Facultad de Ciencias
UNAM

Seminario de Estadística

Tarea 2B

Selección de variables

Enríquez Hernández Leobardo

21 de mayo de 2024

Índice

1. Monte Carlo y Bootstrap no paramétrico

Sea una muestra aleatoria $X_1, ..., X_n$ de una población con distribución $Poisson(\theta)$. Se puede mostrar que la estimación de la función parametral de $\tau(\theta) = e^{-\theta} = P(X = 0)$ es $\hat{\tau}(\theta) = (\frac{n-1}{n})^{\sum_{i=1}^{n} X_i}$ y que es su UMVUE, sin embargo no es fácil encontrar la distribución de $\hat{\tau}(\theta)$ o la expresión de su varianza $V(\hat{\tau}(\theta))$.

a. Método Monte Carlo

Para estimar $E(\hat{\tau}(\theta))$, $V(\hat{\tau}(\theta))$ y el histograma de $\hat{\tau}_1, ..., \hat{\tau}_B$ como datos de la distribución de $\hat{\tau}(\theta)$, se generan diez mil muestras, cada muestra tiene 20 observaciones, de la variable aleatoria $\widehat{\tau} \sim \text{Poisson}(\theta = 1)$.

De este modo, al estimar $E(\widehat{\tau})$, $V(\widehat{\tau})$ y la distribución de $\hat{\tau}$ se obtienen los siguientes resultados (los códigos se pueden consultar en el archivo RMarkdown en los chunks *estamationT* y *histogram1* en las lineas 50 y 83 respectivamente).

$$\mathbb{E}\left[\widehat{\tau}\right] \approx \frac{\sum_{i=1}^{10000} \widehat{\tau}_i}{10000} \approx 0{,}3681426 \qquad y \qquad \mathbb{V}\left[\widehat{\tau}\right] = \mathbb{E}\left[\widehat{\tau}^2\right] - \mathbb{E}\left[\widehat{\tau}\right]^2 \approx 0{,}0069788$$

b. *Bootstrap* no paramétrico

Para el método de *bootstrap* no paramétrico, se generan 20 números aleatorios de una distribución Poisson($\theta = 1$). Hacemos la estimación de $\tau(\theta) = e^{-\theta} = P(X = 0)$ usando $\hat{\tau}(\theta) = (\frac{n-1}{n})^{\sum_{i=1}^{n} X_i}$, estimamos la esperanza y varianza de $\hat{\tau}$ usando bootstrap no paramétrico con $B = 10,000$, y el histograma de $\hat{\tau}^*_{(1)}, ..., \hat{\tau}^*_{(n)}$.

Se obtuvieron los siguientes resultados (el código se puede consultar en el chunk *Bootstrap* en la línea 103 y 139 del archivo RMarkdown).

$$\mathbb{E}\left[\widehat{\tau}\right] \approx 0{,}3073569 \qquad \mathbb{V}\left[\widehat{\tau}\right] \approx 0{,}0037737$$

Los métodos difirieron en aproximadamente 0.060786 para la esperanza del estimador y 0.003205 para su varianza. Los histogramas representan distribuciones muy parecidas.

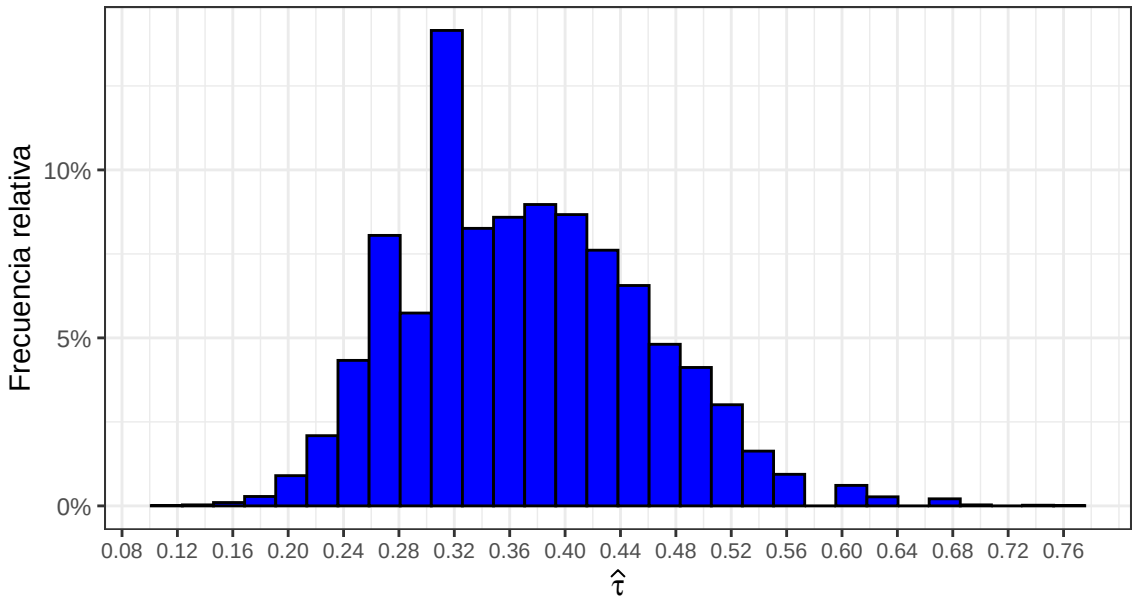


Figura 1: Histograma para las muestras generadas por Monte Carlo

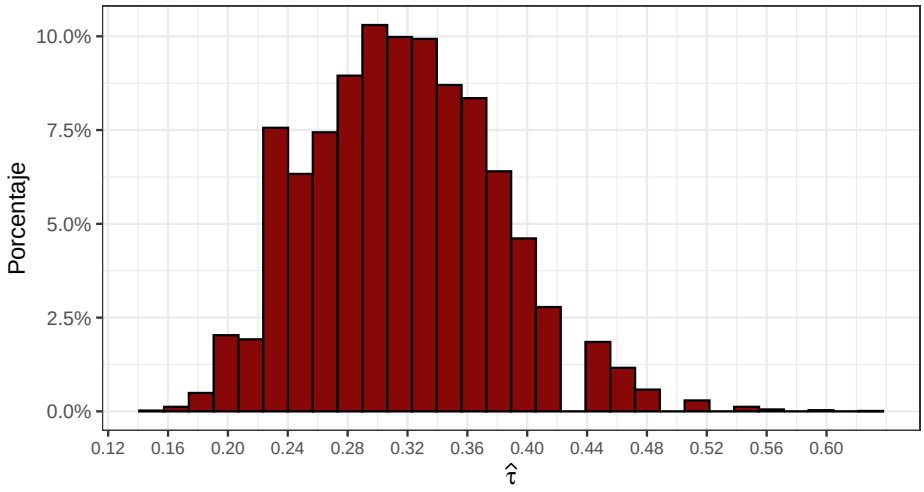
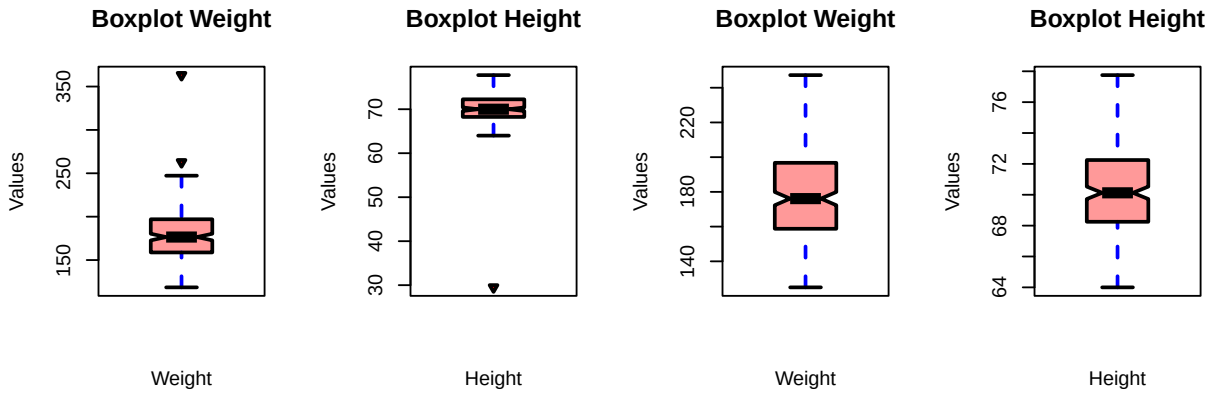


Figura 2: Histograma para las muestras generadas por *bootstrapping*

2. Selección de variables.

Nos interesa es usar las variables clínicas observadas en pacientes de la base de datos `fat` del paquete `faraway` para estudiar cuales son los factores que ayudan a modelar mejor el promedio del porcentaje de grasa corporal en Hombres (brozek). Omitiremos las variables siri, density y free, se eliminaron los valores nulos de la variable brozek, y los outliers de weight y height. Esto último se puede apreciar en la siguiente Gráfica.



Con el pre procesamiento realizado lo que sigue es crear subconjuntos de modelos para datos continuos con liga identidad y distribución Gaussiana, además de hacer selección de variables considerando efectos principales usando el mejor subconjunto, un método stepwise y lasso, con el criterio BIC para el mejor modelo. Además se sonsiderarán los subconjuntos con interacciones, términos cuadráticos para las variables, etc.

```
## (Intercept)      height       abdom       wrist
##   8.5938000  -0.4184531   0.7231168  -1.4836737
```

```
## [1] "BIC: 1412.141592"
```

En un primer subconjunto de ajuste (`fitBestSubset`) con la función `regsubsets` se hizo una selección de las mejores combinaciones de variables de las 14, el mejor resultado fue la combinación de tres variables height, abdom, wrist, con las cuales se obtuvo un menor BIC de 1412.141592. (Chunk fitBestSubset, linea 150)

```
##  (Intercept)       weight        abdom        wrist
## -25.32118026  -0.08768903   0.89001276  -1.22551890
```

```
## [1] "BIC: 1412.254566"
```

En el segundo subconjunto `modeloforward` con el ajuste del modelo stepwise(forward) se obtuvo un BIC de 1412.2545657el cual es muy parecido pero ligeramente mayor al obtenido con el del primer ajuste realizado con la seleccion de variables.(Chunk modeloforward, linea 198)

```
##  (Intercept)          age        abdom        wrist
## -10.86151335   0.07109929   0.71409295  -2.16060130
```

```
## [1] "BIC: 1415.872346"
```

En el tercer subconjunto `modelobackward` con el ajuste Backward obtuvimos un BIC de 1415.8723455 el cual comparado con los dos anteriores BIC resulta mas alto.(Chunk modelobackward, linea 212)

```
## [1] 77
```

```
## (Intercept)          age       height        abdom        wrist
##  4.57171015  0.04702087 -0.32711867  0.71391555 -1.68478028
```

```
## [1] "BIC: 1413.106619"
```

El cuarto subconjunto de modelos `AjusteModeloLasso`, corresponde al modelo lasso, donde se obtuvo un BIC de 1413.1066192. (Chunk AjusteModeloLasso, linea 235)

Con los métodos anteriormente realizados obtuvimos BIC muy similares entre si por lo que escoger uno como mejor modelo seria usar el mas parsimonioso, es decir, que resulte fácil de construirse y de interpretarse.

Ahora ajustaremos modelos parecidos a los anteriormente realizados con la diferencia de que incluiremos **interacciones** para ver si mejoran los modelos.

```
##   (Intercept)         abdom  height:wrist     chest:hip
## -24.906984524   0.873134763  -0.018543532  -0.001293605
```

```
## [1] "BIC: 1405.595944"
```

Para el quinto subconjunto `Ajusteforward2`, el resultado del forward con interacciones muestra un BIC de 1405.5959444.(Chunk Ajusteforward2, linea 292)

```
##   (Intercept)           hip    height:hip    neck:abdom      neck:hip
## -44.092260115   1.141649113  -0.004176813   0.020750934  -0.024928240
```

```
## [1] "BIC: 1416.310922"
```

Para el sexto subconjunto `Ajustebackward2`, el resultado del backward con interacciones muestra un BIC de 1416.3109222. (Chunk Ajustebackward2, linea 318)

```
##    (Intercept)          abdom      abdom:age      age:thigh   height:wrist
## -20.6020768954   0.7065839712  -0.0005744683   0.0018443405  -0.0220104071
```

```
## [1] "BIC: 1411.984695"
```

Para el séptimo subconjunto `AjusteLassoInteracciones`, con los nuevos cambios en el modelo lasso con interacciones obtuvimos un BIC de 1411.9846953. (Chunk AjusteLassoInteracciones, linea 345)

Con las interacciones notamos una pequeña mejoría del BIC.

Ahora, probaremos con distintas funciones ligas (identidad, log) en combinación con el modelo Gama con el fin de ver si con esto logramos mejorar el puntaje de BIC obtenidos hasta este momento.

```
## [1] 7
```

```
##   (Intercept)           hip    hip:height    neck:abdom      hip:neck
## -0.5645385544  0.0598604549 -0.0001725943  0.0011757588 -0.0014368866
```

```
## [1] "BIC: 1490.059783"
```

Para el octavo subconjunto de modelos `GamaLigasBackForLasso`, el mejor modelo considerando el modelo Gama con distintas ligas (identidad, log) y distintos métodos tales como backward, forward y lasso, es el que tiene un BIC de 1490.0597827, el cual es una Gama con liga log. (Chunk GamaLigasBackForLasso, linea 396)

```
##   (Intercept)        height   I(height^2)         abdom    I(abdom^2)
## -34.506777195   1.051899470  -0.010616136   1.429932998  -0.003731687
##         wrist    I(wrist^2)
##  -5.892918565   0.118790824
```

```
## [1] "BIC: 1423.088935"
```

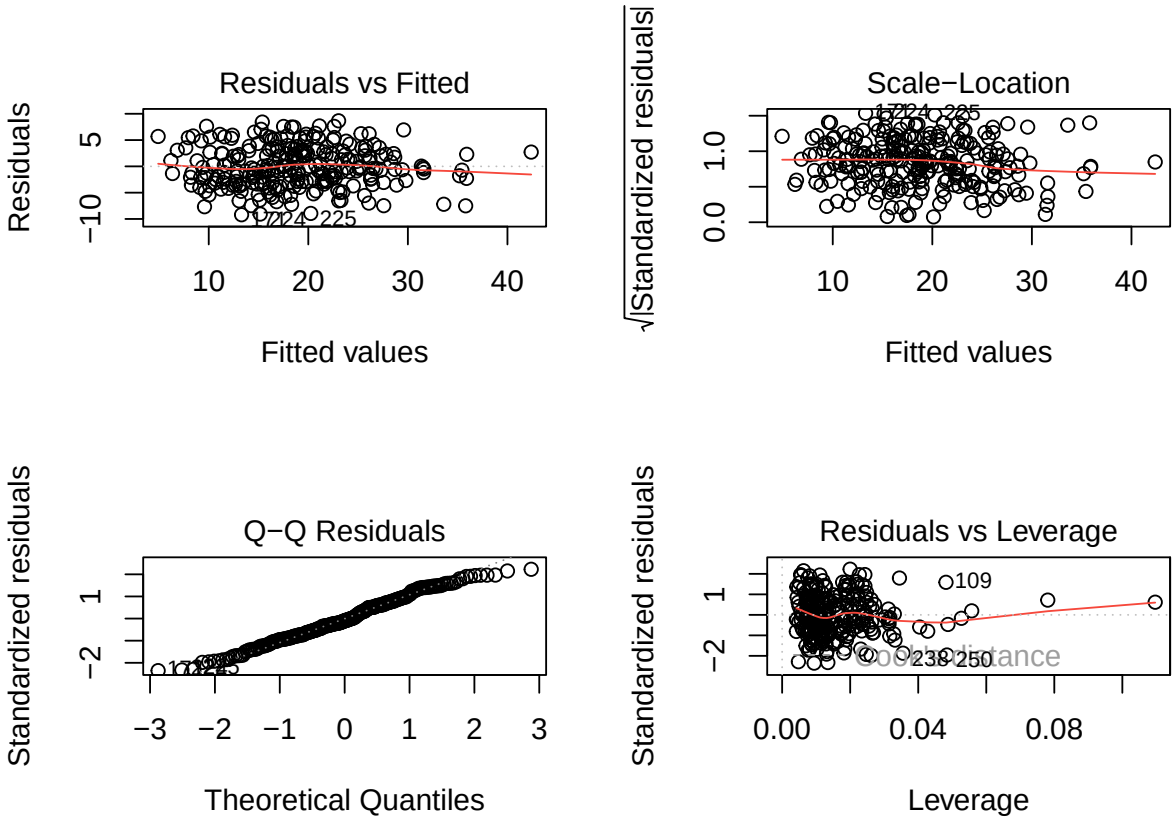
Por último, en el subconjunto noveno de modelos `ajusteCuadraticosubset`, usando una versión extendida que integra el cuadrado de las variables, se tiene un BIC de 1423.0889354 como el mejor. (Chunk ajusteCuadraticosubset, linea 463)

Presentamos a contunuación un Cuadro con los mejores modelos obtenidos en cada subconjunto con su respectivo BIC. Es posible observar que el modelo con el menor BIC de 1405.596 es uno con interacciones, en donde se consideran las covariables `abdom`, y las interacciones de `height:wrist` y `chest:hip`. La variable presente en todos los modelos es `abdom`, seguido de `wrist` en 7 modelos.

No.	Método de selección	Covariables y coeficientes estimados	BIC
1	fitBestSubset	(Intercept), height, abdom, wrist 8.5938, -0.4184531, 0.7231168, -1.4836737	1412.142
2	modeloforward	(Intercept), weight, abdom, wrist -25.3211803, -0.087689, 0.8900128, -1.2255189	1412.255
3	modelobackward	(Intercept), age, abdom, wrist -10.8615133, 0.0710993, 0.714093, -2.1606013	1415.872
4	AjusteModeloLasso	(Intercept), age, height, abdom, wrist 4.5717101, 0.0470209, -0.3271187, 0.7139155, -1.6847803	1413.107
5	Ajusteforward2	(Intercept), abdom, height:wrist, chest:hip -24.9069845, 0.8731348, -0.0185435, -0.0012936	1405.596
6	Ajustebackward2	(Intercept), hip, height:hip, neck:abdom, neck:hip -44.0922601, 1.1416491, -0.0041768, 0.0207509, -0.0249282	1416.311
7	AjusteLassoInteracciones	(Intercept), abdom, abdom:age, age:thigh, height:wrist -20.6020769, 0.706584, -0.0005745, 0.0018443, -0.0220104	1411.985
8	GamaLigasBackForLasso	(Intercept), hip, hip:height, neck:abdom, hip:neck -0.5645386, 0.0598605, -0.0001726, 0.0011758, -0.0014369	1490.06
9	ajusteCuadraticosubset	(Intercept), height, I(height∧2), abdom, I(abdom∧2), wrist, I(wrist∧2) -34.5067, 1.0519, -0.0106, 1.4299, -0.0037, -5.8929, 0.1188	1423.089

Resultados de los métodos de selección

Para inferencia e interpretación de los coeficientes del modelo elegido, es necesario el cumplimiento de los supuestos. En la prueba gráfica de los supuestos, tales como la linealidad (Residuals vs Fitted), homocedasticidad (Scale-Location), normalidad (Q-Q Residuals) y presencia de outliers influyentes (Residuals vs Leverage), se observa que no hay problemas graves con los supuestos. (Chunk plotsmodelo, linea 566)



La linealidad se comprueba con la siguiente prueba. (Chunk linealidad, linea 586)

```
##        abdom height_wrist    chest_hip   Tukey test
##   0.26485820   0.71940474   0.08328658   0.50212569
```

De acuerdo con la prueba `studentized Breusch-Pagan` se tiene un p-value de 0.3265423 por lo que no se rechaza la hipótesis nula de homocesaticidad, por otra parte las pruebas de normalidad Jarque-Bera y Kolmogorov-Smirnov no rechazan la hipótesis nula de normalidad, con p-value de 0.1360153 y 0.1877934, respectivamente. (Chunk pruebasmodelo, linea 575)

Con esto, podemos argumentar que por cada unidad de incremento en la circunferencia del abdomen `abdom` en cm, el porcentaje de grasa corporal (brozek) incrementa en 0,8731348. Por otra parte, con el incremento en una unidad de la interacción estatura - curcunferencia de la muñeca (`height:wrist`) disminuye el porcentaje de grasa corporal en $-0{,}0185435$, y el incremento en una unidad de la interacción circunferencia del pecho - circunferencia de cadera disminuye el porcentaje de grasa corporal en $-0{,}0012936$.

3. Componentes principales y análisis factorial exploratorio

Se analiza la personalidad de de 228 estudiates de una universidad de los Estados Unidos a partir de una encuesta resumida en `Dat3Ex.csv`. Las respuestas de 1 "muy en desacuerdo", 2 "un poco en desacuerdo", 3 "ni de acuerdo ni en desacuerdo", 4 "un poco de acuerdo" y 5 "muy de acuerdo", para un grupo de 44 preguntas, de las cuales tomaremos 15: $V1, V2, V4, V6, V9, V12, V14, V16, V17, V26, V27, V29, V31, V34, V37$ (para mayor detalle ver el cuestionario). Los objetivos son los de obtener los componentes principales y hacer un análisis exploratorio factorial, para identificar dimensiones interesantes de los datos en su escala original y transformada.

Con la ayuda de la librería `factoextra` se obtuvieron los *Componentes Principales* con la función `prcomp` (ver Chunk factorCP en la línea 64).

Posteriormente se usa la función `fviz_eig` para el número de componentes a considerar según varianzas y en la siguiente Figura se muestran para los datos escalados y no escalados, se suguieren entre 4 o 5 componentes pues después de estos ya no hay mucho cambio en la varianza que aportan. Además se acumula en los tres casos un aproximado de 62 % a 63 % de la varianza total cuando consideramos 4 componentes (Chunk Grafica13, linea 79).

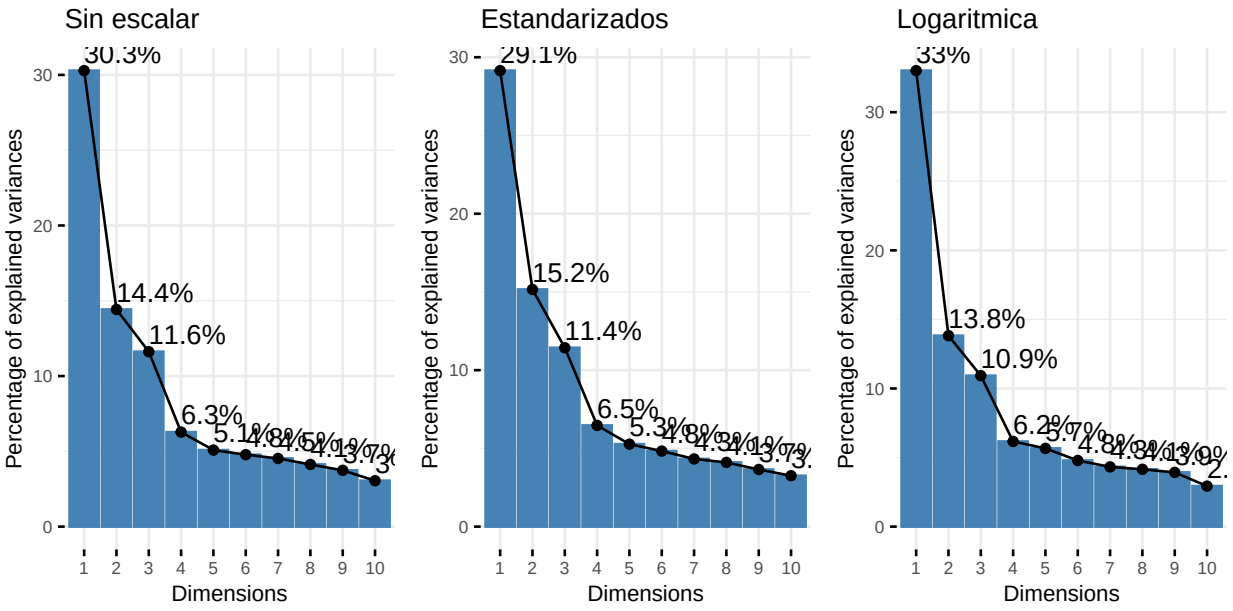


Figura 3: Índices para número de componentes principales

Analizamos las correlaciones de las primeras cuatro componentes con las variables originales (Chunk Correlation, linea 87). Se describen los siguientes resultados generales, considerando correlaciones mayores a 0.5 en valor absoluto, para dar una mayor comprensión y contexto de las variables y componentes principales. La siguiente descripción solamente se presenta para los valores originales, también se hace el ejercicio para datos estandarizados y en logaritmos, sin embargo los resultados son similares por lo que no se describen.

Para los datos sin escalar, las variables Deprimido, Tenso, Malhumorado y Grosero son las que tienen mayor asociación positiva en el componente 1, y por otro lado Relajado, Calmado y Entusiasta son las que tienen mayor asociación negativa con el componente 1. Las variables Parlanchin, Asertivo y Entusiasta son las de mayor asociación positiva para el componente 2, y Tímido y Reservado son las de mayor asociación negativa para el componente 2. Para el componente 3 las de mayor relación positiva son Relajado, Frío y Calmado, mientras que para la relación negativa con el componente 3 no hay valores mayores a 0.5 en valor absoluto. Y para el componente 4 no hay valores mayores a 0.5 en valor absoluto (sin embargo, mencionaremos que las de mayor relación positiva son Tímido, Indulgente y Entusiasta, mientras que las únicas con relación negativa son Frío, Peleonero y Victimista).

Para mayor interpretabilidad visual tenemos la siguiente Gráfica (Chunk Grafica23, línea 118), sólo se presentan los datos originales y los de escala logaritmica, las estandarizadas son iguales a las originales. Estas son las proyecciones de las variables de mayor peso en los primeros 2 componentes principales, rescatan la mayor varianza, podemos observar el sentido y magnitud de las fechas para visualizar la influencia de cada variables en cada componente.

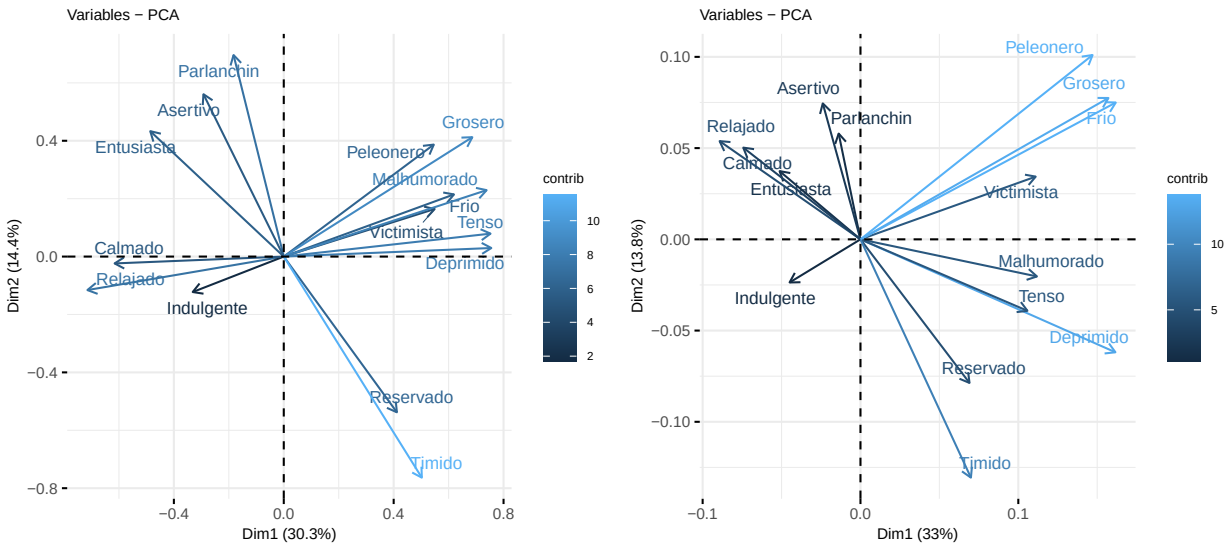
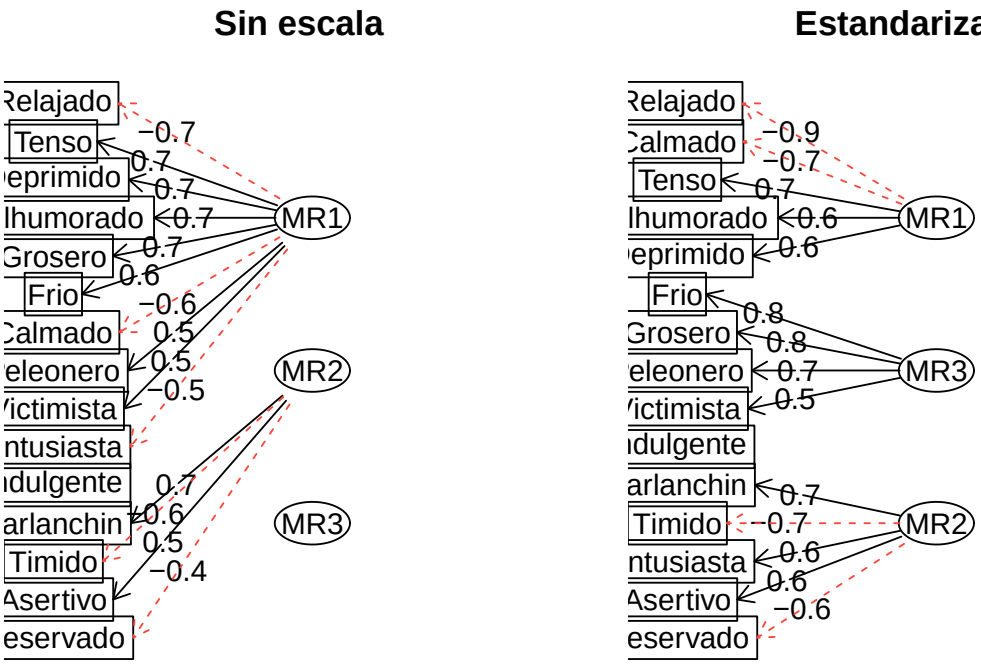


Figura 4: Proyeccion en componentes (izquierda: escala original; derecha:escala log)

Para continuar con el análisis consideramos el enfoque de *Análisis Factorial Exploratorio*, para ello nos apoyamos de la librería `psych` y la función `fa`. De nuevo consideramos datos sin escalar y estandarizados, optamos por considerar 3 factores, en los 2 casos Indulgente no queda en ninguno (Chunk Factorial, linea 131).



De las gráficas anteriores podemos notar, 3 componentes parecen ser suficiente para resumir la información, en contraste con componentes principales hemos reducido un poco más la dimensionalidad, además los resultados son muy similares a los componentes principales pues las variables de mayor peso se repiten casi todos los casos.

Para decidirnos por un modelo se probaron varias rotaciones como varimax y simplimax, también se consideraron a las variables como ordinales y de nuevo con ayuda de fa se obtuvieron las variables latentes mientras que con principal las componentes principales (ver Chunks RotacionesCP, RotacionesAFE y Ordinales; lineas 170, 203 y 228). Optamos por un modelo de Componente principales pues estos recuperan más varianza y dentro de estos el que usa la rotación "cluster" y maneja las variables como ordinales es el mejor rankeado pues recupera un 66 % de varianza total, además nos restringimos a considerar sólo 3 componentes pues el cuarto sólo está relacionado con una variable (Indulgente).

Ya con nuestro modelo seleccionado pasamos a la interpretación, según la Figura 5. El componente 1 corresponde a alumnos victimistas, fríos, groseros y peleoneros. En el componente 2, tenemos alumnos para los que ser asertivo, parlanchín y entusiasta se tiene un mayor relación positiva con el componente y ser tímidos y reservados los mayores valores negativos. Finalmente, en el componente 3 podemos notar mayores relaciones de alumnos deprimidos, malhumorados y tensos, mientras tenemos negativamente a alumnos calmados y relajados.

4. Análisis de conglomerados

El objetivo del analisis es identificar grupos de clientes para focalizar la publicidad de Oddjob Airways, a partir de una encuesta resumida en `Dat4ExB.csv`, y cuyas respuestas van de 1 a 100 (100 es que la persona considera que un aspecto es crucial en el servicio, mientras que 1 corresponde a que no lo es). Estos aspectos son puntualidad (e1), servicio según lo ofrecido (e2), experiencia placentera (e5), comodidad (e8), seguridad (e9), estado del avión (e10), comida adecuada (e16), hospitalidad (e17), viajar de forma sencilla (e21) y entretenimiento a bordo (e22). Como primer paso vamos a considerar que las variables son continuas, entonces dado ese supuesto obtendremos algunos grupos considerando el método k-means.

Aún cuando el indicador de `Average Silhouette width` y los indicadores de `Connectivity` y `Dunn` muestran que el número óptimo de clusters es de 2, el indicador de `Hubert statistic values` muestra que deben de

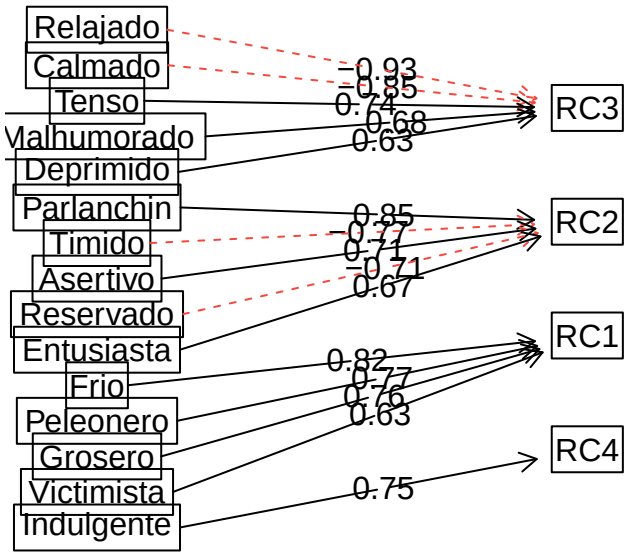
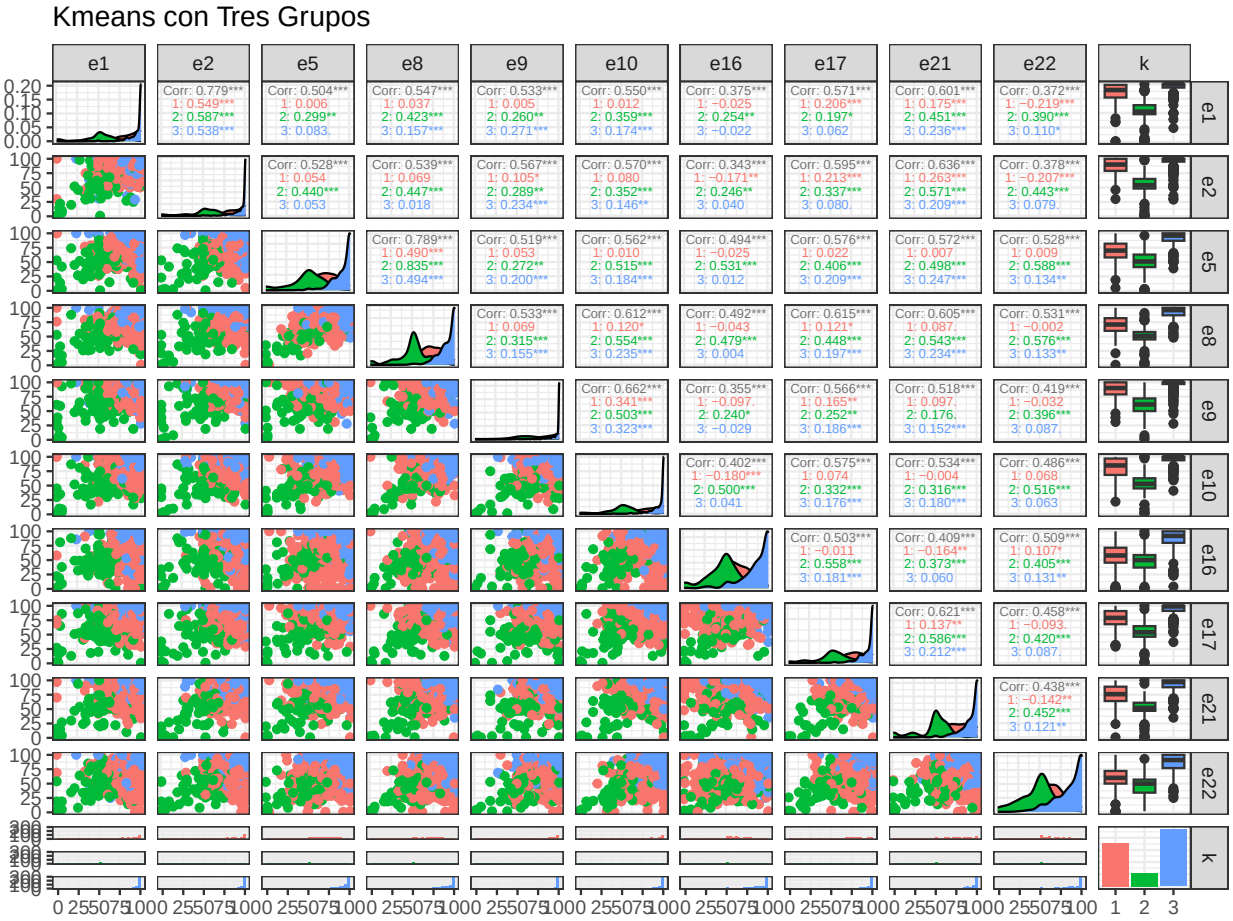


Figura 5: Componentes principales modelo seleccionado

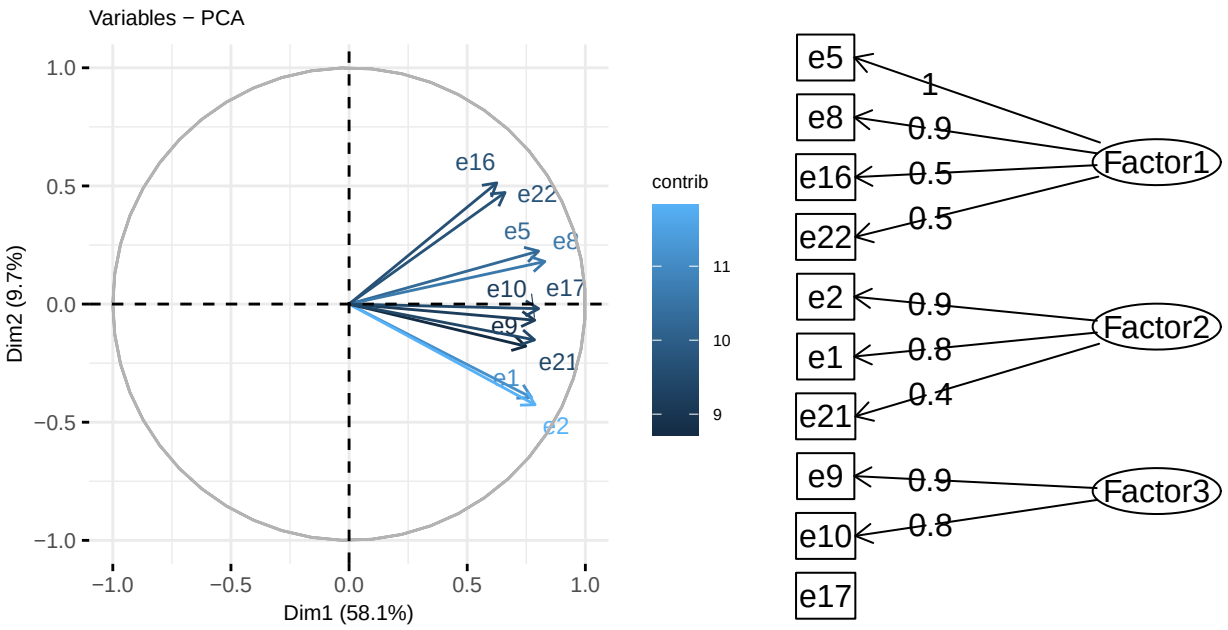
ser 3, y el indicador de `Dindex values` que deben de ser 5. Por lo que no hay un concenso indiscutible del número de clusters a considerar como óptimos. (Chunks clValid, fviz_nbclust_kmeans_silhouette y NbClust, lineas 66, 78 y 87).

Se decidió tomar al menos tres aspectos generales del servicio detectados en las variables: puntualidad y servicio según lo ofrecido; seguridad y estado del avión; y comodidad, experiencia, entretenimiento, hospitalidad y comida. Podemos focalizar la publicidad de la empresa en 3 grupos de clientes con base a estos tres aspectos.

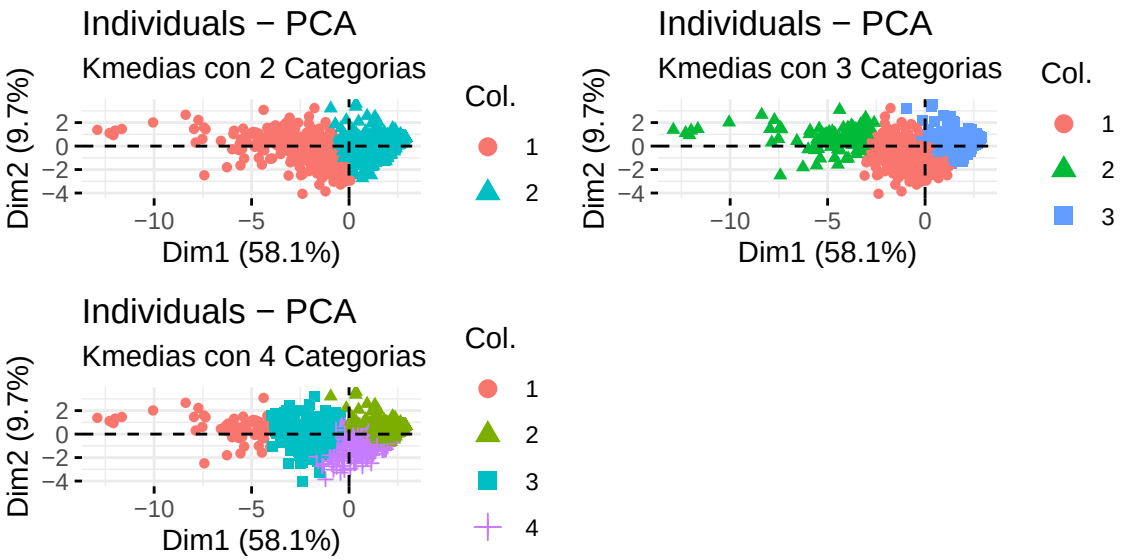
En la siguiente Gráfica podemos observar la asignación de clusters para tres grupos, y las correlaciones entre los aspectos: las correlaciones más altas son entre puntualidad (e1) y servicio acorde a lo ofrecido (e2) por un lado, seguridad (e9) y avión en buen estado (e10) por otro, y por otro lado experiencia placentera (e5) con comodidad (e8).



En los siguientes resultados auxiliares a este análisis, tenemos que la primera y segunda componente principal conservan una varianza de 58.1 % y 9.7 % respectivamente. Además si consideramos 3 factores, tenemos consistencia en lo planteado con los grupos. En ambos casos podemos observar e1 y e2 muy correlacionados o en el mismo factor; e5, e8, e16 y e22 por otra parte; y e9 y e10 por otra parte.



En la primera gráfica siguiente, vemos que a la derecha se encuentran los clientes potenciales con buenas expectativas en general en todas las preguntas, y a la izquierda los de regular y mala, según el primer componente principal.



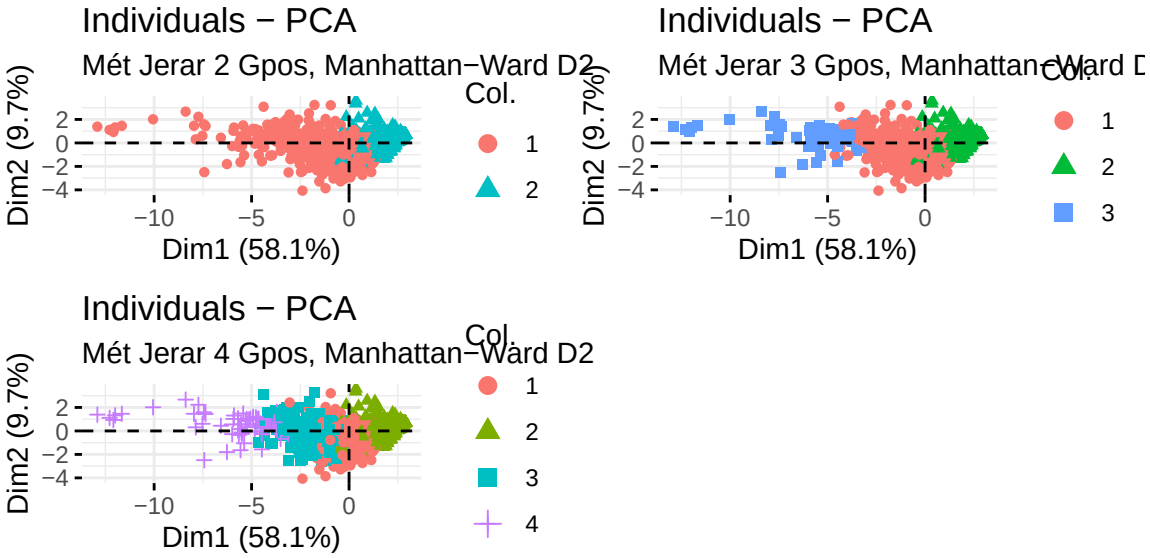
De acuerdo al método silhouette, se nos sugiere tomar dos grupos, pero en nuestro caso para mejorar la interpretación decidimos que es recomendable usar más de 2. Adicionalmente podemos ver que no cambia mucho la interpretación si nos quedamos con 3 grupos o con 4, pues cuando agrupamos en 4 grupos, el grupo 4 combina parte de los grupos 1 y 3.

Observando los componentes principales, podemos decir que es mejor focalizar la publicidad en 3 grupos de clientes: los que esperan puntualidad y un servicio acorde a lo contratado; los que esperan seguridad y buen

mantenimiento y estado del avión; y los que esperan comodidad, experiencia, hospitalidad y entretenimiento. Tal como se decidió agrupar desde un inicio.

Metodo Jerarquico Aglomerativo

Para esto vamos a tomar que las variables son continuas como se hizo anteriormente y tomando tanto las escalas dadas como haciendo transformaciones. Ademas agregaremos las disimilaridades entre clientes y clusters.



En esta figura podemos observar las mismas comparaciones que realizamos en el ejercicio 1 donde se puede ver que el resultado obtenido en este caso aplicando el método aglomerativo resulto ser muy similar al obtenido con K-means. En esta ocasión los 3 grupos de clientes son: los que esperan puntualidad y servicio acorde a lo ofrecido, los que esperan seguridad y buen estado y mantenimiento del avión, y los que esperan experiencia placentera, comodidad, hospitalidad y entretenimiento.

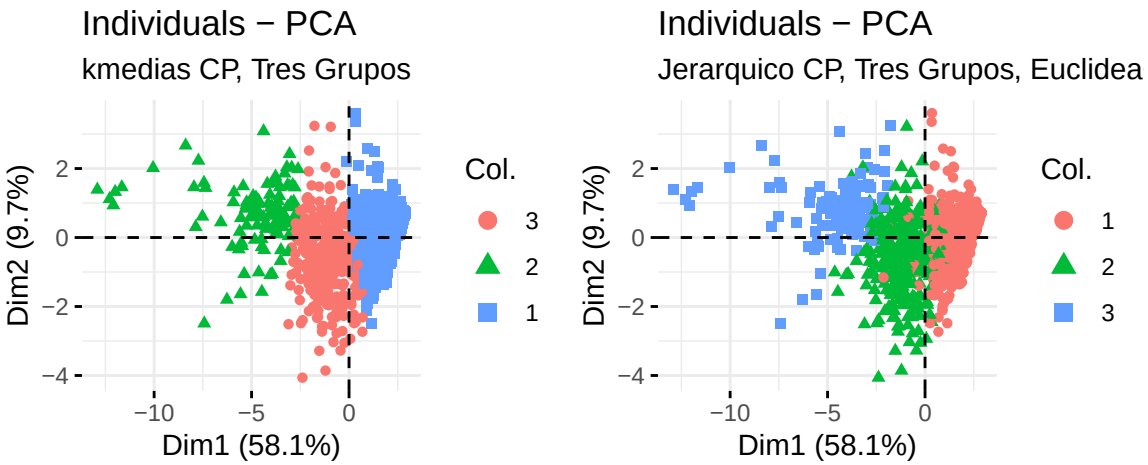
Modificaciones y uso de Componentes principales

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5   PC6    PC7    PC8
## Standard deviation     2.410 0.9830 0.7944 0.7565 0.7025 0.656 0.5895 0.5759
## Proportion of Variance 0.581 0.0966 0.0631 0.0572 0.0493 0.043 0.0348 0.0332
## Cumulative Proportion  0.581 0.6775 0.7406 0.7978 0.8472 0.890 0.9249 0.9581
##                            PC9   PC10
## Standard deviation     0.4773 0.4372
## Proportion of Variance 0.0228 0.0191
## Cumulative Proportion  0.9809 1.0000
```

Primero vamos a hacer el proceso de K-means con las 4 componentes principales que se escogieron.

Análogamente a los ejercicios anteriores vamos a probar usando clusters jerarquicos y conservando las disimilaridades que se usaron en el caso anterior

Obtuvimos que los mejores modelos a usar para 3 clusters fueron Euclidean, Minkowski y Ward D2.



Conclusiones

Como pudimos ver a lo largo de todo este análisis y aplicando distintos métodos de evaluación como fue usar K-means, algoritmos de jerarquía y componentes principales decidimos conservar el de Componentes principales ya que ademas de permitirnos conservar las variables que conservan mayor información dadas las originales y así poder reducir el estudio a estas los resultados obtenidos fueron mas cercanos a lo que deseamos, por ejemplo, la clusterizacion que obtuvimos con la primer componente fue mejor, lo mismo pasó para la segunda componente. Hablando en términos mas generales tenemos que el primer grupo tiene mayor promedio en todas las respuestas, seguido por el segundo grupo y por ultimo se queda el tercer grupo.

Finalmente, creemos que el modelo a utilizar para focalizar la publicidad al publico siempre dependerá en gran medida de el numero de la cantidad de publico que quiera alcanzar la empresa y conforme a esto lanzar los distintos tipos de publicidad, ya que nosotros decidimos tomar 3 clasificaciones sobre 2 0 4, esto con el fin de mantener un equilibrio entre las preferencias de todos los clientes que buscan seguridad, puntualidad y un buen trato por parte de los trabajadores, cosas que sin duda son fundamentales para que la empresa logre atraer nuevos clientes potenciales que le den una gran importancia a estos criterios.