Análisis de Factores

Resumen

El procedimiento **Análisis de Factores** esta diseñado para extraer factores comunes de un conjunto de *p* variables cuantitativas *X*. En muchas situaciones, un número pequeño de factores común pueden representar un gran porcentaje de la variabilidad de las variables originales. La habilidad para expresar la covarianza entre las variables en términos de un número pequeño de factores significativos puede ser de gran ayuda para profundizar en los datos que son analizados.

El procedimiento realiza ambos: componentes principales y análisis de factor clásico. Las cargas de Factor pueden ser extraídas de la matriz de covarianzas muéstrales o de la de correlaciones maestrales. Las cargas iniciales son rotadas usando varimax, equimax, o rotación quartimax.

Ejemplo StatFolio: *factor analysis.sgp*

Datos del Ejemplo:

El archivo *93cars.sf6* contiene información acerca de 26 variables para *n* = 93 marcas y modelos de automóviles, tomadas de Lock (1993). La siguiente tabla muestra una lista parcial de los datos de este archivo:

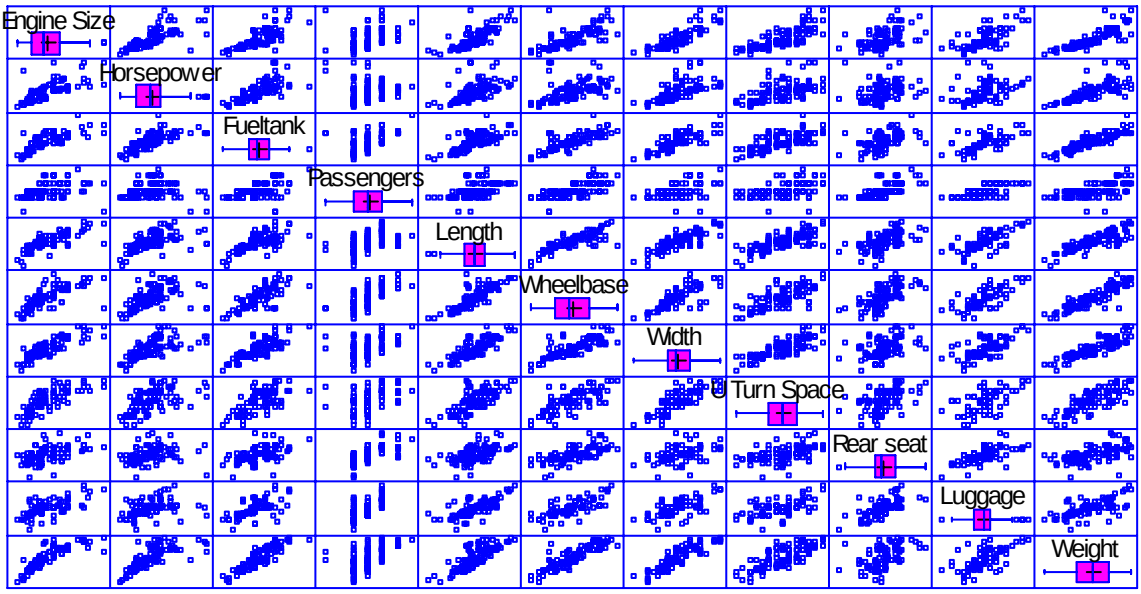
Make (Marca)	*Model (Modelo)*	*Engine Size (Tamaño del motor)*	*Horsepower (Caballos de Fuerza)*	*Fuel Tank (Depósito de gasolina)*	*Passengers (Pasajeros)*	*Length (Longitud)*
Acura	Integra	1.8	140	13.2	5	177
Acura	Legend	3.2	200	18	5	195
Audi	90	2.8	172	16.9	5	180
Audi	100	2.8	172	21.1	6	193
BMW	535i	3.5	208	21.1	4	186
Buick	Century	2.2	110	16.4	6	189
Buick	LeSabre	3.8	170	18	6	200
Buick	Roadmaster	5.7	180	23	6	216
Buick	Riviera	3.8	170	18.8	5	198
Cadillac	DeVille	4.9	200	18	6	206
Cadillac	Seville	4.6	295	20	5	204
Chevrolet	Cavalier	2.2	110	15.2	5	182

Se desea realizar un análisis de factor para las siguientes variables:

Engine Size
Horsepower
Fueltank
Passengers
Length
Wheelbase
Width
U Turn Space
Rear seat

Luggage
Weight

Una grafica matriz de los datos se muestra abajo:



Como es de esperarse, las variables son altamente correlacionadas ya que muchos son relacionados al tamaño del vehículo.

Entrada de Datos

La caja de dialogo de entrada requiere los nombres de las columnas que contiene los datos:

- **Datos:** Las observaciones originales o la matriz de covarianzas muestral $\hat{\Sigma}$. Si se introducen las observaciones originales, introduce *p* columnas numéricas que contengan *n* valores para cada columna de X. Si se introduce la matriz de covarianzas muestral, introduce *p* columnas numéricas que contengan los *p* valores para cada columna de $\hat{\Sigma}$. Si la matriz de covarianzas es introducida, algunas de las tablas y graficas no estarán disponibles.

- **Etiquetas de Puntos**: Etiquetas opcionales para cada observación.

- **Selección:** Selección de un subconjunto de los datos.

Modelo Estadístico

El objetivo del análisis de factor es caracterizar las *p* variables en *X* en términos de un numero pequeño de *m factores comunes F*, los cuales impactan a todas las variables, y un conjunto de *errores o factores específicos ε*, los cuales afectan solo a la variable *X*. Siguiendo Johnson y Wichern (2002), el modelo ortogonal de factor común expresa las variables observadas como

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \varepsilon_2 \\ &\cdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \varepsilon_p \end{aligned} \qquad (1)$$

En notación matricial,

$$X - \mu = LF + \varepsilon \qquad (2)$$

donde μ es un vector de medias y *L* es llamada matriz de *cargas de factores*. Se asume que los factores comunes y los factores específicos son independientes unos de todos. Para evitar ambigüedad en el escalamiento, las varianzas de los factores comunes se asumen iguales a 1, mientras que la matriz de covarianzas de los factores específicos Ψ es una matriz diagonal con elementos diagonales Ψ_j. La matriz de covarianza Σ de las observaciones originales X esta relacionada a la matriz de cargas de factores por

$$\Sigma = LL' + \Psi \qquad (3)$$

Un resultado importante del modelo anterior es la relación entre las varianzas de las variables originales *X* y las varianzas de los factores deseados. En particular,

$$Var(X_j) = l_{j1}^2 + l_{j2}^2 + ... + l_{jm}^2 + \Psi_j \qquad (4)$$

Esta varianza es expresada como la suma de las dos cantidades:

1. La *comunidad*: $l_{j1}^2 + l_{j2}^2 + ... + l_{jm}^2$
2. La *varianza específica*: Ψ_j

La comunidad es la varianza atribuida a los factores que todas las variables *X* tienen en común, mientras que la varianza especifica es especifica a un solo factor.

Se debería notar que las cargas de factores *L* no son únicas. Multiplicación por alguna matriz ortogonal permite otros conjuntos aceptables de cargas de factores. Seguida a la extracción del factor inicial, es común rotar las cargas de factores hasta que ellas no pueden ser fácilmente interpretadas.

Resumen del Análisis

La tabla del *Resumen del Análisis* se muestra abajo:

Análisis de Factor

Datos/Variables:
- Engine Size (liters)
- Horsepower (maximum)
- Fueltank (gallons)
- Passengers (persons)
- Length (inches)
- Wheelbase (inches)
- Width (inches)
- U Turn Space (feet)
- Rear seat (inches)
- Luggage (cu. ft.)
- Weight (pounds)

Entrada de datos: observaciones
Número de casos completos: 82
Tratamiento de valores perdidos: eliminación listwise
Estandarizar: sí

Tipo de Factorización: componentes principales
Número de factores extraídos: 2

Análisis de Factores

Factor		*Porcentaje de*	*Porcentaje*
Número	*Eigenvalor*	*Varianza*	*Acumulado*
1	7.92395	72.036	72.036
2	1.32354	12.032	84.068
3	0.47071	4.279	88.347
4	0.353248	3.211	91.559
5	0.269048	2.446	94.004
6	0.190242	1.729	95.734
7	0.172892	1.572	97.306
8	0.107148	0.974	98.280
9	0.0824071	0.749	99.029
10	0.0694689	0.632	99.660
11	0.0373497	0.340	100.000

	Inicial
Variable	*Comunalidad*
Engine Size	1.0
Horsepower	1.0
Fueltank	1.0
Passengers	1.0
Length	1.0
Wheelbase	1.0
Width	1.0
U Turn Space	1.0
Rear seat	1.0
Luggage	1.0
Weight	1.0

Desplegados en la tabla están:

- **Variables de Datos**: Los nombres de las *p* columnas de entrada.

- **Entrada de Datos**: Cualquier observación *o matriz*, dependen de si los datos contienen las observaciones originales o la matriz de covarianza muestral.

- **Numero de Casos Completos**: El numero de casos *n* para los cuales ninguna de las observaciones es perdida.

- **Tratamiento de Valores Perdidos**: Como los valores perdidos son tratados en la estimación de la matriz de covarianza o correlación. Si selecciona *Lista Completa*, los estimadores serán basados solo en los casos completos. Si selecciona L*ista Par*, todos los pares de datos no perdidos serán usados para obtener los estimadores.

- **Estandarización**: *Si,* si el análisis se basa en la matriz de correlación. *No,* si se basa en la matriz de covarianza.

- **Tipo de Factorización**: Es C*omponentes Principales,* si la extracción de los factores fue hecha directamente en la matriz de covarianza o correlación muestral, o C*lásico,* si los elementos de la diagonal fueron ajustados usando estimadores de las comunidades.

- **Numero de Componentes Extraídos**: El numero de componentes *m* extraídos de los datos. Este numero es basado en la configuración sobre la caja de dialogo *Opciones del Análisis*.

Una tabla también será desplegada para mostrar información de cada uno de *p* posibles factores:

- **Numero de Factor**: El numero de factor *j*, de1 hasta *p*.

- **Eigenvalor**: El eigenvalor de la matriz de covarianza o correlación estimada, $\hat{\lambda}_j$, después de ajustar las comunidades estimadas, si se usa el método *clásico*.

- **Porcentaje de Varianza**: El porcentaje total de la varianza estimada representada por este factor, es igual a

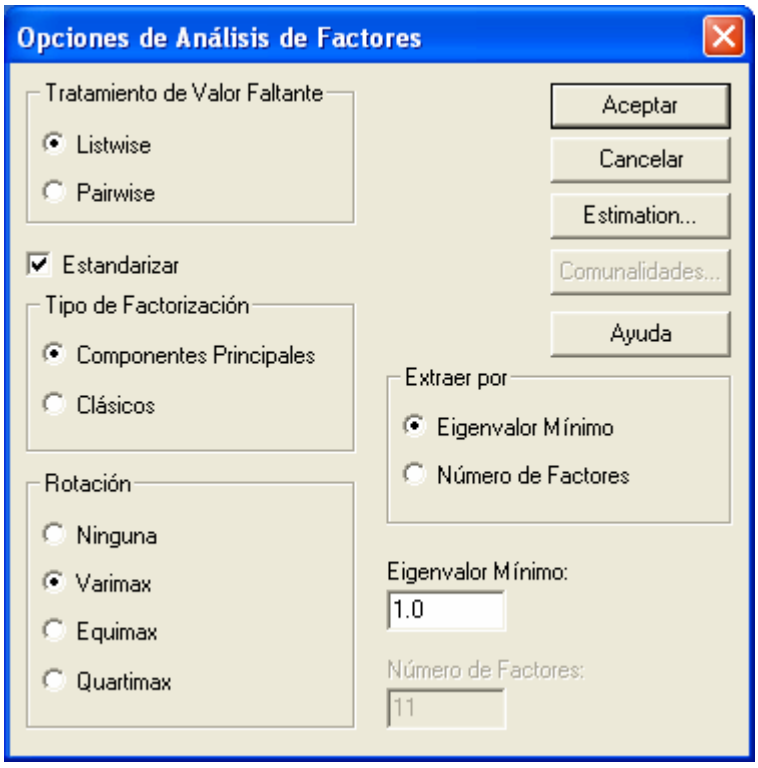
$$100\left(\frac{\hat{\lambda}_j}{\hat{\lambda}_1+\hat{\lambda}_2+\ldots+\hat{\lambda}_m}\right)\% \qquad (5)$$

- **Porcentaje Acumulado**: El porcentaje acumulado del total de la varianza estimada en la población acumulado por los primeros *j* factores.

- **Comunidad Inicial**: La comunidad inicial utilizada en los cálculos, ya sea entrada por el usuario o estimada de las covarianzas o correlaciones muéstrales.

En el ejemplo, los primeros *m* = 2 factores acumulan mas del 84% de toda la varianza entre las 11 variables.

Opciones del Análisis



- **Tratamiento de Valores Perdidos**: Método para manejar valores perdidos cuando se estima las covarianzas o las correlaciones muéstrales. Especifique *Lista Completa* para usar solo los casos que no tiene valores perdidos para cualquier entrada de variables. Especifique *Lista Par* para usar todos los pares de observaciones en los cuales ningún valor fue perdido.

- **Estandarizar**: Activar esta caja para basar el análisis en la matriz de correlación en lugar de covarianza. Esto corresponde a estandarizar cada variable de entrada antes de calcular las variables, sustrayendo su media y dividiendo entre la desviación estándar.

- **Tipo de Factorización**: Seleccione *Componentes Principales* para extraer los factores directamente de la matriz de covarianza o correlación. Seleccione *clásico* para remplazar los elementos de la diagonal con las comunidades estimadas. Si se usa el método *clásico*, se pueden especificar las comunidades presionando el botón *Comunidad* o permitir que el programa use un método iterativo para estimarlos.

- **Rotación**: El método usado para rotar la matriz de cargas de los factores después de que estos han sido extraídos. La rotación *Varimax* maximiza la varianza de las cargas cuadradas en cada columna. *Quartimax* maximiza la varianza de las cargas cuadradas en cada fila. *Equimax* intenta alcanzar un balance entre filas y columnas.

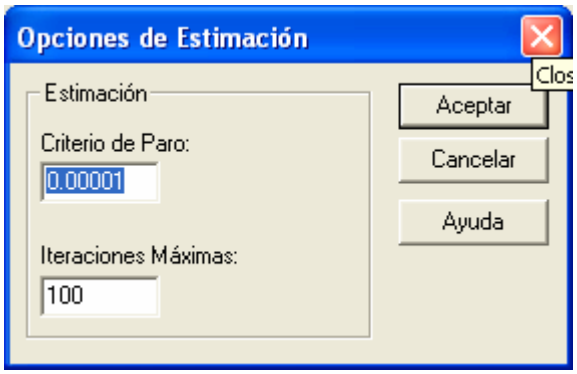
- **Extraídos Por**: El criterio usado para determinar el número de factores a extraer.

- **Eigenvalor Mínimo**: Si extraemos por la magnitud de los eigenvalores, el eigenvalor mínimo con el cual el factor será extraído.

- **Numero de Factores**: Si extraemos por numero de factores, el numero *k*.

Existen también dos botones que acceden a cajas de dialogo adicionales:

Botón de Estimación

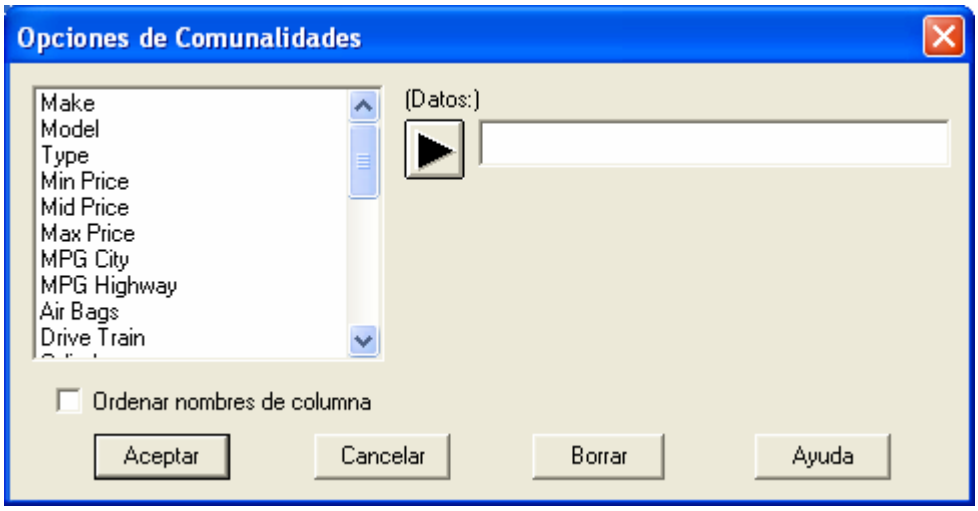


Estos campos controlan las iteraciones utilizadas en:

1. El método C*lásico* de la extracción de factor. Las comunidades estimadas son revisadas hasta que el cambio proporcional en su suma es menor que el C*riterio de Paro*, o que el M*áximo de Iteraciones* es pasado.

2. *Rotación* de las cargas del factor. El criterio de paro aplica a la varianza de los elementos cuadrados en la diagonal de la matriz de cargas de los factores.

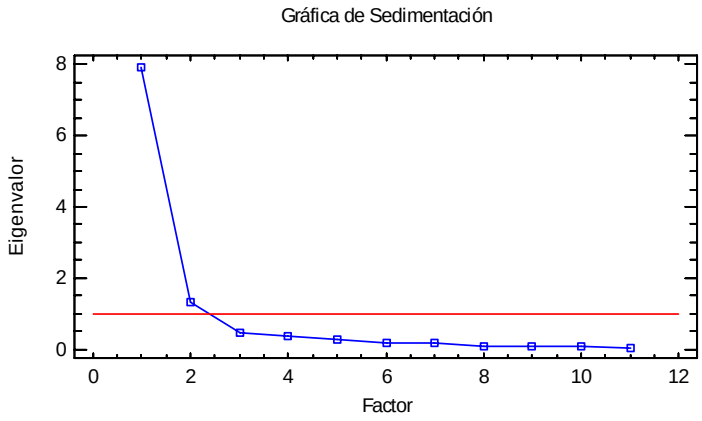
Botón Comunidades



Cuando se usa el método de estimación C*lásico*, se puede especificar una columna que contenga las comunidades en lugar de que el programa las estime por iteración.

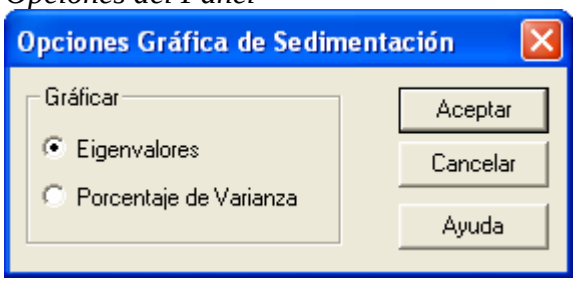
Grafico Scree

El *Grafico Scree* puede ser de gran ayudar para determinar el número de factores a extraer. Por defecto, grafica el tamaño de los eigenvalores correspondientes a cada uno de los posibles *p* factores:



Una línea adicional es agregada en el mínimo valor especificado sobre la caja de dialogo *Opciones del Análisis*. En la grafica anterior, observe que solo los primeros 2 factores tienen eigenvalores grandes.

Opciones del Panel



- **Graficar**: Valor graficado en el eje vertical.

Estadísticas de Extracción

El panel *Estadísticas de Extracción* muestra el valor estimado de los coeficientes *l* para cada factor extraído, antes de que cualquier rotación sea aplicada:

Matriz de Cargas Antes de Rotar

	Factor	*Factor*
	1	*2*
Engine Size	0.936606	-0.154035
Horsepower	0.754754	-0.50948
Fueltank	0.876138	-0.241737
Passengers	0.671882	0.610074
Length	0.944075	0.0244126
Wheelbase	0.944096	0.0702147
Width	0.914567	-0.154446
U Turn Space	0.842284	-0.0955416
Rear seat	0.650975	0.613778
Luggage	0.778316	0.371338
Weight	0.948687	-0.237682

	Estimado	*Específico*
Variable	*Comunalidad*	*Varianza*
Engine Size	0.900958	0.0990419
Horsepower	0.829223	0.170777
Fueltank	0.826054	0.173946
Passengers	0.823616	0.176384
Length	0.891874	0.108126
Wheelbase	0.896247	0.103753
Width	0.860287	0.139713
U Turn Space	0.71857	0.28143
Rear seat	0.800491	0.199509
Luggage	0.743667	0.256333
Weight	0.9565	0.0435005

También se despliegan las comunidades y las varianzas específicas. Las ponderaciones dentro de cada columna frecuentemente tienen interpretaciones interesantes. En el ejemplo, observe que las ponderaciones en la primera columna son todas aproximadamente iguales. Esto implica que el primer componente es básicamente un promedio de todas las variables de entrada. El segundo componente es ponderado mas pesadamente en una dirección positiva en el número de *Passengers*, el sitio *Rear Seat*, y la cantidad de espacio *Luggage*, y en una dirección negativa a *Horsepower*. Esto parece diferenciar entre los distintos tipos de vehículos. Note también que *U Turn Space* y *Luggage* tienen una varianza específica más grande que los demás, implicando que ellos no son muy bien tomados en cuenta por los dos factores extraídos.

Estadísticas de Rotación

El panel de *Estadísticas de Rotación* muestra los valores estimados de los coeficientes *l* después de que la rotación requerida fue aplicada:

Matriz de Cargas del Factor Después Varimax Rotación

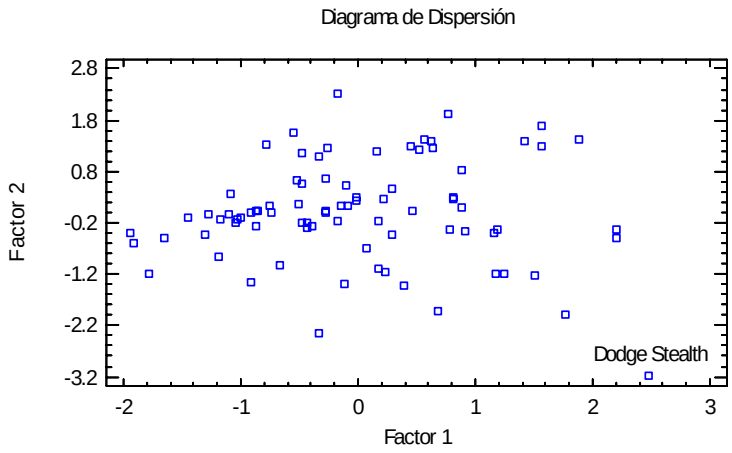
	Factor	*Factor*
	1	*2*
Engine Size	0.859769	0.402188
Horsepower	0.910596	0.00617243
Fueltank	0.859441	0.295661
Passengers	0.209571	0.883004
Length	0.765091	0.553632
Wheelbase	0.739226	0.591432
Width	0.841818	0.389395
U Turn Space	0.748896	0.397145
Rear seat	0.190229	0.874245
Luggage	0.43229	0.746186
Weight	0.917004	0.340004

	Estimado	*Específico*
Variable	*Comunalidad*	*Varianza*
Engine Size	0.900958	0.0990419
Horsepower	0.829223	0.170777
Fueltank	0.826054	0.173946
Passengers	0.823616	0.176384
Length	0.891874	0.108126
Wheelbase	0.896247	0.103753
Width	0.860287	0.139713
U Turn Space	0.71857	0.28143
Rear seat	0.800491	0.199509
Luggage	0.743667	0.256333
Weight	0.9565	0.0435005

Note que la rotación tiene decrecimiento substancial en las cargas de *Passengers*, *Rear seat*, y *Luggage* en el primer vector y los hace variables dominantes del segundo factor. El segundo factor parece distinguir familias de vehículos grandes tal como minivans y SUV's entre los otros automóviles.

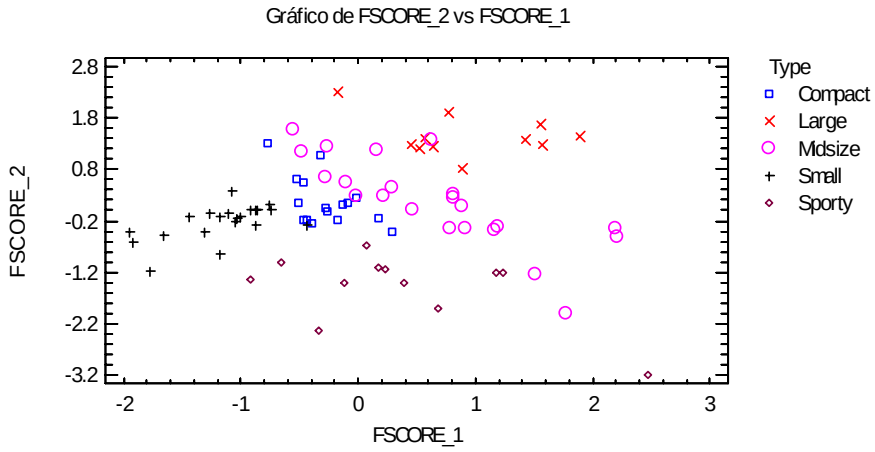
Gráficos de Dispersión 2D y 3D

Los *Gráficos de Dispersión 2D y 3D* muestran 2 o 3 factores seleccionados para cada uno de los *n* casos, después de la rotación.



Es usual examinar algunos puntos que están lejos de otros, tales como los resaltados *Dodge Stealth*, los cuales tiene un valor muy pequeño para el segundo factor.

Una variación interesante de este grafico es codificación de las variables de acuerdo a otra columna, tal como el tipo de vehiculo:

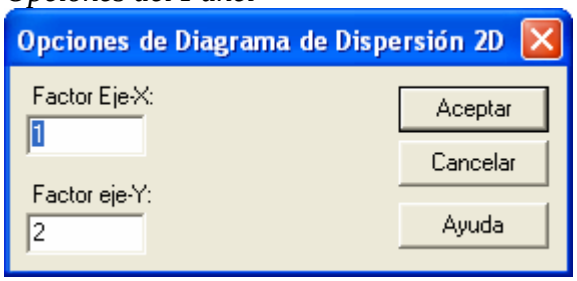


Para producir la grafica anterior:

1. Presionar el botón *Guardar Resultados* y grabar *Puntuación del Factor* en nuevas columnas sobre la hoja de datos.
2. Seleccionar el procedimiento *Grafico X-Y* en la parte superior del menú e introducir las nuevas columnas.
3. Seleccionar *Opciones del Análisis* y especificar *Tipo* en el campo *Puntos de Códigos*.

Ahora es claro que el primer factor esta relacionado al tamaño del vehículo, mientras que el segundo factor separa los carros deportivos de los demás.

Opciones del Panel



Especifique los factores a graficar en cada eje.

Puntuación del Factor

El panel *Puntuación del Factor* despliega las puntuaciones de los factores rotados para cada uno de *n* casos.

Tabla de Puntuación de Factores

		Factor	*Factor*
Fila	*Etiqueta*	*1*	*2*
1	Integra	-0.440603	-0.294691
2	Legend	0.817275	0.299261
3	90	0.177176	-0.154546
4	100	0.155524	1.17616
5	535i	1.5048	-1.23631
6	Century	-0.474803	1.14786
7	LeSabre	0.63412	1.25438
8	Roadmaster	1.88652	1.43271
9	Riviera	1.18707	-0.321997
…	…	…	…

Las puntuaciones del factor muestran donde cae cada observación con respecto a los factores extraídos.

Coeficientes del Factor

La tabla de *Coeficientes del Factor* muestra los coeficientes usados para crear los valores de los factores en las variables originales.

Coeficientes del Factor

	Factor	*Factor*
	1	*2*
Engine Size	0.163284	0.29611
Horsepower	-0.0292234	-0.263759
Fueltank	19.7073	14.343
Passengers	4.48584	10.7923
Length	39.5473	46.3067
Wheelbase	8.18626	26.1997
Width	-59.5975	-13.7139
U Turn Space	3.83938	15.7456
Rear seat	-17.5316	-16.1991
Luggage	-6.00197	19.3445
Weight	-114.779	-181.049

Si la matriz de covarianza muestral *S* ha sido factorizada, entonces los coeficientes son los términos cargas multiplicadas por la desviación de cada variable con respecto a su media en

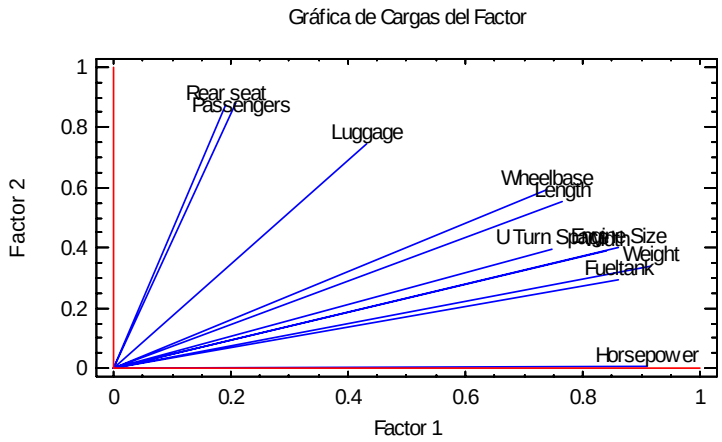
$$\hat{f}_j = \hat{L}' S^{-1}(x_j - \bar{x}) \qquad (6)$$

Si la matriz de correlación muestral *R* ha sido factorizada, entonces los coeficientes son los términos cargas multiplicadas por los valores estandarizados de cada variable de acuerdo a

$$\hat{f}_j = \hat{L}' R^{-1} z_j \qquad (7)$$

Gráfico Factor 2D y 3D

El *Gráfico Factor* muestra la localización de cada variable en el espacio de 2 o 3 factores seleccionados:



Las variables más lejos sobre de la línea de referencia en 0 provoca la contribución mas grande de los factores.

Grabar Resultados

Los siguientes resultados pueden ser guardados en una hoja de datos:

1. *Eigenvalores* – Los *m* eigenvalores.

2. *Matriz de Factores* – Las *m* matrices, cada una contiene *p* estimadores de los coeficientes *l* antes de la rotación.

3. *Matriz de Factores Rotados* – Las *m* columnas, cada una contiene *p* estimadores de los coeficientes *l* después de la rotación.

4. *Matriz de Transición* – La matriz *m* por *m* que multiplica las cargas de los factores originales para calcular las cargas de los factores rotados.

5. *Comunalidades* – Las *p* comunidades estimadas después de la rotación.

6. *Varianzas Específicas* – Las *p* varianzas especificas después de la rotación.

7. *Puntuación de Factores* – Las *m* columnas, conteniendo cada uno de *n* valores correspondientes a los factores extraídos.

8. *Coeficientes de Calificación de Factor* – Las *m* columnas, cada una conteniendo los *p* valores de los coeficientes del factor.

