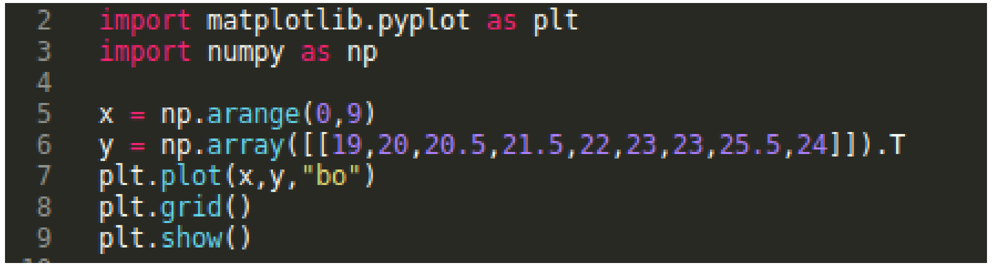
Linear Regression and PCA

Linear Regression

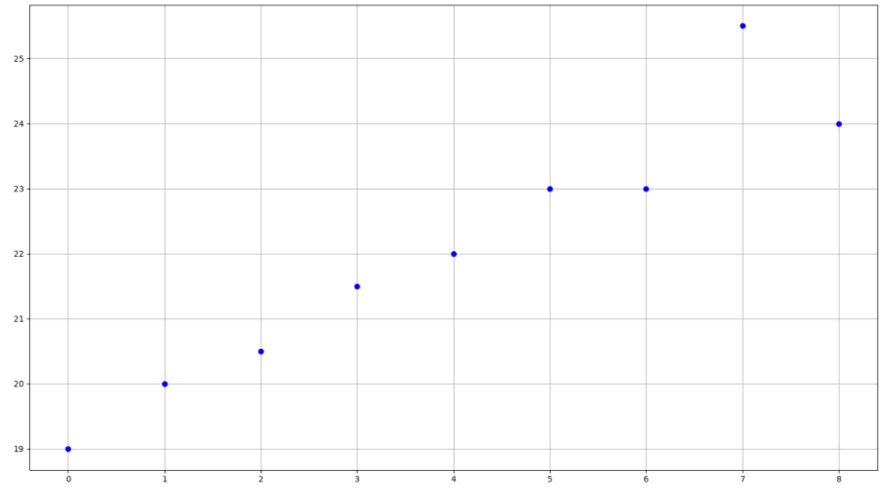
Small example

Suppose you have 9 points on the plane, defined by their (x,y) coordinates

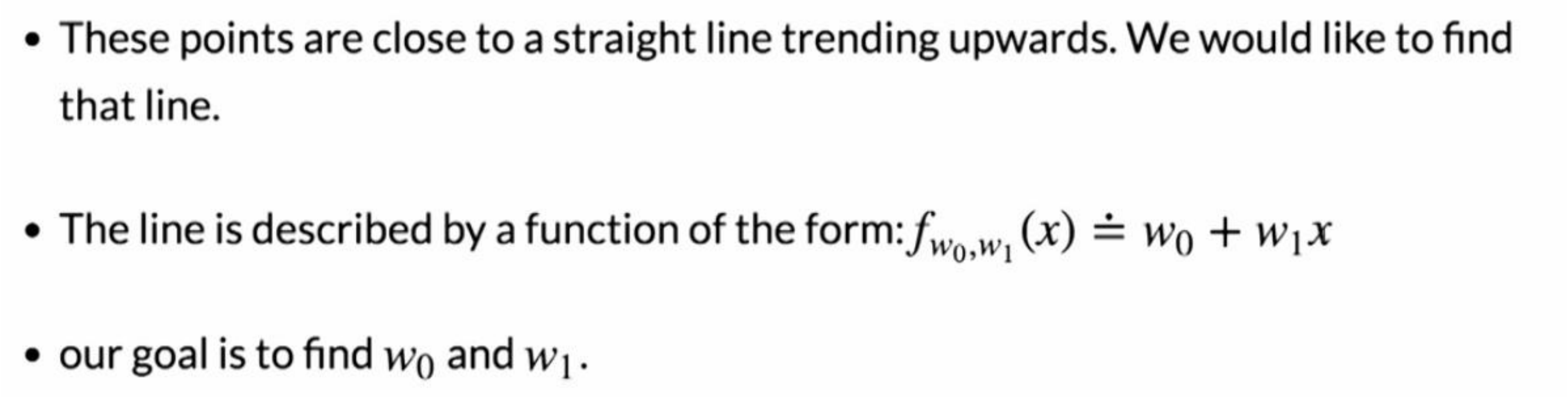


```python
import matplotlib.pyplot as plt
import numpy as np

x = np.arange(0,9)
y = np.array([[19,20,20.5,21.5,22,23,23,25.5,24]]).T
plt.plot(x,y,"bo")
plt.grid()
plt.show()
```

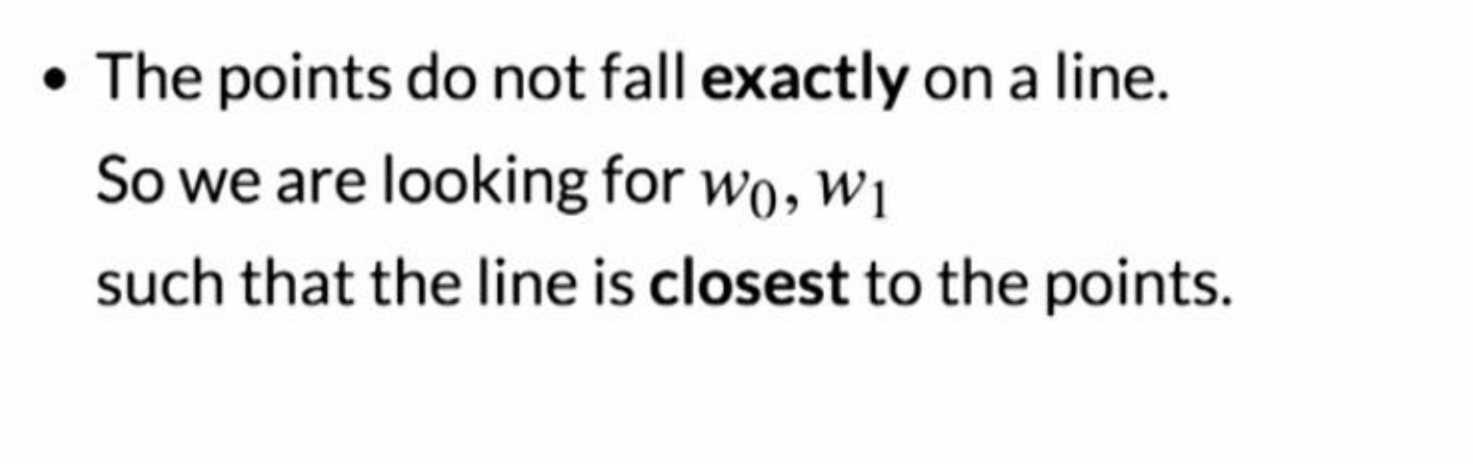


Linear Regression



- These points are close to a straight line trending upwards. We would like to find that line.
- The line is described by a function of the form: $f_{w_0,w_1}(x) \doteq w_0 + w_1 x$
- our goal is to find w_0 and w_1.

Linear Regression

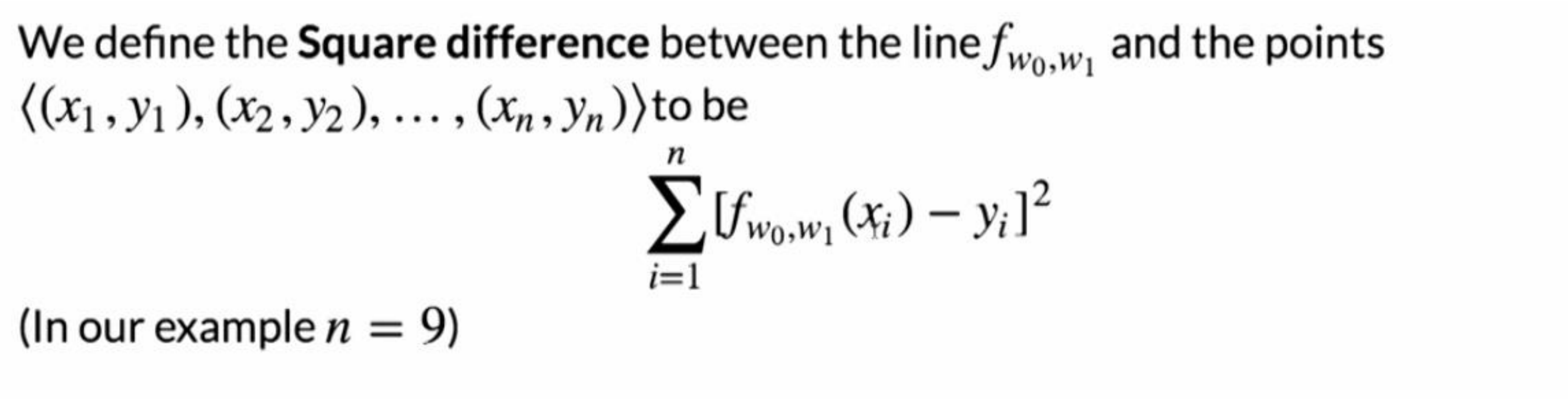


- The points do not fall **exactly** on a line.

  So we are looking for w_0, w_1

  such that the line is **closest** to the points.

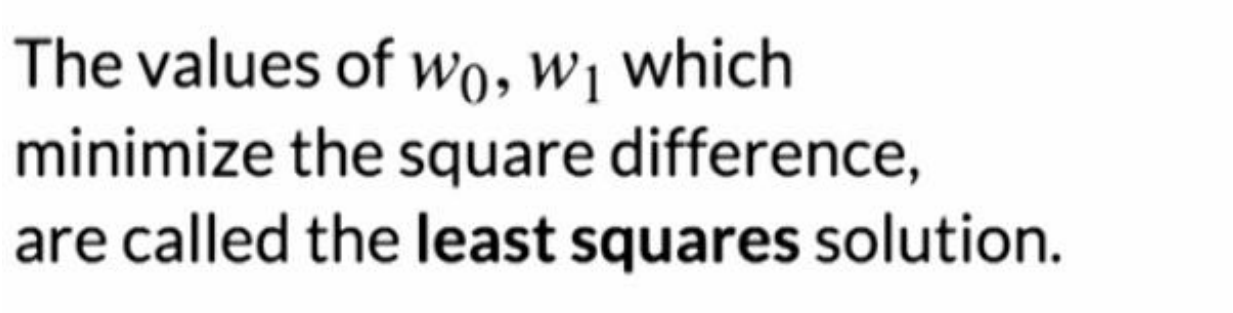
Linear Regression



We define the **Square difference** between the line f_{w_0,w_1} and the points $\langle (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \rangle$ to be

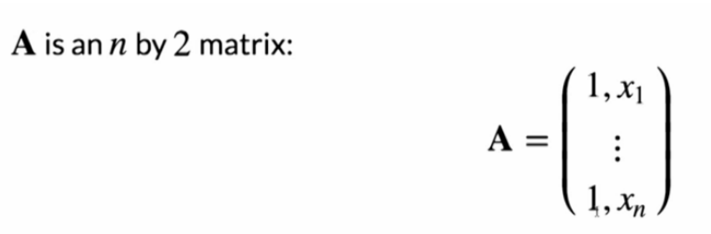
$$\sum_{i=1}^{n} [f_{w_0,w_1}(x_i) - y_i]^2$$

(In our example $n = 9$)



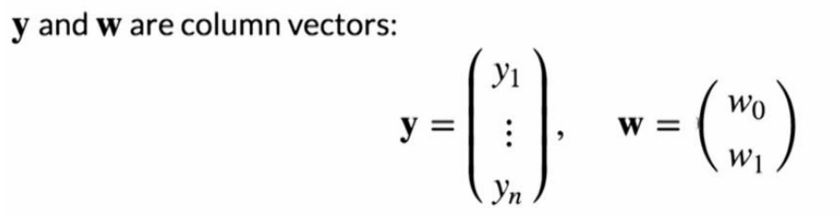
The values of w_0, w_1 which
minimize the square difference,
are called the **least squares** solution.

Linear Regression



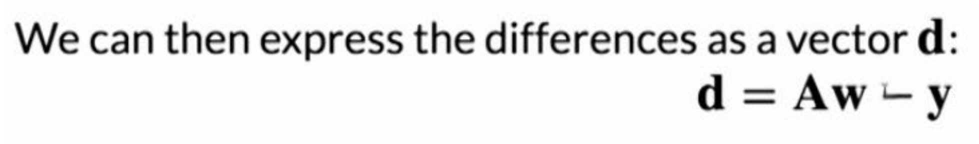
$\mathbf{A}$ is an n by 2 matrix:

$$\mathbf{A} = \begin{pmatrix} 1, x_1 \\ \vdots \\ 1, x_n \end{pmatrix}$$



$\mathbf{y}$ and $\mathbf{w}$ are column vectors:

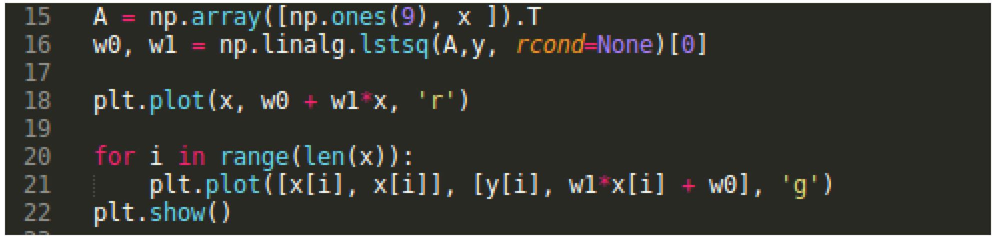
$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$$



We can then express the differences as a vector $\mathbf{d}$:

$$\mathbf{d} = \mathbf{A}\mathbf{w} - \mathbf{y}$$

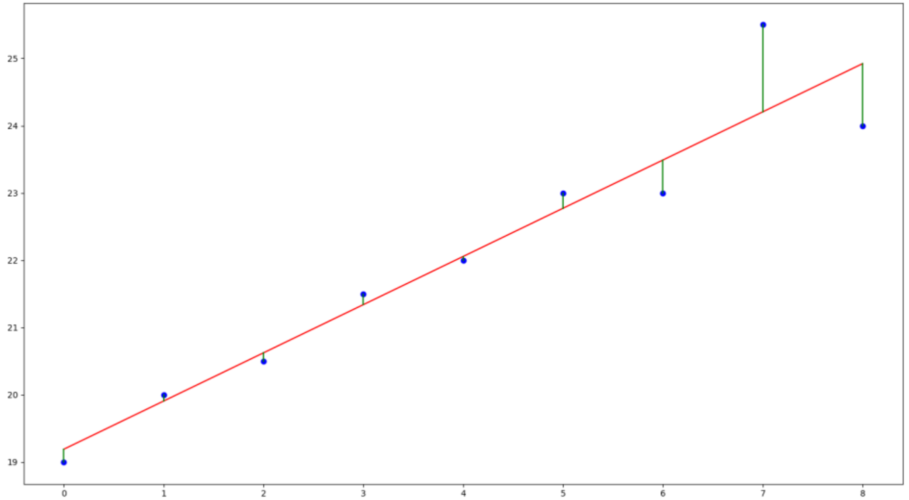
Linear Regression



```
A = np.array([np.ones(9), x ]).T
w0, w1 = np.linalg.lstsq(A,y, rcond=None)[0]

plt.plot(x, w0 + w1*x, 'r')

for i in range(len(x)):
    plt.plot([x[i], x[i]], [y[i], w1*x[i] + w0], 'g')
plt.show()
```



Principal Component Analysis

WHY PCA?

1. Try to train the models on original number of features, which take days or weeks if the number of features is too high.
2. Reduce the number of variables by merging correlated variables.
3. Extract the most important features from the dataset that are responsible for maximum variance in the output. Different statistical techniques are used for this purpose e.g. linear discriminant analysis, factor analysis, and principal component analysis.

Example: Places Rated

In the Places Rated Almanac, Boyer and Savageau rated 329 communities according to the following nine criteria:

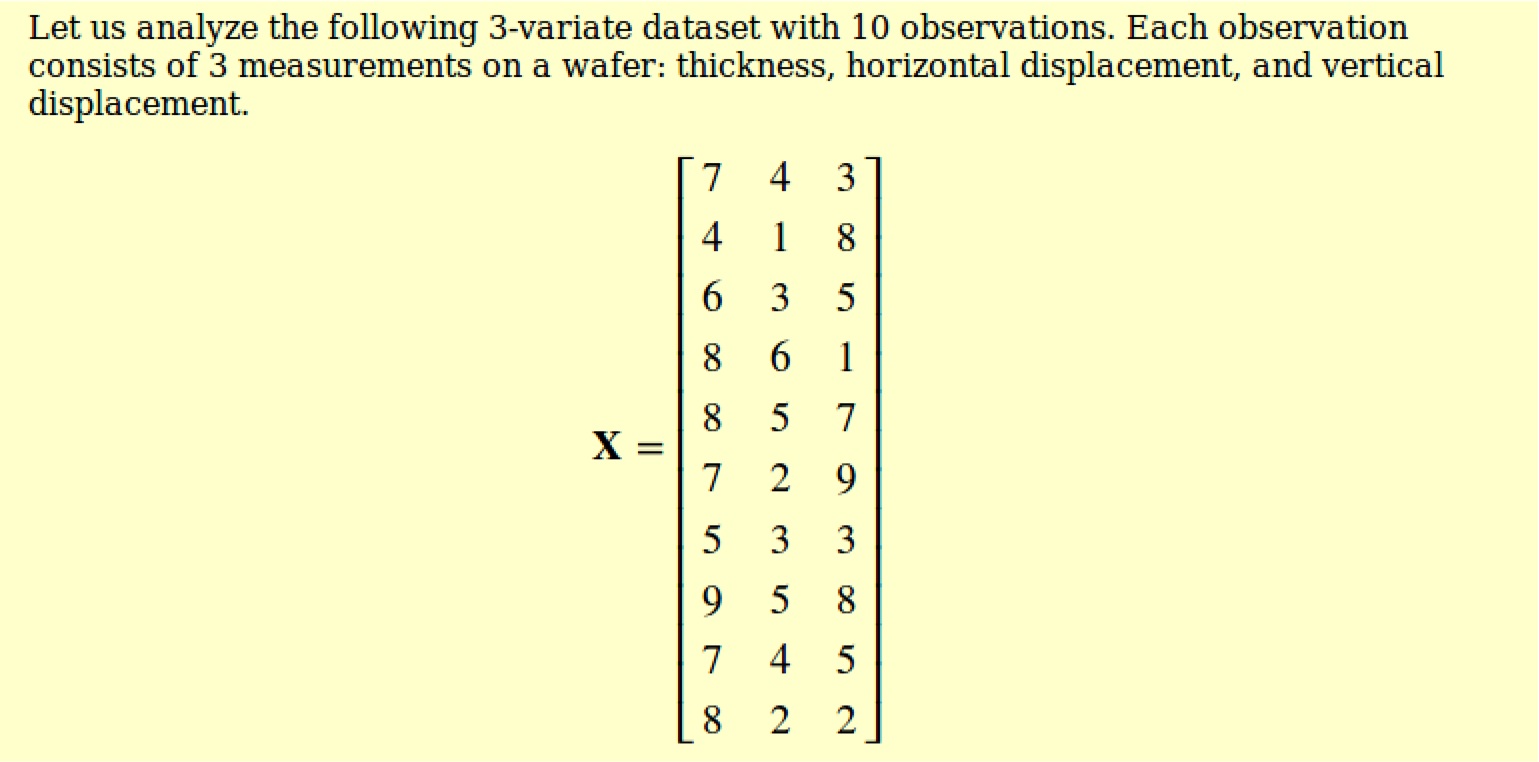
1. **Climate and Terrain**
2. **Housing**
3. **Health Care & the Environment**
4. **Crime**
5. **Transportation**
6. **Education**
7. **The Arts**
8. **Recreation**
9. **Economics**

Objective

With a large number of variables, the dispersion matrix may be too large to study and interpret properly. There would be too many pairwise correlations between the variables to consider.

To interpret the data in a more meaningful form, it is necessary to reduce the number of variables to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component.

Numerical Example



Let us analyze the following 3-variate dataset with 10 observations. Each observation consists of 3 measurements on a wafer: thickness, horizontal displacement, and vertical displacement.

$$\mathbf{X} = \begin{bmatrix} 7 & 4 & 3 \\ 4 & 1 & 8 \\ 6 & 3 & 5 \\ 8 & 6 & 1 \\ 8 & 5 & 7 \\ 7 & 2 & 9 \\ 5 & 3 & 3 \\ 9 & 5 & 8 \\ 7 & 4 & 5 \\ 8 & 2 & 2 \end{bmatrix}$$

Compute the correlation matrix

First compute the correlation matrix.

$$\mathbf{R} = \begin{bmatrix} 1.00 & 0.67 & -0.10 \\ 0.67 & 1.00 & -0.29 \\ -0.10 & -0.29 & 1.00 \end{bmatrix}$$

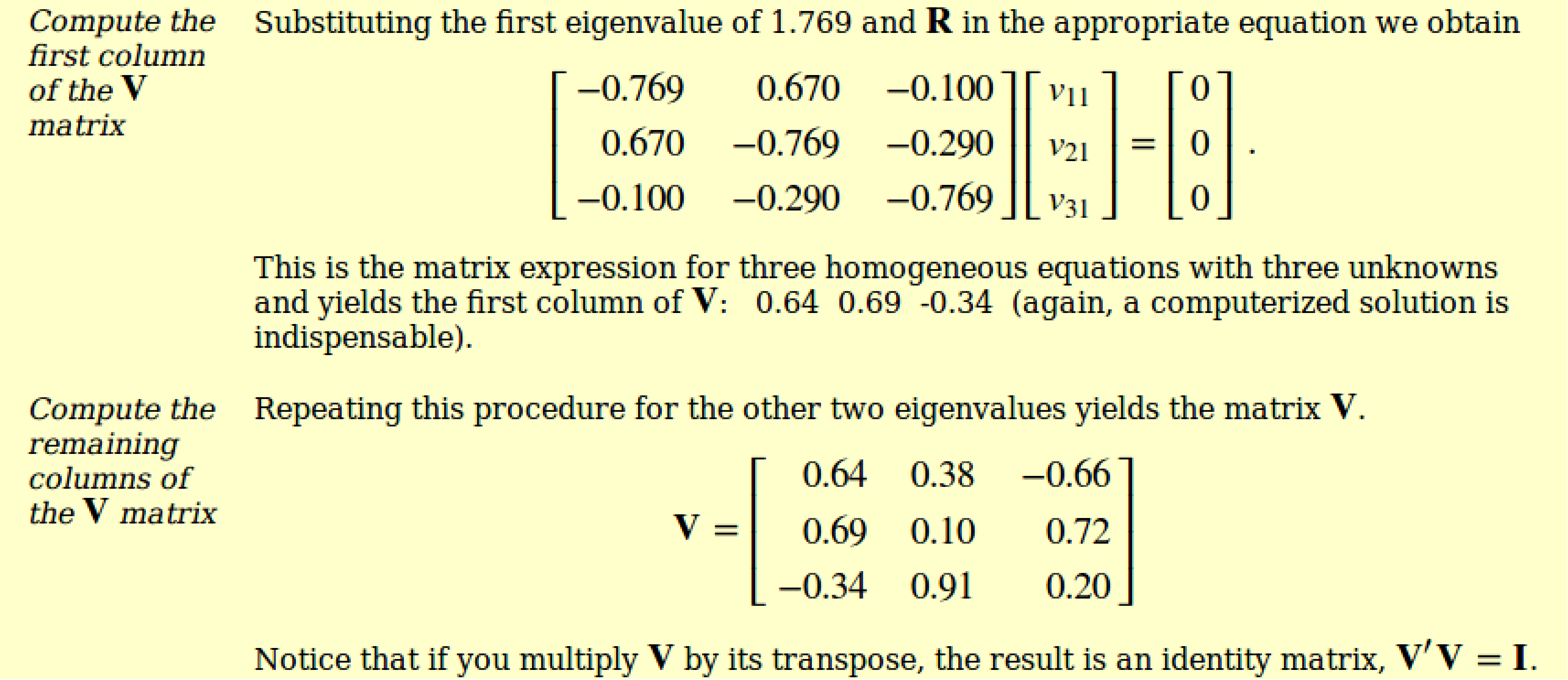
Solve for the roots of **R**

Next solve for the roots of **R** using software.

λ	**value**	**proportion**
1	1.769	0.590
2	0.927	0.899
3	0.304	1.000

Notice that:

- Each eigenvalue satisfies $|\mathbf{R} - \lambda\mathbf{I}| = 0$.
- The sum of the eigenvalues $= 3 = p$, which is equal to the trace of **R** (i.e., the sum of the main diagonal elements).
- The determinant of **R** is the product of the eigenvalues.
- The product is $\lambda_1 \times \lambda_2 \times \lambda_3 = 0.499$.



Compute the first column of the **V** *matrix*

Substituting the first eigenvalue of 1.769 and **R** in the appropriate equation we obtain

$$\begin{bmatrix} -0.769 & 0.670 & -0.100 \\ 0.670 & -0.769 & -0.290 \\ -0.100 & -0.290 & -0.769 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{21} \\ v_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} .$$

This is the matrix expression for three homogeneous equations with three unknowns and yields the first column of **V**: 0.64 0.69 -0.34 (again, a computerized solution is indispensable).

Compute the remaining columns of the **V** *matrix*

Repeating this procedure for the other two eigenvalues yields the matrix **V**.

$$\mathbf{V} = \begin{bmatrix} 0.64 & 0.38 & -0.66 \\ 0.69 & 0.10 & 0.72 \\ -0.34 & 0.91 & 0.20 \end{bmatrix}$$

Notice that if you multiply **V** by its transpose, the result is an identity matrix, $\mathbf{V'V} = \mathbf{I}$.

Compute the $\mathbf{L}^{1/2}$ *matrix*

Now form the matrix $\mathbf{L}^{1/2}$, which is a diagonal matrix whose elements are the square roots of the eigenvalues of $\mathbf{R}$. Then obtain $\mathbf{S}$, the factor structure, using $\mathbf{S} = \mathbf{V}\mathbf{L}^{1/2}$.

$$\begin{bmatrix} 0.64 & 0.38 & -0.66 \\ 0.69 & 0.10 & 0.72 \\ -0.34 & 0.91 & 0.20 \end{bmatrix} \begin{bmatrix} 1.33 & 0 & 0 \\ 0 & 0.96 & 0 \\ 0 & 0 & 0.55 \end{bmatrix} = \begin{bmatrix} 0.85 & 0.37 & -0.37 \\ 0.91 & 0.10 & 0.40 \\ -0.45 & 0.88 & 0.11 \end{bmatrix}$$

So, for example, 0.91 is the correlation between the second variable and the first principal component.

Compute the communality

Next compute the communality, using the first two eigenvalues only.

$$\mathbf{SS}' = \begin{bmatrix} 0.85 & 0.37 \\ 0.91 & 0.09 \\ -0.45 & 0.88 \end{bmatrix} \begin{bmatrix} 0.85 & 0.91 & -0.45 \\ 0.37 & 0.09 & 0.88 \end{bmatrix} = \begin{bmatrix} 0.8662 & 0.8140 & -0.0606 \\ 0.8140 & 0.8420 & -0.3321 \\ -0.0606 & -0.3321 & 0.9876 \end{bmatrix}$$

Diagonal elements report how much of the variability is explained

Communality consists of the diagonal elements.

var	
1	0.8662
2	0.8420
3	0.9876

This means that the first two principal components "explain" 86.62 % of the first variable, 84.20 % of the second variable, and 98.76 % of the third.

Compute the coefficient matrix

The coefficient matrix, $\mathbf{B}$, is formed using the reciprocals of the diagonals of $\mathbf{L}^{1/2}$.

$$\mathbf{B} = \mathbf{VL}^{-1/2} = \begin{bmatrix} 0.48 & 0.40 & -1.20 \\ 0.52 & 0.10 & 1.31 \\ -0.26 & 0.95 & 0.37 \end{bmatrix}$$

Compute the principal factors

Finally, we can compute the factor scores from **ZB**, where **Z** is **X** converted to standard score form. These columns are the *principal factors*.

$$\mathbf{F} = \mathbf{ZB} = \begin{bmatrix} 0.41 & -0.69 & 0.06 \\ -2.11 & 0.07 & 0.63 \\ -0.46 & -0.32 & 0.30 \\ 1.62 & -1.00 & 0.70 \\ 0.70 & 1.09 & 0.65 \\ -0.86 & 1.32 & -0.85 \\ -0.60 & -1.31 & 0.86 \\ 0.94 & 1.72 & -0.04 \\ 0.22 & 0.03 & 0.34 \\ 0.15 & -0.91 & -2.65 \end{bmatrix}$$

Principal factors control chart

These factors can be plotted against the indices, which could be times. If time is used, *the resulting plot is an example of a principal factors control chart.*