STA 5207 Assignment 8

Due Friday November 5

Consider the data set Assignment.txt in Canvas->Files->Data Sets (and SAS Studio). There is a response variable and 3 predictors (no column headings):

*y*: number of worker hours over a month

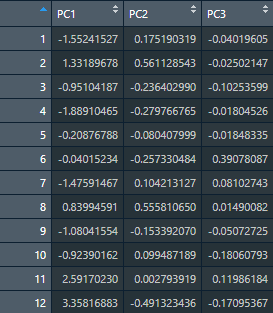
*x*1: number of cases handled

*x*2: eligible population for the hospital

*x*3: number of operating rooms

1. (10 points) Perform multicollinearity diagnostics. Does it appear that multicollinearity is a problem? (no output needed)
   1. Yes it appears there is a problem with multicollinearity
2. (5 points) Standardize the predictors. (no code or output need to be submitted)
3. (5 points) What is the maximum number of principal components that can be created?
   1. 3 principal components are the maximum amount because 3 predictors
4. (10 points) Which principal component accounts for the most variation in the predictors? What proportion? (no output needed)
   1. PC1 accounts for the most variation in the predictors at 95.76%
5. (10 points) Based on the cumulative proportion of variation, how many principal components do you believe are adequate? Why? (no output needed)
   1. I believe only the first 2 PCs are adequate because they account for 99.237% of the variation, so the 3rd component accounts for hardly any variation
6. (10 points) Obtain the principal directions (submit R output)
   1. Rotation (n x k) = (4 x 4):

|  |  |  |  |
| --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 |
| x1 | 0.5833505 | -0.2991168 | -0.7551366 |
| x2 | 0.5794513 | -0.4982279 | 0.6449847 |
| x3 | 0.5691559 | 0.8138170 | 0.1173178 |

1. (10 points) Using the principal directions and the predictor values from the data, obtain the first three elements of Principal component 1 by hand.
   1. 0.5833505(-.9276511) + 0.5794513(-1.01275953) + 0.5691559(-.7457092) = -1.55241
   2. 0.5833505(.6280143) + 0.5794513(0.47606092) + 0.5691559(1.2117774) = 1.331897
   3. 0.5833505(-.4066499) + 0.5794513(-.49943399) + 0.5691559(-.7457092) = -0.95104
2. (10 points) Obtain all the principal components from R (submit R output)
   1. 
3. (15 points) Perform principal component regression using the components you chose in 5. Give the fitted equation and R2. No output needed.
   1. y = 1582.35 + 686.01x1 + 78.02x2
   2. R2 = 0.9857
4. (15 points) Perform principal component regression using all the components. Give the fitted equation and R2. Compare the coefficient estimates and R2 with the model from question 9.
   1. y = 1582.35 + 686.01x1 + 78.02x2 – 59.37x3
   2. R2 = 0.9858
   3. The coefficient estimates are the same for PC1 and PC2 because the PCs are orthogonal, but the R2 when using all PCs is 0.0001 higher, so it appears that using the first 2 principal components is sufficient.

No R code needed.