Overview of SAS Project:

+Rising Health Care costs in US prompts need for change in Health Care Policy

+Want to assess healthcare cost associated with having diabetes and analyze ways to reduce incidence of diabetes as a means of reducing healthcare costs

+Need to determine ideal target population of individuals who are at risk for diabetes and enact measures that reduce risk of diabetes below a critical point

+Proposed measure: reduce BMI in a target population (those individuals with BMI >=25)

+Provided with 43 parameters for 50,000+ patients, including whether or not the patient has diabetes

Overall Plan:

+Reduce the number of parameters to a core group of parameters that are the most important in determining whether or not an individual will have diabetes or not

1st, remove those parameters that may be caused by diabetes, as these are not useful in predicting who will get diabetes in the future

2nd, perform Partial Least Squares analysis on the remaining parameters to determine the most important parameters, i.e., those most associated with persons having diabetes

+PLS is an extension of the multiple linear regression model, using the dominant eigenvector corresponding to matrices that are the product of some factor matrix and some response matrix

For each *h*=1,…,*c*, where ***A0*=*X'Y*, *M0*=*X'X*, *C0*=*I***, and *c* given,

1. 1) compute ***qh***, the dominant eigenvector of ***Ah'Ah***
2. 2) ***wh=Ahqh***, ***ch=wh'Mhwh***, ***wh=wh/sqrt(ch***), and store ***wh*** into ***W*** as a column
3. 3) ***ph=Mhwh***, and store ***ph*** into ***P*** as a column
4. 4) ***qh=Ah'wh***, and store ***qh*** into ***Q*** as a column
5. 5) ***vh=Chph***, and ***vh=vh/||vh||***
6. 6) ***Ch+1=Ch - vhvh'*** and ***Mh+1=Mh - phph'***

7) ***Ah+1=ChAh***

The factor scores matrix ***T*** is then computed as ***T*=*XW*** and ***B*** for the regression of ***Y*** on ***X*** is computed as ***B*=*WQ'***.

\*from http://www.statsoft.com/textbook/partial-least-squares/