SPIDERFINANCIAL
www.spiderfinancial.com

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Tutorial: Principal Component 102

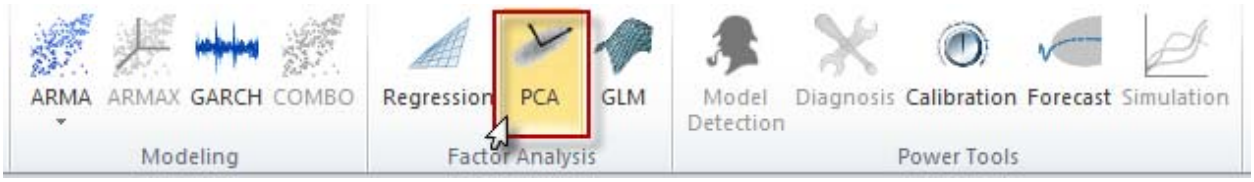
This is the second entry in our principal components analysis (PCA) series. In this tutorial, we will resume our discussion on dimension reduction using a subset of the principal components with a minimal loss of information. We will use NumXL and Excel to carry out our analysis, closely examining the different output elements in an attempt to develop a solid understanding of PCA, which will pave the way to a more advanced treatment in future issues.

In this tutorial, we will continue to use the socioeconomic data provided by Harman (1976). The five variables represent total population ("Population"), median school years ("School"), total employment ("Employment"), miscellaneous professional services ("Services"), and median house value ("House Value"). Each observation represents one of twelve census tracts in the Los Angeles Standard Metropolitan Statistical Area.

District	population	median school yrs	total employment	misc professional services	median house value
1	5700	12.8	2500	270	$ 25,000
2	1000	10.9	600	10	$ 10,000
3	3400	8.8	1000	10	$ 9,000
4	3800	13.6	1700	140	$ 25,000
5	4000	12.8	1600	140	$ 25,000

Process

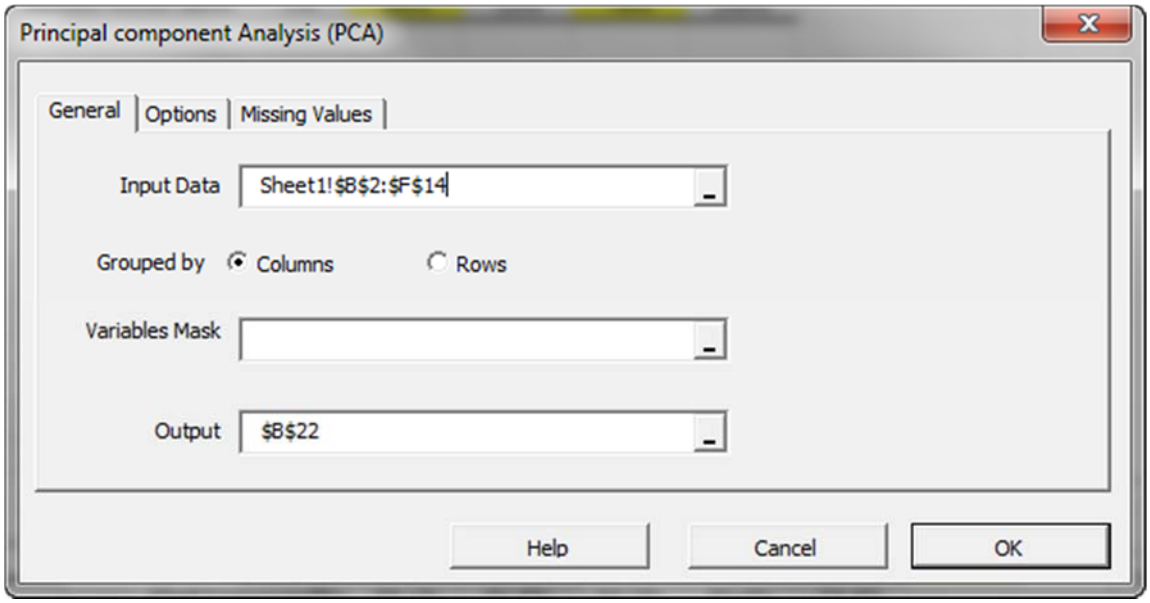
Now we are ready to conduct our principal component analysis. First, select an empty cell in your worksheet where you wish the output to be generated, then locate and click on the principal component (PCA) icon in the NumXL tab (or toolbar).



The Regression Wizard will appear.

 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

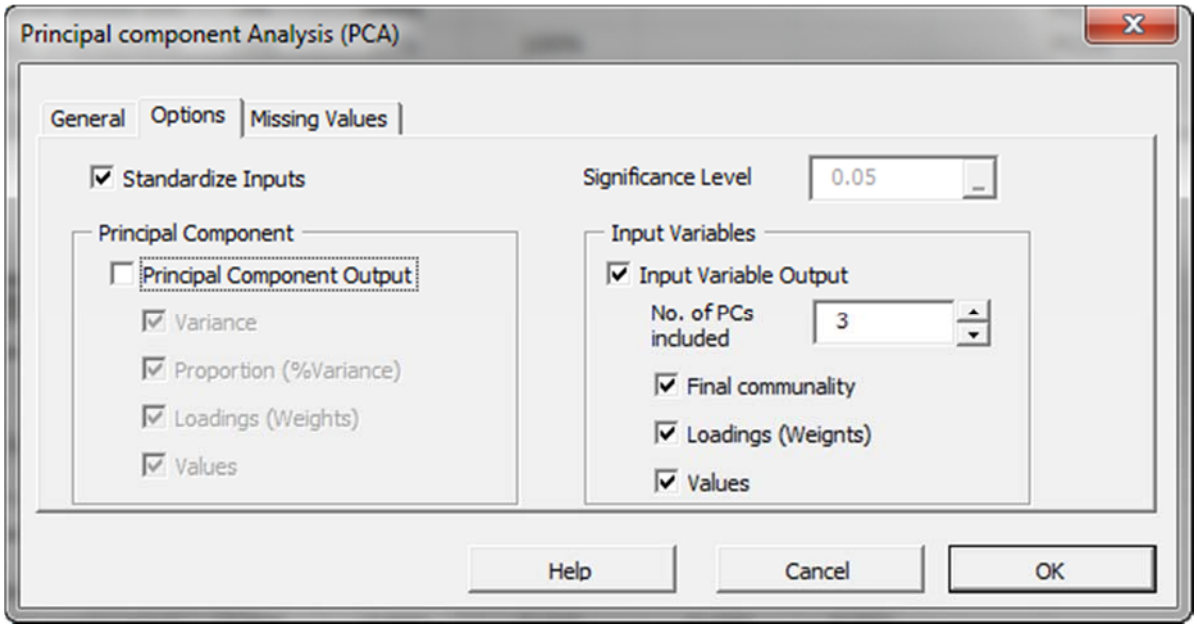


Select the cells range for the five input variable values.

Notes:

1. The cells range includes (optional) the heading (Label) cell, which would be used in the output tables where it references those variables.
2. The input variables (i.e. X) are already grouped in columns (each column represents a variable), so we don't need to change that.
3. Leave the "Variables Mask" field blank for now. We will revisit this field in later entries.

Next, select the "Options" tab.



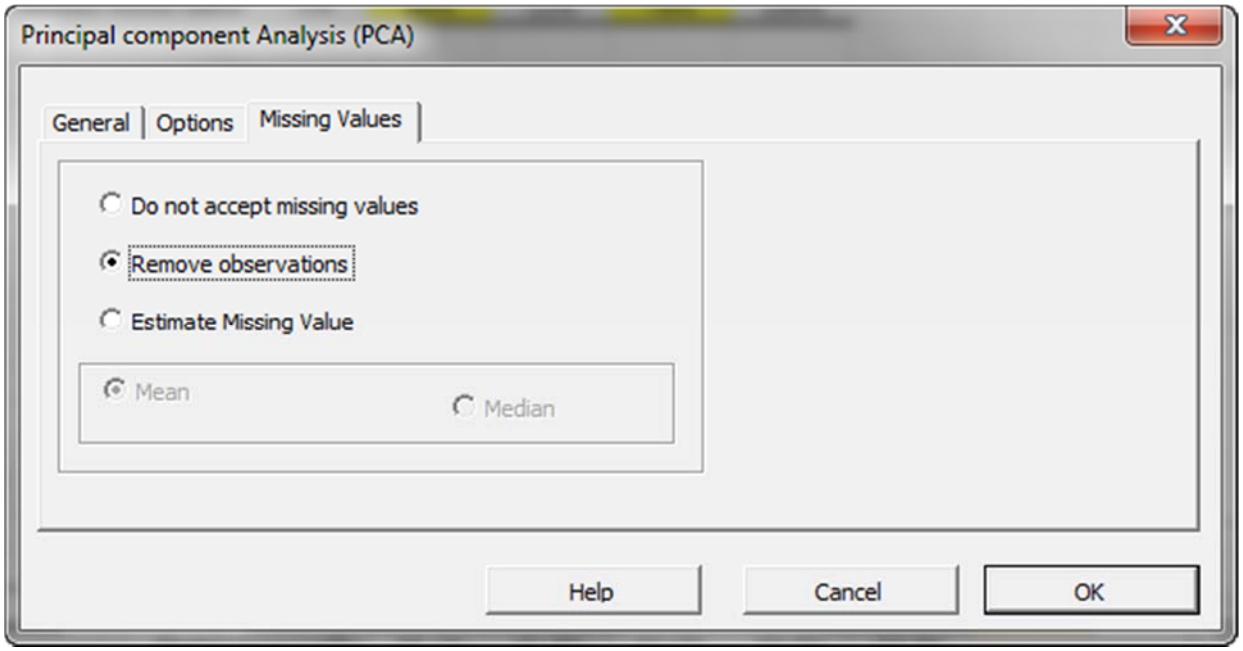
 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Initially, the tab is set to the following values:

- "Standardize Inputs" is checked. **Leave this option checked.**
- "Principal Component Output" is checked. **Uncheck it.**
- The significance level (aka. α) is set to 5%.
- "Input Variables" is unchecked. **Check this option.**
- Set "No. of PCs included" to 3. This action can be done now or altered later in the output tables, as our formulas are dynamic.
- Under "Input Variables", check the "Values" option, so the generated output tables include a fitted value for the input variables using a reduced set of components.

Now, click the "Missing Values" tab.



In this tab, you can select an approach to handle missing values in the data set (X and Y). By default, any missing value found in any observation would exclude the observation from the analysis.

This treatment is a good approach for our analysis, so let's leave it unchanged.

Now, click "OK" to generate the output tables.

 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

	population	median school yrs	total employment	misc professional services	median house value	3
Final Communality	98.1%	91.8%	94.2%	93.8%	93.9%	

Loadings	population	median school yrs	total employment	misc professional services	median house value
PC(1)	0.23	0.50	0.34	0.56	0.52
PC(2)	-0.66	0.32	-0.59	0.01	0.34
PC(3)	-0.64	-0.38	0.43	0.49	-0.15
PC(4)	#N/A	#N/A	#N/A	#N/A	#N/A
PC(5)	#N/A	#N/A	#N/A	#N/A	#N/A

Values	population	median school yrs	total employment	misc professional services	median house value
	5428.42	13.18	2662.21	257.87	24105.14
	1280.87	10.11	505.20	8.07	12914.74
	3083.11	9.03	1229.27	-12.10	9046.57
	4236.77	13.43	1356.03	176.00	24173.37
	4163.44	13.07	1408.92	165.85	22990.44
	7725.77	9.20	2839.69	47.45	9249.81
	8160.75	11.79	789.46	-27.46	16060.27
	9669.86	10.95	2913.47	94.62	14619.26
	10038.31	12.16	3343.86	180.93	19177.30
	9035.92	13.86	4054.07	341.55	26335.62
	9536.34	9.82	3313.54	82.01	11122.77
	10040.43	10.68	3584.27	135.20	14204.72

Analysis

1. Statistics

	population	median school yrs	total employment	misc professional services	median house value	3
Final Communality	98.1%	91.8%	94.2%	93.8%	93.9%	

 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

In this table, we show the percentage of variance of each input variable accounted for (aka final communality) using the first three (3) factors. Unlike the cumulative proportion, this statistic is related to one input variable at a time.

Using this table, we can detect which input variables are poorly presented (i.e. adversely affected) by our dimension reduction. In this example, the "median school years" has the lowest value, yet the final communality is still around 92%.

2. Loadings

In the loading table, we outline the weights of the principal component in each input variable:

Loadings	population	median school yrs	total employment	misc professional services	median house value
PC(1)	0.23	0.50	0.34	0.56	0.52
PC(2)	-0.66	0.32	-0.59	0.01	0.34
PC(3)	-0.64	-0.38	0.43	0.49	-0.15
PC(4)	#N/A	#N/A	#N/A	#N/A	#N/A
PC(5)	#N/A	#N/A	#N/A	#N/A	#N/A

To compute the values of an input variable using PC values, we use the weights above to linearly transform them back. For example, the population factor is expressed as follows:

$$X_1 = 0.23PC_1 - 0.66PC_2 - 0.64PC_3$$

Notes:

- This table is basically the transposed table (row turned into columns) that we saw in the variables' loadings for PCs.
- The sum of squares of each row must be 1.
- $PC_1, PC_2, PC_3, .., PC_m$ are uncorrelated, so to compute the variance of X_1 (standardized) using first k-components:

$$\text{Var}[\hat{X}_i] = \gamma_1^2\sigma_1^2 + \gamma_2^2\sigma_2^2 + \ldots + \gamma_k^2\sigma_k^2$$

$$X_i = \frac{x_i - \overline{x}_i}{\sigma_{x_i}}, \text{ thus } \text{Var}[X_i] = 1$$

$$\gamma_1^2\sigma_1^2 + \gamma_2^2\sigma_2^2 + \ldots + \gamma_k^2\sigma_k^2 \le 1$$
$$\gamma_1^2\sigma_1^2 + \gamma_2^2\sigma_2^2 + \ldots + \gamma_m^2\sigma_m^2 = 1$$

 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Where:

- γ_k is the loading of the k-th principal component for the input variable X_i
- σ_k^2 is the variance of the k-th principal component.
- $\hat{X}_i$ is the estimate for the standardized input variable using the first k-components.

- By definition, the $\text{Var}[\hat{X}_i]$ is the final communality.
- The variance of the fitted input variable in the original scale (non-standardized) is expressed as follows:

$$\text{Var}[\hat{x}_i]=(\gamma_1^2\sigma_1^2+\gamma_2^2\sigma_2^2+\ldots+\gamma_k^2\sigma_k^2)\times\sigma_{x_i}^2$$

- Reducing the dimension, in essence, reduces the variance of the input variables (low pass filter).
- What about the correlation among original variables? How are they affected by reducing the dimensions?

$$\hat{X}_i=\gamma_1 PC_1+\gamma_2 PC_2+\ldots+\gamma_k PC_k$$
$$\hat{X}_j=\omega_1 PC_1+\omega_2 PC_2+\ldots+\omega_k PC_k$$
$$\text{Cov}[\hat{X}_i,\hat{X}_j]=\rho_{\hat{x}_i\hat{x}_j}=E[\hat{X}_i\times\hat{X}_j]=E[\gamma_1\omega_1 PC_1^2+\gamma_2\omega_2 PC_2^2+\ldots+\gamma_k\omega_k PC_k^2]=\sum_{l=1}^{k}\gamma_l\omega_l\sigma_l^2$$
$$\text{Cov}[\text{X}_i,\text{X}_j]=\rho_{x_ix_j}=\sum_{l=1}^{m}\gamma_l\omega_l\sigma_l^2$$
$$\rho_{\hat{x}_i\hat{x}_j}=\rho_{x_ix_j}-\sum_{l=k+1}^{m}\gamma_l\omega_l\sigma_l^2$$

Where:

- $\hat{X}_i$ is the standardized fitted i-th variable.
- X_i is the original standardized i-th variable.
- x_i is the original non-standardized ith variable.
- $\hat{x}_i$ is the fitted non-standardized i-th variable.

The correlation between the fitted variables (i.e. reduced dimensions) may be slightly different depending on the variance of the dropped factors and the loading of those factors in each variable.

- How about covariance between the variables (non-standardized)?

 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

$$\hat{X}_i = \frac{\hat{x}_i - \overline{x}_i}{\sigma_{x_i}} \Rightarrow \hat{x}_i - \overline{x}_i = \hat{X}_i \times \sigma_{x_i}$$

$$\text{Cov}[\hat{x}_i.\hat{x}_j] = \text{E}[(\hat{x}_i - \overline{x}_i)(\hat{x}_j - \overline{x}_j)] = \sigma_{x_i}\sigma_{x_j} \times E[\hat{X}_i \hat{X}_j]$$

$$\text{Cov}[\hat{x}_i.\hat{x}_j] = \sigma_{x_i}\sigma_{x_j} \times \sigma_{\hat{x}_i\hat{x}_j} = \sigma_{x_i}\sigma_{x_j} \times (\sigma_{x_i x_j} - \sum_{l=k+1}^{m} \gamma_l \omega_l \sigma_l^2)$$

$$\text{Cov}[\hat{x}_i.\hat{x}_j] = \text{Cov}[x_i.x_j] - \sigma_{x_i}\sigma_{x_j} \times \sum_{l=k+1}^{m} \gamma_l \omega_l \sigma_l^2$$

In sum, the relative change in covariance is equal to the change in correlation between the two variables.

Note: Reducing the number of factors alters the statistical characteristics of the underlying data set, so extreme care must be taken.

3. Fitted Input Values

Using the first three (3) principal components, NumXL calculates the fitted value for each input variable:

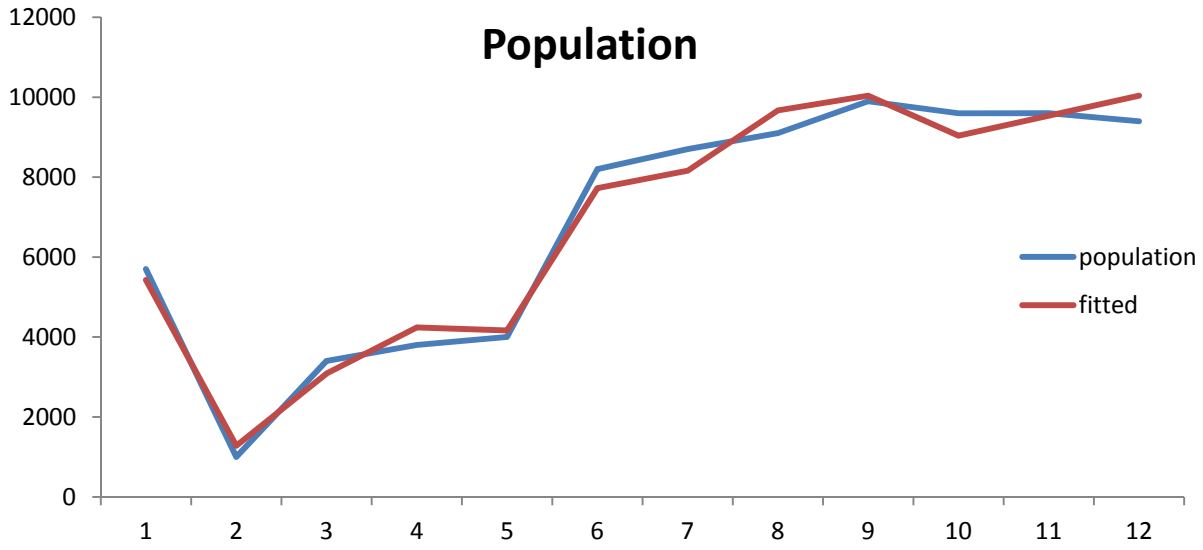
Values	population	median school yrs	total employment	misc professional services	median house value
	5428.42	13.18	2662.21	257.87	24105.14
	1280.87	10.11	505.20	8.07	12914.74
	3083.11	9.03	1229.27	-12.10	9046.57
	4236.77	13.43	1356.03	176.00	24173.37
	4163.44	13.07	1408.92	165.85	22990.44
	7725.77	9.20	2839.69	47.45	9249.81
	8160.75	11.79	789.46	-27.46	16060.27
	9669.86	10.95	2913.47	94.62	14619.26
	10038.31	12.16	3343.86	180.93	19177.30
	9035.92	13.86	4054.07	341.55	26335.62
	9536.34	9.82	3313.54	82.01	11122.77
	10040.43	10.68	3584.27	135.20	14204.72

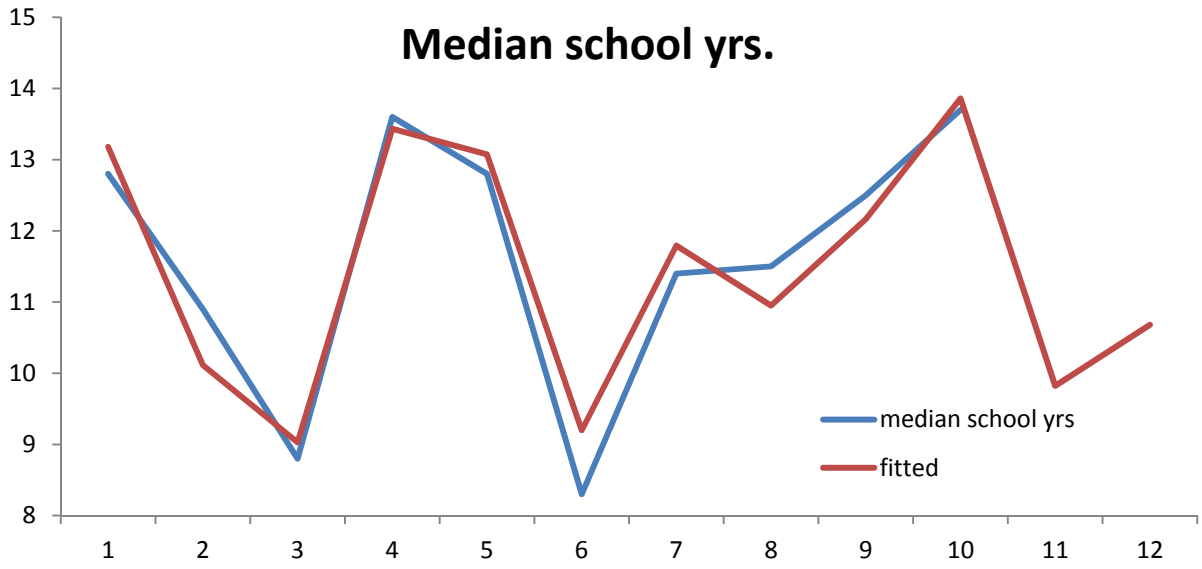
Note: Although the PCA uses the standardized version of the input variables, it computes the fitted values in the original non-standardized format.

 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Let's plot population (highest final communality) and median school years (lowest final communality) for the original data and for the fitted one.





 © Spider Financial Corp, 2013

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Conclusion

In this tutorial, we examined the dimension reduction proposition from 5 PCs to 3 PCs without significant loss of information.

What do we do now?

In the first two tutorials, we focused on delivering the key ideas behind the principal component analysis and, to some extent, the rationale behind the dimension reduction proposition. The cross-section socio-economic sample data, although not a time series, served to demonstrate the theory and to show NumXL's different output tables.

In the third entry of this series, we are ready to look into a set of correlated time series, apply PCA technique to derive a reduced core set of uncorrelated drivers. Next, we forecast the values (mean and standard error) for the uncorrelated drivers, and using the PCA Loadings, imply the corresponding forecast (mean and error) for each input variable.

 © Spider Financial Corp, 2013