CSCI_B 565 DATA MINING
Homework Number 1
Morning Class Computer Science Core
Spring
Indiana University,
Bloomington, IN

Puneet Loya
ploya

Jan 21,2015

All the work here is solemnly mine

Answer for Question 1

δ	l
1.0	-1.93
1.4	-1.135
2.0	-0.681
2.2	-0.341
2.8	0.401
2.8	0.501
3.2	0.690
3.8	1.310
4.4	18.99
4.8	2.499

Part 1

As $\delta \rightarrow l$ *is linear (from the question)* $l = \beta_0 + \beta_1 \delta$
Solving β_0, β_1 *by the least square method,*

$$n\beta_0 + \beta_1 \Sigma_i \delta_i - \Sigma_i l_i = 0$$
$$\beta_0 \Sigma_i \delta_i + \beta_1 \Sigma_i {\delta_i}^2 - \Sigma_i \delta_i l_i = 0$$

From the above table,
$\Sigma_i \delta_i = 28.4$ $\Sigma_i l_i = 20.304$ $\Sigma_i \delta_i l_i = 99.63$ $\Sigma_i {\delta_i}^2 = 94.56$
So the equations looks like:

$$10\beta_0 + 28.4\beta_1 - 20.304 = 0$$
$$28.4\beta_0 + 94.56\beta_1 99.63 = 0$$

Solving these simultaneous equations,:
$\beta_0 = -6.54\beta_1 = 3.0183$
Taking β_1 *as approximately* 3.02

δ	Expected(l)	Observed(l)	Difference
1.0	-1.93	-3.52	1.59
1.4	-1.135	-2.312	1.77
2.0	-0.681	-0.5	0.181
2.2	-0.341	0.104	0.445
2.8	0.401	1.916	1.515
2.8	0.501	1.916	1.415
3.2	0.690	3.124	2.434
3.8	1.310	4.936	3.626
4.4	18.99	6.748	12.242
4.8	2.499	7.956	5.457

So a difference of 12.242 for $\delta = 4.4$ *seems to be an outlier.*
There can be two options:
1) Remove the outlier.
2) Predict the correct value by observation.
For the 2nd option we can take mean of the previous and the next value. As it is an increasing pattern with no regular differences a guess would be mean of the predecessor and successor.
By taking mean we get the corrected value as ***1.901****.*

Part 2

For the cleaned data, predicting the intervals was done by using matrix multiplication.

$$log_{10}l = \beta_0 + \beta_1\delta$$

The above equation is a linear equation. The brute force approach used to get a range for the β_0&β_1 is taking two data points of the cleaned data and solving for $\beta_0 and \beta_1$ by matrix multiplication.
Example:
$y_1 = \beta_0 + \beta_1 x_1$
$y_2 = \beta_0 + \beta_1 x_2$

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}$$

$$X = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$B = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix}$$

AX = B
$X = A^{-1}B$
The intervals are set for both β_0 *and* β_1 as $(\infty, -\infty)$ initially.
As from the above example, the inverse of the matrix is calculated and multiplied to the other matrix i.e, the matrix with the dependent variables. The output of this matrix multiplication will give the tenative β_0 *and* β_1. On getting this result, the above two intervals are to be modified. The modification is done as follows:

$$rangeMin\beta_0 = min(current\beta_0, previous\beta_0)$$
$$rangeMax\beta_0 = max(current\beta_0, previous\beta_0)$$

This (rangeMin,rangeMax) is first calculated for first two data points and then this window keeps getting braoder with new data point coming in for both β_0 *and* β_1. Every calculation of β_0 *and* β_1 the weighted mean of regressor variables and response varaibles respectively and use them as one of the data points for the next calculation. This calculation of mean was a heuristic used to get the resultant approximate range of β_0 *and* β_1. The java code is attached with the assignment.
The final values are $\beta_0 = -1.19$ *and* $\beta_1 = 0.33$
The Running Complexity will be O(n^4). With increasing number of parameters, the matrix dimension is all that will increase, the complexity will remain the same i.e,O(n^4). Limitation of the program is it does not accept negative values for l.

Part 3

The lm function for the cleaned data returns the coefficients $\beta_0 = -1.27$ *and* $\beta_1 = 0.35$. Even in R, no negative values were considered.

Part 4

The results from the heuristics are nearly equal to the values given by the lm function of R. From the heuristics: $\beta_0 = -1.19$ *and* $\beta_1 = 0.33$
From the lm function $\beta_0 = -1.27$ *and* $\beta_1 = 0.35$

Answer for Question 2

Data Mining: The process of knowledge discovery in data i.e, finding patterns within data by transforming it to run algorithms on it and validate the findings. These findings may turn out to be useful for the business to make informed decisions. The finding of patterns may involve application of Artificial intelligence, Machine learning and Statistics.

Analytics: Data analytics is a way to get inferences from the data that is already known. There may not be any transformation of data involved.

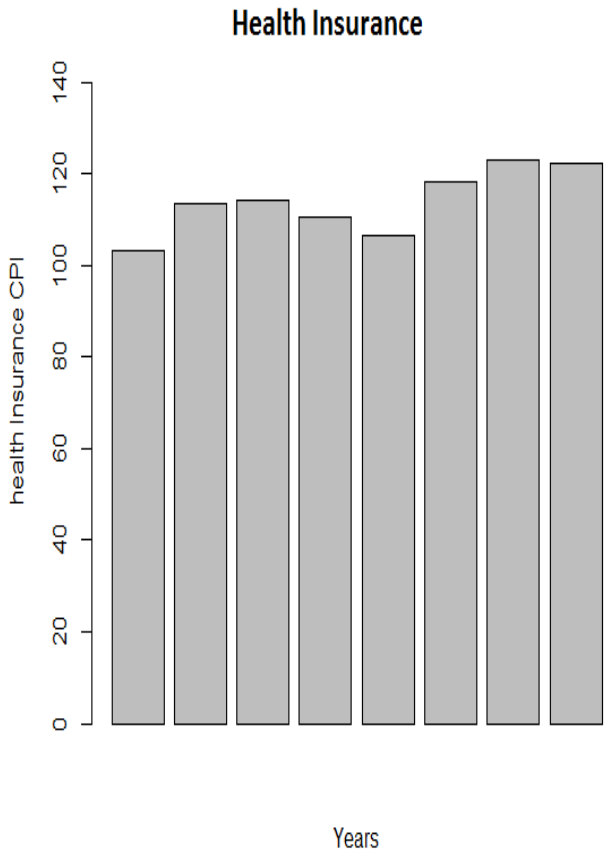
Statistics: A branch of Mathematics which deals with interpretation of data collection, its organization and presentation.

Machine Learning: The study of constructing and improving computer algorithms automatically by learning from heuristics.

Database: It is a data store which can organize data in a systematic model and supports querying of the stored information.

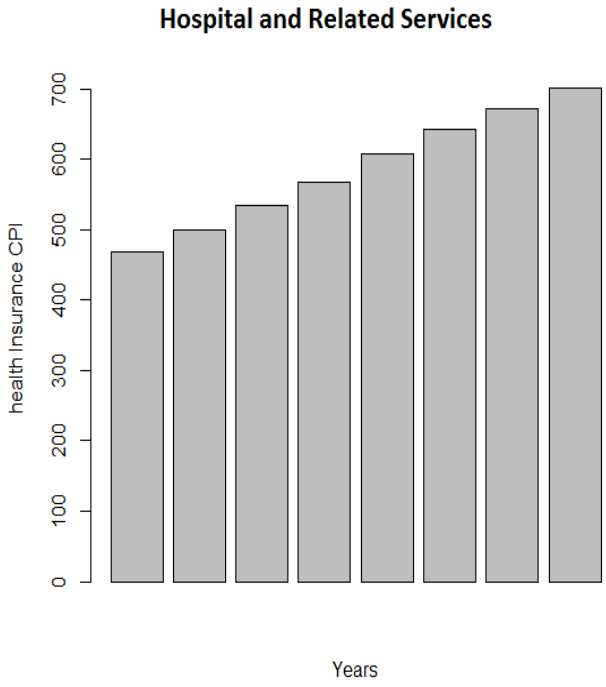
Answer for Question 3

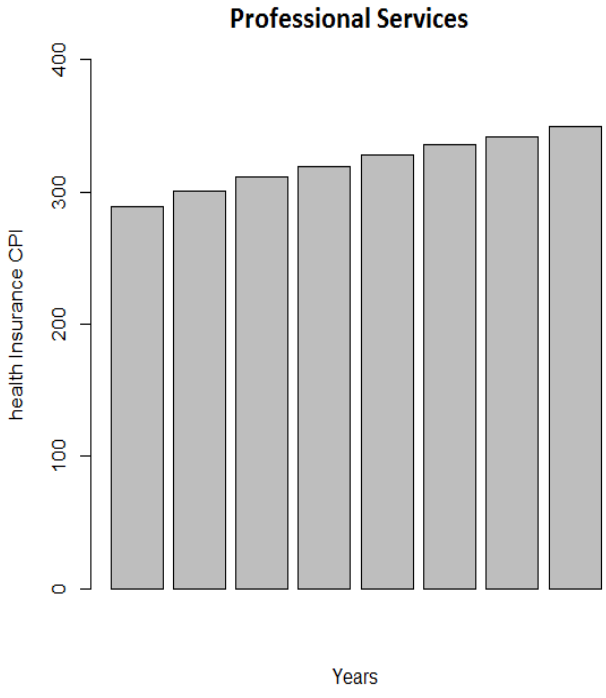
I looked into the medical data for all US cities average (Area Code :0000) of the survey. The medical care services include: health insurance, professional services and hospital and related services. The data compared is for years 2006 - 2011. The pattern of data for health insurance has decreased continuously from 2008 to 2011. The decrease during this period amounts to about 7.8%(114 105). Later increased rigorously from 2011 to 2012(105 - 118) i.e in a year. My prediction would be a weak economy, people did not insure themselves. Later after 2011, things look back on track.



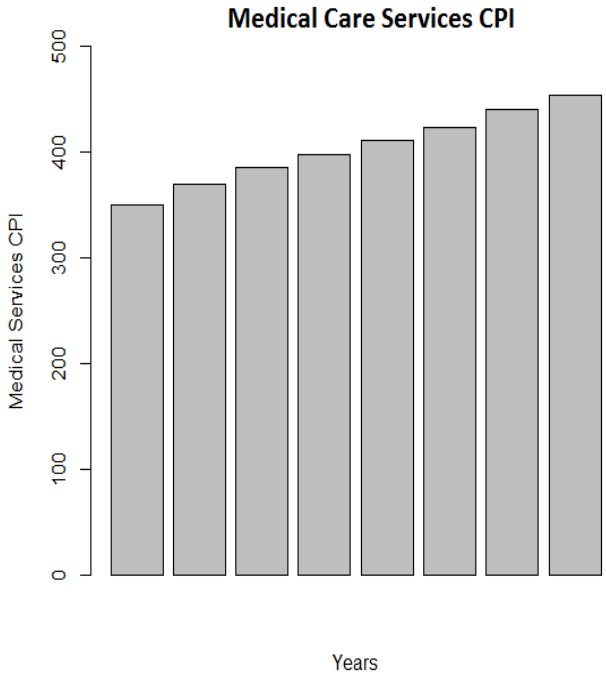
Year	CPI
2006	103.1
2007	113.52
2008	114.221
2009	110.527
2010	106.627
2011	118.28
2012	122.973
2013	122.108

But the other two components in Medical Care Services, professional services and hospital and related services do not show any decrease in consumer price index even during weaker economy.





Even the overall Medical Care Services consumer price index has never decreased.



References

http://en.wikipedia.org/wiki/United_States_Consumer_Price_Index
To get a background of Medical Care Services: http://www.bls.gov/cpi/cpifact4.htm

Credits

I discussed the approach for the 2nd part of the first question in the assignment with a class mate : Suhas Gulur Ramakrishna