The Analytics Edge Summer 2024

Test Your Knowledge of PCA, SVD and Censoring

1. In this exercise, you will study a technique called latent semantic indexing, which applies singular value decomposition to create a low dimensional representation of data that is designed to capture semantic similarity of words. A list of all 460 unique words/terms that occurs in a set of 9 documents is provided in `lsiWords.txt`. A document by term matrix is in `lsiMatrix.txt`.

   (a) Use the `read.table()` function to read in the data. Create a matrix $\boldsymbol{X}$ that is the transpose of the lsiMatrix, so that each column represents a document. Compute the singular value decomposition of $\boldsymbol{X}$ and make an approximation to it using the first two singular values and vectors. Plot the low dimensional representation of the 9 documents in two dimensions by using the right singular vectors. Which two documents appear to be closest to each other in the low dimensional representation?

   (b) Consider finding documents that are about alien abductions. There are three versions of this word in the data - term 23 ("abducted"), term 24 ("abduction"), term 25 ("abductions"). Suppose you want to find documents containing the word "abducted"; now documents 2 and 3 contain it, but, document 1 does not. However document 1 is clearly related to the topic. Thus LSI should also find document 1. Create a test document q containing the one word "abducted" and project it into the 2D subspace to make $\hat{q}$. Now compute the cosine similarity between $\hat{q}$ and the low dimensional representation of all the documents. What are the top 3 closest matches?

2. This question involves the use of principal component analysis on the well-known `iris` dataset. The dataset is available in R.

   (a) How many observations are there in the dataset? What are the different fields/attributes in the data set?

   (b) Create a new dataset `iris_data` by removing the `Species` column and store its content as `iris_sp`.

   (c) Compare the various pair of features using a pairwise scatterplot and find correlation coefficients between the features. Which features seem to be highly correlated?

   (d) Conduct a principal component analysis on `iris_data` without standardizing the data. You may use `prcomp(..., scale=F)`.

      (i) How many principal components are required to explain at least 90 % of the variability in the data? Plot the cumulative percentage of variance explained by the principal components to answer this question.

(ii) Plot the data along the first two principal components and color the different species separately. Does the first principal component create enough separation among the different species? To plot, you may use the function `fviz_pca_ind` and `fviz_pca_biplot` in textttlibrary(factoextra). You can play around with the `habillage` option in the command and also try to create confidence bounds on the ellipses around the groups to see how much they are separated.

(e) Do the same exercise as in (d) above, now after standardizing the dataset. Comment on any differences you observe.

3. The data in the file `mroz.csv` includes data on hours worked for 753 married women. This dataset is taken from the paper by Mroz, T.A. (1987) titled *The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, Econometrica, 55, 765-799.* The variables in the dataset include:

- `hours`: Hours worked
- `kidslt6`: Number of children less than six years old
- `kidsge6:` Number of kids between 6 and 18 years of age
- `age`: Age
- `educ`: Years of education
- `exper`: Past years of labor market experience
- `nwifeinc`: Husband's earnings, measured in thousands of dollars

(a) Quadratic functions are often used in labor economics to capture increasing or decreasing marginal rates. Start by defining a new variables `exper2` which is defined as the square of the `exper` variable. Run a linear regression on the hours worked using all the variable including `exper2` and report on the R-squared and adjusted R-squared. Write down the fitted equation. Does the result indicate that the `exper2` variable is significant at the 5% level? Does the results indicate that experience has an increasing or diminishing marginal effect on wage?

(b) What is the range of fitted values for hours from the result in part (a)? How many of the fitted values are below 0? How many observations in the dataset have hours = 0? How do these two numbers compare?

(c) Estimate a Tobit model to predict hours using all the variables including `exper2`. Compare the signs of the coefficients with the results in part (a) and check if the results match?

(d) We will now compare the R-squared values from the results in part (a) and (c). Remember that for the linear regression model, the R-squared value is the squared correlation between the fitted values of hours and the actual value. We now define the R-squared for a Tobit model in a similar manner. Given the estimated values of β and the scale

parameter σ, the predicted mean value of the dependent variable is given as:

$$\mathbb{E}(y_i|x_i) = \beta' x_i \Phi(\beta' x_i/\sigma) + \sigma\phi(\beta' x_i/\sigma).$$

Compute this value for all observations (remember that the `predict` function in `survreg` only returns the $\beta' x_i$ values). Now define the R-squared value by computing the correlation of these predicted value with the actual value of the hours. Compare the R-squared values of the two models and comment on which is preferred.

(e) We now compare the estimates from the two models as a function of education. Assume that all the variables other than `educ` are set at their mean values. Write down the linear equation that describes the average hours worked as a function of the education level from the linear regression model. What is the estimated value at 8 and 12 years of education from the linear regression model? Compare this with the estimated values from the Tobit model. Is there increasing or decreasing marginal effect of education on the hours worked in the Tobit model?