



# STAT 505 - Applied Multivariate Statistical

## Start Here!

- ▶ [Welcome to STAT 505!](#)
- ▶ [Faculty login \(PSU Access Account\)](#)

## Lessons

- ▶ [Lesson 0: Matrices and Vectors](#)
- ▶ [Lesson 1: Graphical Display of Multivariate Data](#)
- ▶ [Lesson 2: Measures of Central Tendency, Dispersion and Association](#)
- ▶ [Lesson 3: Linear Combinations of Random Variables](#)
- ▶ [Lesson 4: Multivariate Normal Distribution](#)
- ▶ [Lesson 5: Sample Mean Vector and Sample Correlation and Related Inference Problems](#)
- ▶ [Lesson 6: Multivariate Conditional Distribution and Partial Correlation](#)
- ▼ [Lesson 7: Principal Components Analysis \(PCA\)](#)
  - ◉ [7.1 - Principal Component Analysis \(PCA\) Procedure](#)
  - ◉ [7.2 - How do we find the coefficients?](#)
  - ◉ [7.3 - Example: Places Rated](#)
  - ◉ [7.4 - Interpretation of the Principal Components](#)
  - ◉ [7.5 - Alternative: Standardize the Variables](#)
  - ◉ [7.6 - Example: Places Rated after Standardization](#)
  - ◉ [7.7 - Once the Components Have Been Calculated](#)
  - ◉ [7.8 - Summary](#)
- ▶ [Lesson 8: Canonical Correlation Analysis](#)
- ▶ [Lesson 9: Factor Analysis](#)
- ▶ [Lesson 10: Discriminant Analysis](#)
- ▶ [Lesson 11: Inferences Regarding Multivariate Population Mean](#)
- ▶ [Lesson 12: Cluster Analysis](#)
- ▶ [Lesson 13: Multivariate Analysis of Variance \(MANOVA\)](#)

[Home](#) // [Lesson 7: Principal Components Analysis \(PCA\)](#)

## 7.3 - Example: Places Rated

[Printer-friendly version](#)

We will use the Places Rated Almanac data (Boyer and Savageau) which rates 329 communities according to nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

Notes:

- The data for many of the variables are strongly skewed to the right.
- The log transformation was used to normalize the data.

Using SAS

Using Minitab

The SAS program [places.sas](#) will implement the principal component

- ▶ [Lesson 14: Repeated Measures Analysis](#)

## Resources

- [SAS Tour](#)
- [SAS Interactive Data Analysis](#)
- [A Quick Introduction to Minitab](#)
- [Worked Examples from the Course That Use Software](#)

```
options ls=78;
title "PCA - Covariance Matrix - Places Rated";

data places;
  infile "D:\Statistics\STAT 505\data\places.txt";
  input climate housing health crime trans educate arts r
  climate=log10(climate);
  housing=log10(housing);
  health=log10(health);
  crime=log10(crime);
  trans=log10(trans);
  educate=log10(educate);
  arts=log10(arts);
  recreate=log10(recreate);
  econ=log10(econ);
run;

proc princomp cov out=a;
  var climate housing health crime trans educate arts rec
run;

proc corr;
  var prin1 prin2 prin3 climate housing health crime tran
  recreate econ;
run;

proc gplot;
  axis1 length=5 in;
  axis2 length=5 in;
  plot prin2*prin1 / vaxis=axis1 haxis=axis2;
  symbol v=J f=special h=2 i=none color=black;
run;
```



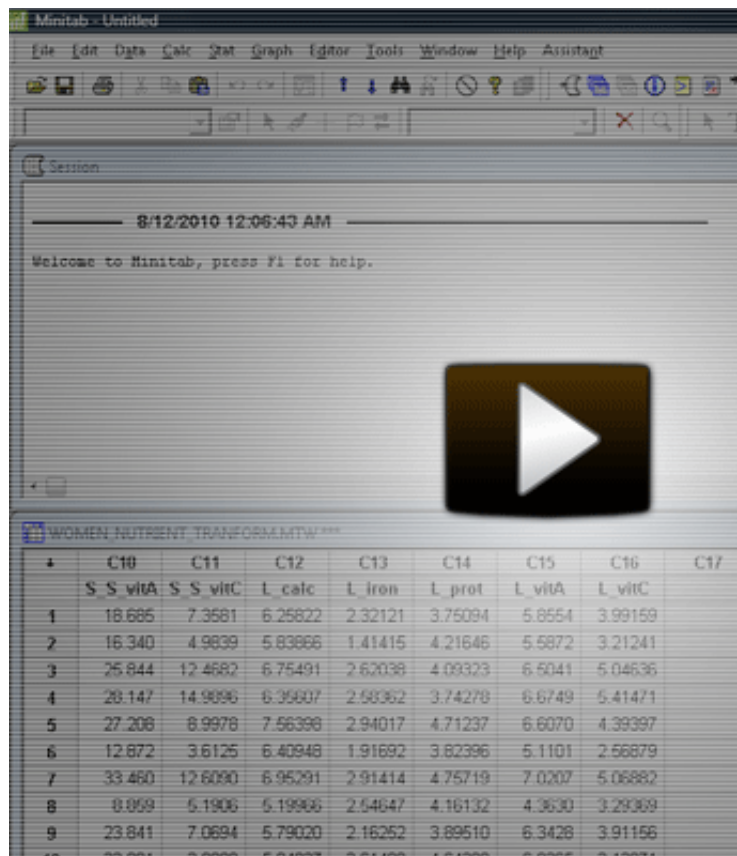
When you examine the output, the first thing that SAS does is to give 329 observations representing the 329 communities in our dataset and simple statistics that report the means and standard deviations for each variable.

Below this is the variance-covariance matrix for the data. You should note that the variance reported for climate is 0.01289.

What we really need to draw our attention to here is the eigenvalues of the covariance matrix. In the SAS output the eigenvalues are ranked order from largest to smallest. We will use the first three eigenvalues into Table 1 below for discussion.

-

Click on the arrow in the window below to see how to perform a principal component analysis on the places Rated data using the Minitab statistical software application.



The screenshot shows the Minitab software interface. At the top is a menu bar with options: File, Edit, Data, Calc, Stat, Graph, Editor, Tools, Window, Help, Assistant. Below the menu bar is a toolbar with various icons. The main window displays a session log with the date and time '8/12/2010 12:06:43 AM' and a welcome message. A large play button icon is centered in the session window. Below the session window, a data table titled 'WOMEN NUTRIENT TRANSFORM.MTW' is visible. The table has 10 rows and 9 columns. The columns are labeled C10 through C17, with sub-labels: S S vitA, S S vitC, L calc, L iron, L prot, L vitA, L vitC. The rows contain numerical data values.

	C10	C11	C12	C13	C14	C15	C16	C17
	S S vitA	S S vitC	L calc	L iron	L prot	L vitA	L vitC	
1	18.686	7.3581	6.25822	2.32121	3.75094	5.8554	3.99159	
2	16.340	4.9839	5.83866	1.41415	4.21646	5.5872	3.21241	
3	25.844	12.4682	6.75491	2.62038	4.09323	6.5041	5.04636	
4	20.147	14.9896	6.36007	2.58362	3.74278	6.6749	5.41471	
5	27.208	8.9978	7.56398	2.94017	4.71237	6.6070	4.39397	
6	12.872	3.6125	6.40948	1.91692	3.82396	5.1101	2.56879	
7	33.460	12.6090	6.95291	2.91414	4.75719	7.0207	5.06882	
8	0.059	5.1906	5.19946	2.54647	4.16132	4.3630	3.29369	
9	23.841	7.0694	5.79020	2.16252	3.89510	6.3428	3.91156	
10	33.081	2.8990	5.94877	2.51408	4.54789	6.9365	2.13871	

Principle components analysis using N

### Data Analysis:

**Step 1:** We examine the eigenvalues to determine how many principal components should be considered:

Table 1. Eigenvalues, and the proportion of variation explained by the principal components.

Component	Eigenvalue	Proportion	Cumulative
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0535	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966

9	0.0018	0.0034	1.0000
<b>Total</b>	0.5225		

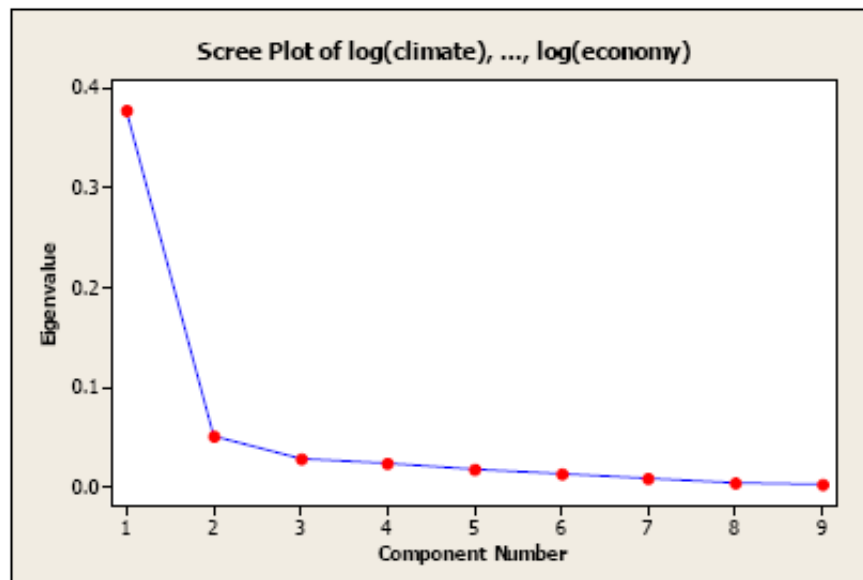
If you take all of these eigenvalues and add them up and you get the total variance of 0.5223.

The proportion of variation explained by each eigenvalue is given in the third column. For example, 0.3775 divided by the 0.5223 equals 0.7227, or, about 72% of the variation is explained by this first eigenvalue. The cumulative percentage explained is obtained by adding the successive proportions of variation explained to obtain the running total. For instance, 0.7227 plus 0.0977 equals 0.8204, and so forth. Therefore, about 82% of the variation is explained by the first two eigenvalues together.

Next we need to look at successive differences between the eigenvalues. Subtracting the second eigenvalue 0.051 from the first eigenvalue, 0.377 we get a difference of 0.326. The difference between the second and third eigenvalues is 0.0232; the next difference is 0.0049. Subsequent differences are even smaller. A sharp drop from one eigenvalue to the next may serve as another indicator of how many eigenvalues to consider.

The first three principal components explain 87% of the variation. This is an acceptably large percentage.

An **Alternative Method** to determine the number of principal components is to look at a Scree Plot. With the eigenvalues ordered from largest to the smallest, a scree plot is the plot of  $\hat{\lambda}_i$  versus  $i$ . The number of component is determined at the point, beyond which the remaining eigenvalues are all relatively small and of comparable size. The following plot is made in Minitab.



The scree plot for the variables without standardization (covariance matrix)

As you see, we could have stopped at the second principal component, but we continued till the third component. Relatively speaking, contribution of the third component is small compared to the second component.

**Step 2:** Next, we will compute the principal component scores. For example, the first principal component can be computed using the elements of the first eigenvector:

$$Y_1 = 0.0351 \times (\text{climate}) + 0.0933 \times (\text{housing}) + 0.4078 \times (\text{crime}) + 0.1004 \times (\text{transportation}) + 0.08743 \times (\text{arts}) + 0.1590 \times (\text{recreation}) + 0.0195 \times (\text{economy})$$

In order to complete this formula and compute the principal component for the individual community of interest, plug in that community's values for each of these variables. A fairly standard procedure is, rather than using the raw data here, to use the difference between the variables and their sample means. This is known as the standardized values of the random variables. Standardization does not affect the interpretation.

Magnitudes of the coefficients give the contributions of each variable to that component. However, the magnitude of the coefficients also depend on the variances of the corresponding variables.

◀ 7.2 - How do we find the coefficients?

7.4 - Interpretation of the Principal Components ▶



[Printer-friendly version](#)



Drupal

© 2014 The Pennsylvania State University. All rights reserved.