{[\Sigma e^2/n]^{.5}}$
- Do an example: residuals are 1,2,-1,2,-2
- RMSE = $[[(1)^2+(2)^2+(-1)^2+(2)^2+(-2)^2]/5]^{.5}$

  $= [[1+4+1+4+4]/5]^{.5}$

  $= [14/5]^{.5}$

  $= 1.67$ approx
- Generally, 68% of residuals are within 1 RSME of regression line.
- Generally , 95% of residuals are within 2 RSME of regression line.
- In this example, 2/5 within 1 RSME, 5/5 within 2 RMSE.
- Look familiar?
- In nice cases, residuals form a normal distribution around the least square line
- Can use Z-table.

Nice Conditions for Least Squared Line

- Eyeball that a line fits the data (as we have done before)
- Random residuals not too far from line (outliers can be a problem)
- The size of the residuals roughly constant (not those that get larger at one end)
- No repeating patterns in the data (e.g. time series)[Look at examples on p. 342 (Fig. 7.13)

Formulas for the line y=mx+b for linear regression

- **b** = $[(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)]/[n(\Sigma x^2) - (\Sigma x)^2]$

- **m** = $[n(\Sigma xy) - (\Sigma x)(\Sigma y)]/[n(\Sigma x^2) - (\Sigma x)^2]$

- **n = number of points**

Diabetes Example – Relationship of Age to Glucose Level

- Frederic Grant Banting (1891 – 1941) born on this day – Nobel Prize for treating diabetes with insulin
- Diabetes is one of the most common diseases – 8% of world's population
- Serious – doubles your risk of early death; damages eyes, kidneys, blood vessels, etc.
- High blood sugar either because pancreas does not create enough insulin to control glucose (type 1) or insulin resistance where cells do not respond properly to insulin (type 2)
- Risk of developing diabetes increases with age
- Observe the data set of 6 points:

(43,99), (21,65), (25,79), (42,75), (57,87), (59,81)

Plot the points.

Sketch in the line by sight.

Eyeball the correlation.

Diabetes Example (cont.)

subject	Age (x)	Glucose (y)	xy	xx	yy
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Plugging in to the formulas

- $b = [(486)(11409) – (247)(20485)]/[(6)(11409) – (247)^2]$
  $= 484947/7445$
  $= 65.14$

$m = [(6)(20485) – (247)(486)]/[(6)(11409) – (247)^2]$
$= 2868/7445$
$= .385225$

$Y = .385225x + 65.14$

What is the meaning of this equation?

Homework Problem

- Use Minitab Express to calculate the line and the root mean square error for the diabetes example that we just did

The word ‘regression’

- Root is ‘regress’
- 1550s – return to a former state – from Latin *regressus*
- 1823 – to move backward
- 1926 – to return to an earlier (and usually worse or less developed) state of life
- Statistics – move away from the random variation in a sample to its primitive state
- Let’s consider a different place where ’regression’ appears in statistical discussion
  - Use a baseball example