ASSIGNMENT 5
DUE DATE: Saturday 10 November, 2012 11:59 PM

In this assignment you will implement the synthetic residual generation algorithm discussed in Lecture 11 (slide 11 onwards). You will also assess if trends between synthetic residuals and synthetic data values are similar to that of residuals and data values in the true data.

The assignment will contribute 9 points to your final grade, but will be graded out of 90 points. You may use one of R or Matlab to implement the algorithm.

Steps

For a given data set of size n, you should perform the following steps:

1. Determine the coefficients of a linear regression model on the data set – \( w = (A^T A)^{-1} A^T y \). This can be achieved in R by using the solve function: \( \text{solve}(t(A)%*%A,t(A)%*%y) \).
2. Obtain \( y_{\text{pred}} = Aw \) and \( \text{residuals}_{\text{true}} = y - y_{\text{pred}} \).
3. For each independent variable \( x_i \) and the dependent variable \( y \), obtain the pdf using the normal kernel density estimation method
   a. divide the range of the variable into, say, 100000 equally spaced points between the minimum and maximum values
   b. obtain the density value (using the kernel method) at each of those points
4. Use the pdfs to generate a synthetic data set
   a. for each variable, sample n values using the pdf for that variable: you can use the inverse-cdf method to do this sampling (see http://en.wikipedia.org/wiki/Inverse_transform_sampling)
5. For each variable, use the method discussed in class to obtain synthetic residuals for the synthetic data values
   a. this step requires fitting a smooth curve to the true value vs. true residual plot (an example is given later)

Data sets

You should create different data sets with different number of variables, with linear and non-linear dependencies. Create at least five such data sets. Example construction:

- two independent variables \( x_1 \) and \( x_2 \)
- \( x_1 \) is uniformly distributed between 0 and 1
- \( x_2 \) is normally distributed with zero mean and unit variance
- \( y = x_1 + x_2^2 \)

The data value vs. residual plot (fitting done using linear regression) for the data set generated using this construction is as follows:
As can be seen, the residual plot indicates that there is a non-linear relationship between y and x_2.

**Fitting curve to data value vs. residual plot**

In order to generate synthetic residuals corresponding to a variable x, you will first have to fit a smooth curve to the x vs. residual (true) plot. The steps to do this in R are shown below. If you are implementing in Matlab, then you will have to find the corresponding functions in Matlab.

```r
library("mgcv")

# create a data frame of x and the residuals
d <- data.frame(x=x,residual=residual)

# run gam; if you want to know the curve-fitted values of x then its in # fitted(x)
g <- gam(y~s(x),data=d)

# some points on which we want to obtain value from fitted curve
s <- data.frame(x=c(30,70))
# this is how you obtain values from the fitted curve
res.oncurve <- predict(g,newdata=s)
```

The following plot shows some hypothetical x and residual values, along with the fitted curve. It also shows the values obtained for x=30 and x=70 on the curve.

![Residual plot](image)

**Assessment**

For each variable (including the dependent variable y) in each data set, create a variable value vs. true residual and a synthetic value vs. synthetic residual plot. Compare the plots to see if similar conclusions can be made from the synthetic data plot. Including the plots and your assessment in a separate document. Also, include in this write up, a brief description of the implemented algorithm and a description of the data sets.
Grading and submission

The 90 points will be distributed as follows: correct implementation – 40; comprehensive experiments – 20; discussion – 20; source code documentation – 10. Submit the source code and the write-up in blackboard by the stated deadline.

You must work alone in this assignment. It is okay to discuss the algorithm with your peers, but the programming must be done individually.