PRINCIPAL COMPONENT ANALYSIS

By:
Dr. Akash Asthana
Assistant Professor,
Dept. of Statistics,
University of Lucknow, Lucknow

PRINCIPAL COMPONENT ANALYSIS

- It is a dimension reduction technique.
- In some situations the measurements are taken over a large number of variables.
- But it is not possible to deal with a large number of variables.
- Therefore instead of these large number of variables their linear combinations, which are linearly independent and orthonormal also, are used which can explain maximum possible variation in the data.
- These linear combinations are called as principal components.

PRINCIPAL COMPONENT ANALYSIS

- Transforming the original vector variable to the vector of principal components amounts to a rotation of coordinate axes to a new coordinate system that has inherent statistical properties.
- The set of principal components yields a convenient set of coordinates, and the accompanying variances of the components characterize their statistical properties.
- The method of principal components is used to find the linear combinations with large variance.

PRINCIPAL COMPONENT ANALYSIS

- Let X be a random vector of p variables having variance-covariance matrix Σ. Without loss of generality the mean vector of X can be taken as 0.
- The main objective of Principal Component Analysis is to obtain the linear combinations of X vector in a manner that the variance of the combination is maximum.
- Let the linear combination of X is $\beta' X$.
- Then $V(\beta' X) = \beta' \Sigma \beta$. (1)
- As these linear combinations are orthonormal we will maximize this variance under the condition $\beta' \beta = 1$.

PRINCIPAL COMPONENT ANALYSIS

- For the purpose we define the function:
  $\varphi = \beta'\Sigma\beta - \lambda(\beta'\beta - 1)$
  Where λ is Lagrange's multiplier (a scalar quantity).
- For maximum variance: $\frac{\partial \varphi}{\partial \beta'} = 0$
- It will give the equation: $\Sigma\beta - \lambda\beta = 0$
  or $(\Sigma - \lambda I)\beta = 0$ *(2)*
- This equation will have a solution if $(\Sigma - \lambda I)$ will be singular i.e. $|\Sigma - \lambda I| = 0$. *(3)*
- In other words we can say that λ is the characteristic root of Σ and β be the characteristic vector of Σ.

PRINCIPAL COMPONENT ANALYSIS

- Also from equation (1), we get: $\beta'\Sigma\beta' = \lambda\beta'\beta = \lambda$.
- Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the characteristic vectors of the matrix Σ.
- As the Linear combination must have maximum variance we take $\lambda = \lambda_1$ and $\beta^{(1)}$ be the characteristic vector associated with it.
- Therefore first principal component for X matrix is given by $U_1 = \beta^{(1)\prime}X$ where $\beta^{(1)}$ is the characteristic vector associated with the maximum characteristic root λ_1 of Σ.
- Let us consider another principal component $\beta'X$ of X which have maximum variance (lower than U_1) and is uncorrelated with $\beta^{(1)\prime}X$.
- Then we would have:

  $V(\beta'X) = \beta'\Sigma\beta;\ Cov(\beta'X, \beta^{(1)\prime}X) = \beta'\Sigma\beta^{(1)} = 0;\ \beta'\beta = 1 \quad (4)$
- Now for obtaining the second principal component we maximize $\beta'\Sigma\beta$ under the conditions $\beta'\beta = 1$ and $\beta'\Sigma\beta^{(1)} = 0$.

PRINCIPAL COMPONENT ANALYSIS

- Now we define a function:

  $\varphi_1 = \beta'\Sigma\beta - \lambda(\beta'\beta - 1) - 2v\beta'\Sigma\beta^{(1)}$

  Where λ and ν are the Lagrange's multipliers (Scalars).
- On maximizing the φ with respect to β' we get:

  $2\Sigma\beta - 2\lambda\beta - 2v\Sigma\beta^{(1)} = 0 \qquad (5)$
- Pre-multiplying it by $\beta^{(1)'}$ we get

  $2\beta^{(1)'}\Sigma\beta - 2\lambda\beta^{(1)'}\beta - 2v\beta^{(1)'}\Sigma\beta^{(1)} = 0$

  or $-2v\lambda = 0$
- As λ can't be we get ν = 0. putting it in (5) we get $(\Sigma - \lambda I)\beta = 0$.
- Which again show that β is the characteristic vector of matrix Σ and λ is its Eigen root.
- We can define this principal component as $U_2 = \beta^{(2)'}X$, where $\beta^{(2)}$ is the solution of equation (2) for the $\lambda = \lambda_2$.

PRINCIPAL COMPONENT ANALYSIS

- Proceeding in same manner at $(k+1)^{th}$ stage we get following conditions:
  $V(\beta' X)=\beta'\Sigma\beta$ (6)
  $\beta'\beta = 1$ (7)
  $Cov(\beta' X, \beta^{(i)} X) = \beta'\Sigma\beta^{(i)} \quad \forall i = 1,2,\dots,k$ (8)
- Proceeding in same manner and using Lagrange's multipier we can define the function $\varphi_{(k+1)}$ as:

$$\varphi_{(k+1)} = \beta'\Sigma\beta - \lambda(\beta'\beta - 1) - 2\sum_{i=1}^{k} v_i \beta'\Sigma\beta^{(i)} \qquad (9)$$

- On maximizing the equation (9) with respect to β' we again get that β is the characteristic vector of the matrix Σ corresponding to $(k+1)^{th}$ largest characteristic root $(\lambda_{(k+1)})$of it.
- Also the variance of the $(k+1)^{th}$ principal component is $\lambda_{(k+1)}$.
- $V(\beta^{(k+1)\prime} X) = \beta^{(k+1)\prime}\Sigma\, \beta^{(k+1)} = \lambda_{(k+1)}\beta^{(k+1)\prime}\, \beta^{(k+1)} = \lambda_{(k+1)}$
  by using equations (1) and (2).

PRINCIPAL COMPONENT ANALYSIS

- In matrix form we can define:

$\boldsymbol{\beta} = (\beta^{(1)}\ \beta^{(2)}\ \ldots\ \beta^{(p)})'$ and $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}$

- Then we can define principal components as:

  " Let **X** be a p-component random vector having mean 0 and variance-covariance matrix Σ. Then there exist an orthogonal linear transformation $\boldsymbol{U}=\boldsymbol{\beta}'\boldsymbol{X}$ such that the covariance matrix of U is Λ, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda p$ are the roots of equation (3). The k^{th} column of $\boldsymbol{\beta}$, $\beta^{(k)}$ satisfies the equation (2). The kth component of U, $U_k= \beta^{(k)}X$ has maximum variance of all normalized linear combinations uncorrelated with U_1, U_2,…, $U_{(k-1)}$."

PRINCIPAL COMPONENT ANALYSIS

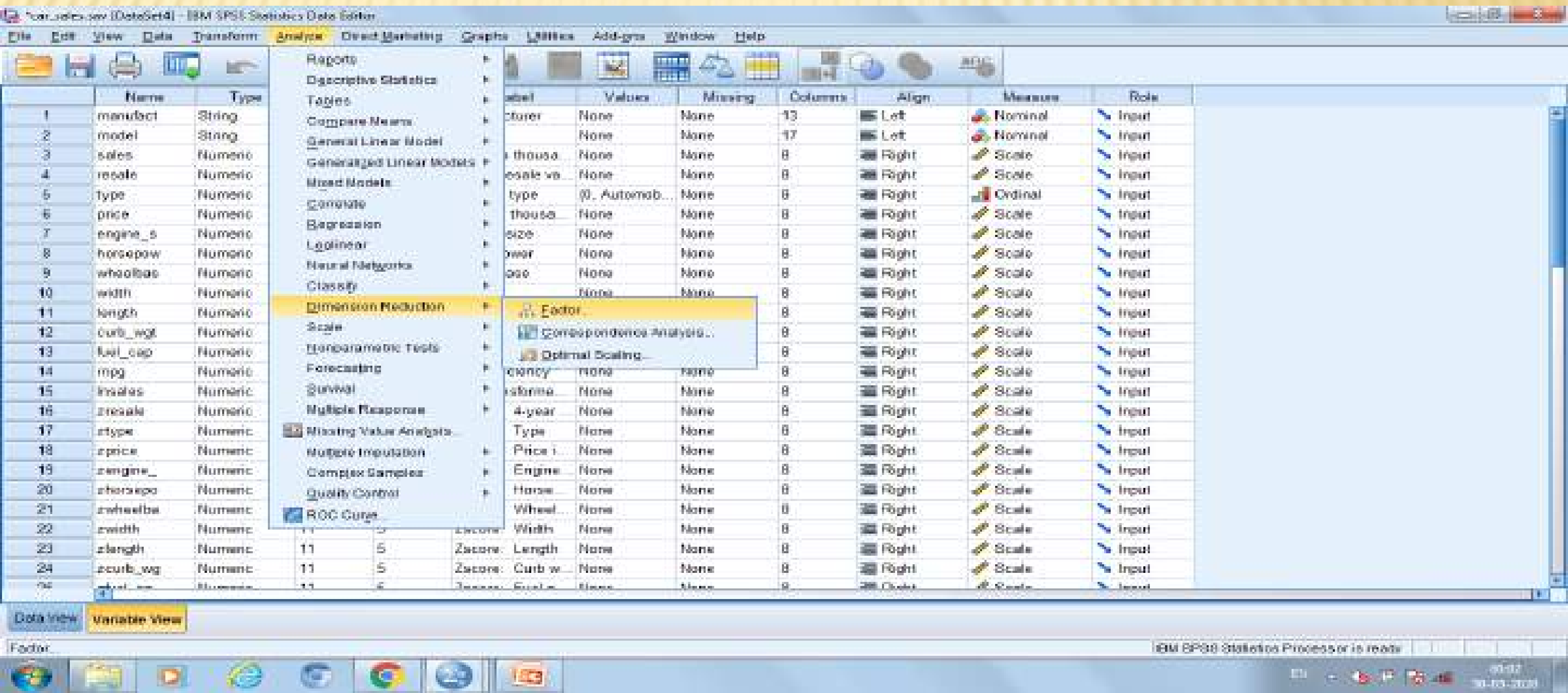
- In general situation the variance-covariance matrix Σ is unknown. Therefore for obtaining the principal components its MLE is used and thus obtained principal components are called as MLE of principal components.
- As we know that for a square matrix of order m there will be at most m characteristic roots. Therefore for a p component matrix X one can obtain at most p-principal components.

STEPS FOR PRINCIPAL COMPONENT ANALYSIS:

1. First transform the matrix of all variables under consideration to a matrix X such that mean of X will be 0.
2. Obtain the Variance-covariance matrix of X, Σ (or its MLE) under the assumption that X is Normally Distributed.
3. Obtain the Characteristic roots of Σ and arrange them in descending order ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$).
4. For each distinct Eigen root obtain Eigen vector.
5. Normalize these Eigen vectors dividing these by their norms ($\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(p)}$).
6. Then obtain the principal components by multiplying these β_i's with X (i.e. $\beta^{(1)}X, \beta^{(2)}X, \ldots, \beta^{(p)}X$)
7. In the situation if the unit of measurements for variables are not same it is better to use correlation matix in place of variance-covariance matrix.

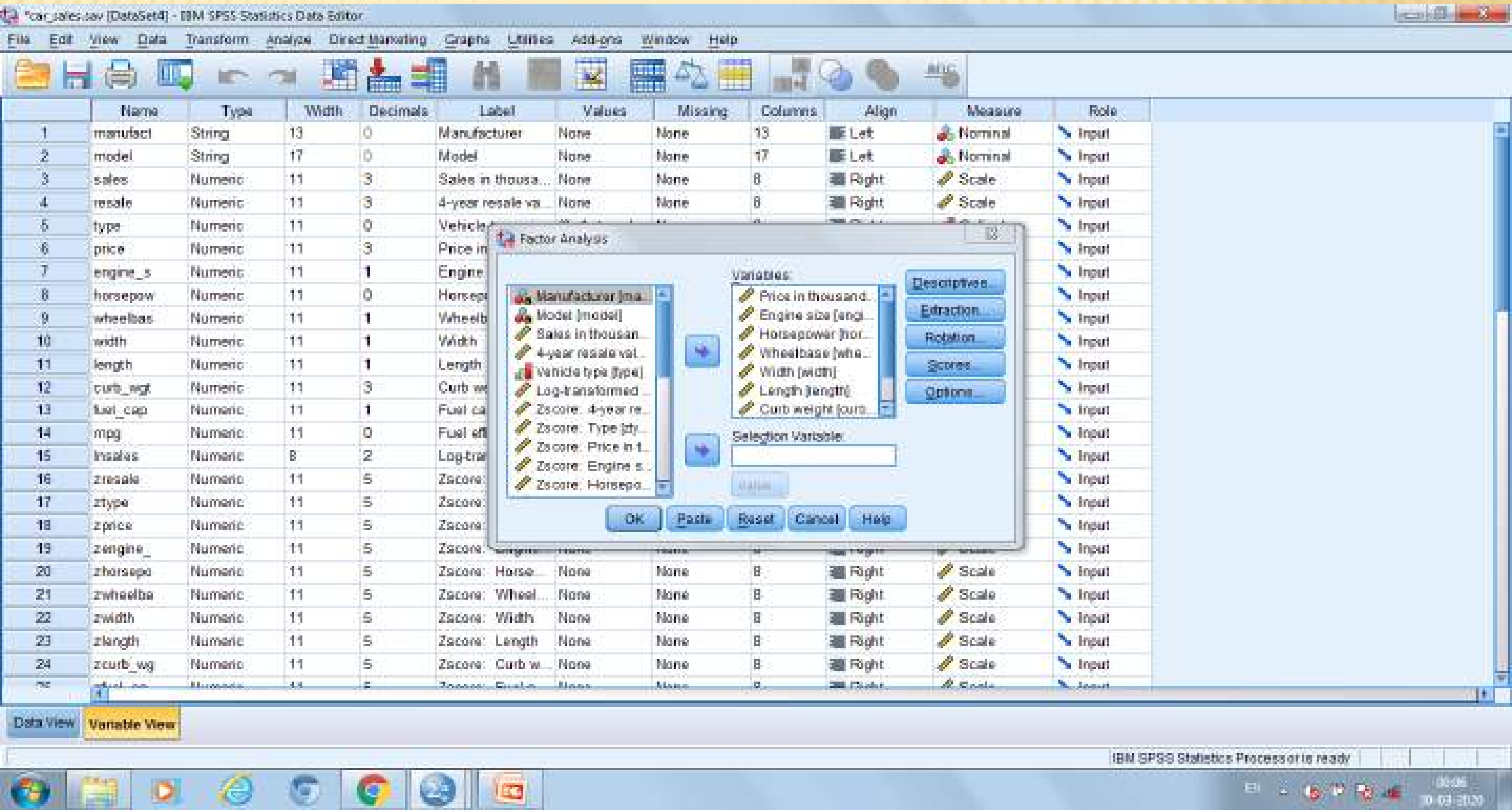
PRINCIPAL COMPONENT ANALYSIS: AN EXAMPLE USING SPSS

- For performing Principal Component Analysis (PCA) using SPSS Following steps are used.
- Click on Analyze → Dimension Reduction → Factor



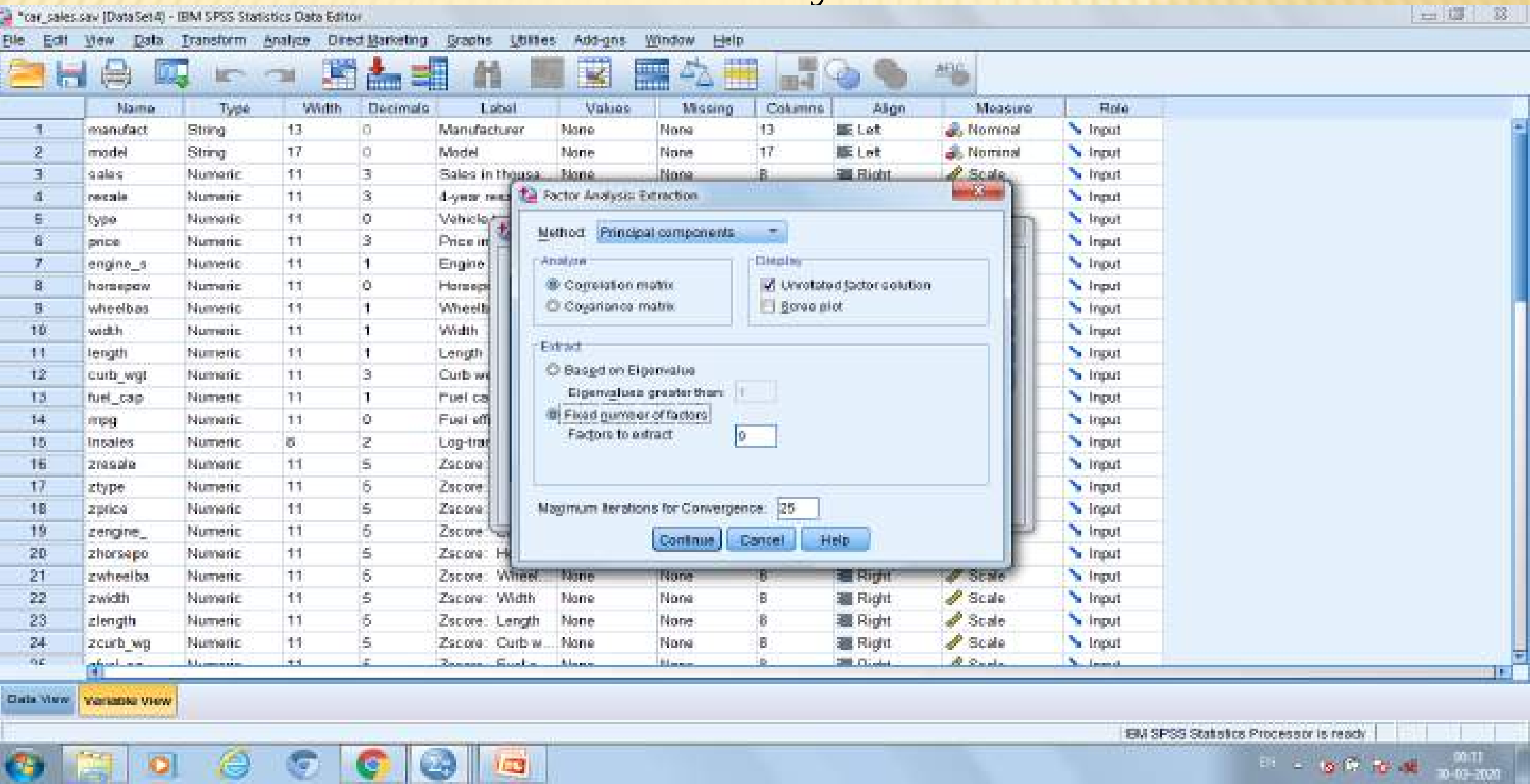
EXAMPLE (CONTD.)

- It will open the factor analysis window put all the variables required for PCA in variable box. Then click on Extraction.



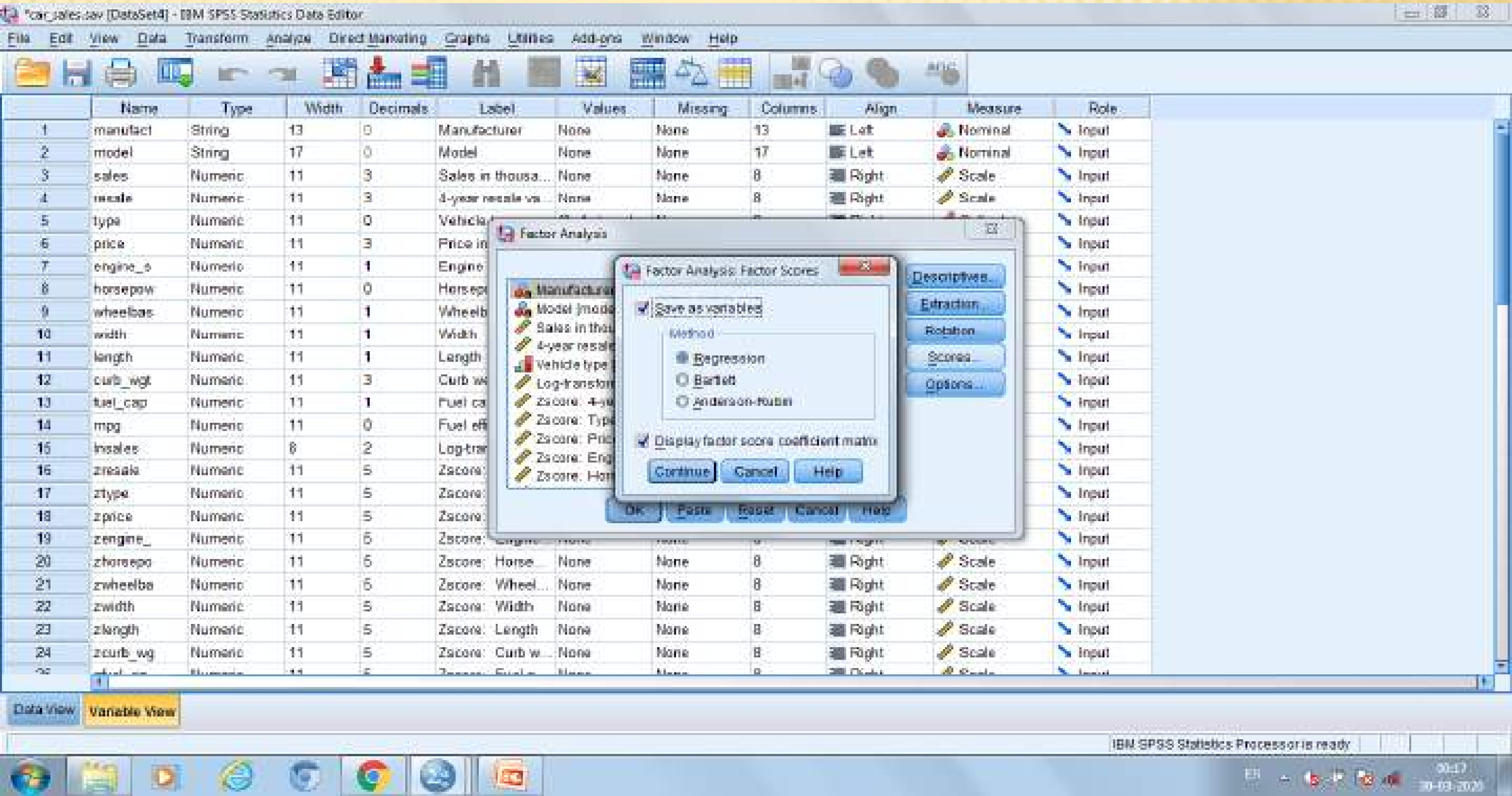
EXAMPLE (CONTD.)

- On clicking Extraction window will be open. Click on Correlation matrix and then fixed number of factors put the number of variables in the analysis in the box shown.



EXAMPLE (CONTD.)

- Click on continue then click on Scores → Save as variable → Display factor score coefficient matrix.



EXAMPLE (CONTD.)

- On applying the PCA using SPSS five tables are generated out of which three are important for explaining PCA.
- Table – 1: Toatl Variance Explained

Component	Eigen Value	% of Variance	Cumulative %	Component	Eigen Value	% of Variance	Cumulative %
1	5.804	64.490	64.490	6	0.155	1.719	96.491
2	1.517	16.860	81.349	7	0.139	1.547	98.038
3	0.623	6.918	88.267	8	0.114	1.266	99.305
4	0.338	3.757	92.025	9	0.063	0.695	100.000
5	0.247	2.747	94.772				

EXAMPLE (CONTD.)

- This table shows that nine principal components are generated against the 9 variables. Second column represents the Eigen value of correlation matrix. Third and fourth column shows the Percentage of variance and cumulative percentage of variance explained by these components.
- As we know that Eigen values are arranged in the descending order. In this example we can see that this process is followed. For first component the Eigen value is 5.804 which is largest among all the Eigen value. It also shows that the variance of first principal component is 5.804. The % of variance is 64.490 which is also equal to cumulative %. It shows that first principal component explain the 64.490% of total variation in the data. For second component cumulative % is 81.439, which shows that both, first and second component collectively explain the 81.439% of total variation in the data.

EXAMPLE (CONTD.)

- Table 2: Component Score Coefficient Matrix

	1	2	3	4	5	6	7	8	9
Price in thousands	0.105	-0.457	0.233	0.822	0.713	-0.399	-0.226	0.153	-1.773
Engine size	0.152	-0.160	0.183	-0.665	-0.838	1.068	0.742	0.779	-1.603
Horsepower	0.133	-0.351	0.435	-0.022	-0.370	-0.233	0.396	-0.660	2.765
Wheelbase	0.124	0.388	0.183	0.655	-0.110	-0.811	1.385	1.115	0.200
Width	0.143	0.159	0.282	-1.026	1.260	-0.510	-0.031	-0.181	-0.143
Length	0.126	0.337	0.545	0.351	-0.552	0.279	-1.344	-1.120	-0.652
Curb weight	0.159	0.025	-0.353	0.190	0.304	0.819	-1.261	1.616	1.367
Fuel capacity	0.149	0.078	-0.605	0.337	0.418	0.989	0.915	-1.578	-0.010
Fuel efficiency	-.0146	0.070	0.654	0.289	0.725	1.570	0.522	0.278	0.517

EXAMPLE (CONTD.)

- The Second table shows the coefficients for each variable in the corresponding principal components.
- It actually represents the Eigen vectors for the correlation matrix of variables.
- First Principal component can be written as:

U_1 = 0.105(Price in Thousand) + 0.152(Engine Size) + 0.133(Horse Power) + 0.124(Wheel Base) + 0.143 (Width) + 0.126(Length) + 0.159(Curb Weight) + 0.149(Fuel Capacity) + (-0.146)(Fuel Efficiency)

- In the same manner other Principal components can be written.

EXAMPLE (CONTD.)

- Table 3: Component score coefficient matrix.

Component	1	2	3	4	5	6	7	8	9
1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

- This table shows that all the principal components are linearly independent of each other and are normalized.

REFERENCES

1. Anderson TW, An introduction to Multivariate Statistical Analysis, 3rd Edition, John Wiley & Sons Inc., New Jersey.
2. Malhotra NK, Birks DF, Marketing Research an Applied Approach, 4th Edition, Prentice Hall, New Delhi.
3. Johnson RA, Wichern DW, Applied Multivariate Statistical Analysis, 3rd Edition, Prentice Hall, New Delhi.
4. Morrison DF, Multivariate Statistical Methods, 2nd Edition, McGraw Hill Publication, India.