Aprendizaje Estadístico Automatizado

Profesor: Dr. Gonzalo Pérez de la Cruz.

Ayudantes: Leonardo Daniel de la Cruz Cuaxiloa,
Fernando Raúl Garay Araujo.

Estudiantes: Carlos Iván Canto Varela,
Marcos Guillermo Isunza Álvarez.

Examen 2A: Selección de variables, *bootstrap* y aprendizaje no supervisado

1. *Bootstrap* no paramétrico

---

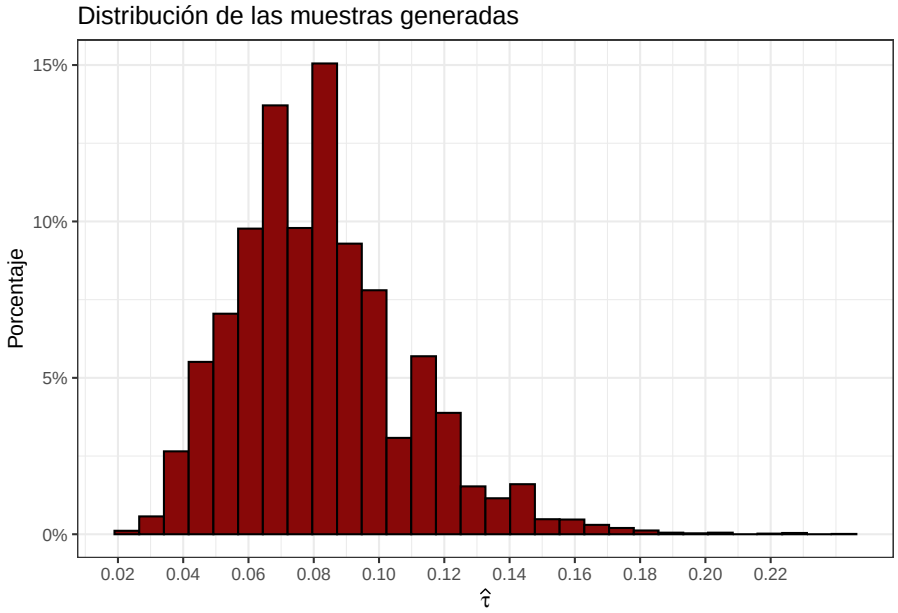
a. Método Monte Carlo

Para estimar el parámetro se generan diez mil muestras —cada una con 25 observaciones— de la variable aleatoria $\widehat{\tau} \sim \text{Poi}(\lambda = 2.5)$.

De este modo, el método Monte Carlo indica que

$$\begin{aligned} \text{E}\left[\widehat{\tau}\right] &\approx \frac{\sum_{i=i}^{10000} \widehat{\tau}_i}{10000} \\ &= 0.0819575, \end{aligned} \qquad \& \qquad \begin{aligned} \text{V}\left[\widehat{\tau}\right] &= \text{E}\left[\widehat{\tau}^2\right] - \text{E}\left[\widehat{\tau}\right]^2 \\ &\approx 0.0007074. \end{aligned}$$

Las muestras generadas se pueden ver en la figura siguiente:



Histograma para las muestras generadas por Monte Carlo

b. *Bootstrap* no paramétrico

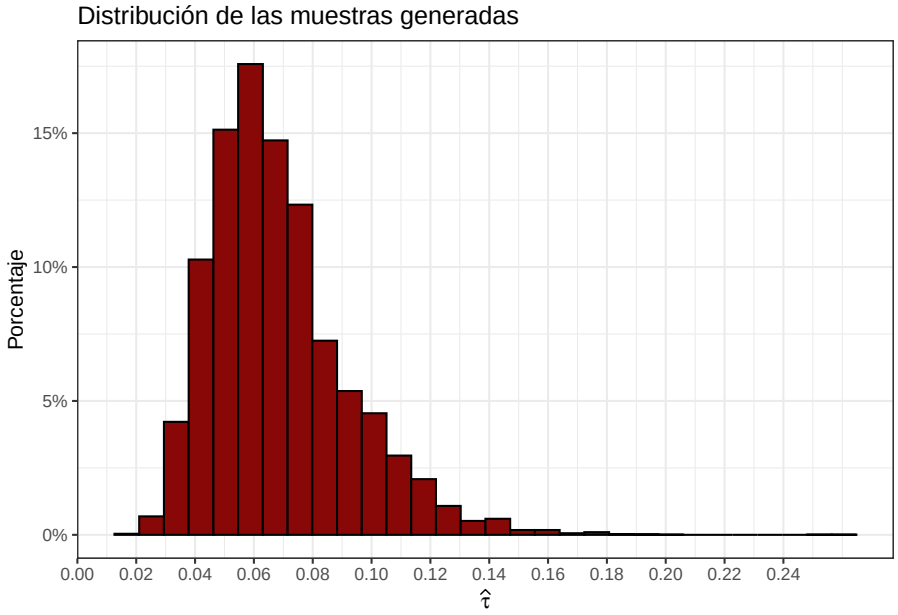
En contraste, para aplicar *bootstrap* sólo se necesita tomar una muestra Poisson, en este caso de tamaño 25.

```
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = b_tau, statistic = tau_stat, R = 10000)
##
##
## Bootstrap Statistics :
##       original      bias    std. error
## t1* 0.06488923 0.003449598  0.02360989
```

De aquí se tiene que

$$\mathrm{E}\left[\widehat{\tau}\right] = 0.0648892 \qquad \wedge \qquad \mathrm{V}\left[\widehat{\tau}\right] = 0.0005574.$$

Luego se presenta la visualización de las muestras del método:



Histograma para las muestras generadas por *bootstrapping*

En este caso, los métodos difirieron en aproximadamente 0.017 para la esperanza del estimador y 0.00015 para su varianza, sin mencionar la similaridad visual hallada. Como se esperaba, ambos se asemejan a una distribución Poisson con parámetros ligeramente modificados.

La cercanía que se alcanzó es encomendable, especialmente al considerar que *bootstrap* no necesita más que los datos crudos para mostrar una estimación aceptable del parámetro y su distribución. Sin embargo, los pocos (25) datos de entrada hacen que los resultados varíen con cada aplicación, de modo que mientras menos observaciones mayor deberá ser la calidad de la muestra proporcionada. Por otra parte, Monte Carlo se mostró más consistente en repeticiones subsecuentes.

2. Selección de variables

Nuestro objetivo es responder a la siguiente pregunta: ¿Que variables modelan de manera óptima el promedio del porcentaje de grasa corporal en los hombres?

Para eliminar del análisis los casos con valores extraños primero necesitamos saber cuales son esos valores extraños. Por un análisis exploratorio establecemos que consideraremos como casos extraños a los valores de peso mayores a 250 lb y a los valores de altura menores a 60 pulgadas. También se desecharon las observaciones con *brozek* cero.

Inicialmente, se propone el modelo de RLS (liga identidad y distribución Gaussiana) para datos continuos:

$$glm : Normal \quad con \quad E(\hat{brozek}) = \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p$$

Donde $E(\hat{brozek})$ representa el promedio del porcentaje de grasa corporal en hombres y x_p representan a las variables que ayudan a modelar de mejor manera el promedio del porcentaje de grasa corporal.

i) Sólo efectos principales

a) Mejor subconjunto

Se realizó la selección de variables por el método de mejor subconjunto usando el algoritmo *exhaustive* de la función *regsubsets* del paquete leaps. Entonces se obtuvieron 14 combinaciones distintas de variables, una por cada subconjunto de covariables; la mejor presentó a las covariables *height, abdom* y *wrist* y un BIC de 1412.142.

b) Método stepwise

Para realizar la selección de variables usando el método stepwise se implemento la misma paquetería pero con los algoritmos *forward* y *backward*.

Los modelos obtenidos compartieron las covariables *abdom* y *wrist*, pero el método forward incluyó *weight* y alcanzó un BIC de 1412.255; el *backward* prefirió a *age* y obtuvo un puntaje de 1415.872 en BIC.

Método Lasso

Para realizar la selección de variables por el método de penalización Lasso se utilizó el paquete "glmnet". Se obtuvo un BIC de 1413.107 con las variables *age, height, abdom* y *wrist*.

ii) Efectos principales e interacción

Los siguientes resultados se consiguieron de manera igual que en la sección anterior. Esta vez, a diferencia del inciso anterior, se tomó en cuenta las interacciones entre las variables así como los efectos de las variables principales.

a) Método stepwise

El ajuste del método backward halló un BIC de 1416.311 con las covariables *hip, height*hip, neck*abdom* y *neck*hip*. Por otro lado, en *forward* se incluyeron *abdom, height*wrist* y *chest*hip* para alcanzar 1405.596 en el criterio BIC.

b) Método Lasso

Junto al criterio de máxima verosimilitud para encontrar los coeficientes del modelo, se intnetó conseguir un modelo por método Lasso. Sin embargo, no se halló la convergencia necesaria con el método "glmnet" utilizado.

Modificaciones: distribución Gamma

Para propósitos de presentación y ahorro del espacio asignado a cada problema, en los siguientes modelos ajustados solo se indica que variables se están utilizando y cual es el valor BIC obtenido para cada modelo ajustado.

Inciso i)

Para realizar la modificación se ajusto un objeto "glm" en R, y usando las variables obtenidas en el método de mejor subconjunto, es decir, las variables height, abdom, wrist.

Para el modelo con liga identidad se obtuvo la puntuación BIC: 1493.01, mientras que el de función liga logarítmica obtuvo el BIC 1512.765.

Para las variables seleccionadas mediante el método forward se consideran las variables *weight, abdom* y *wrist*. Para este modelo ajustado se obtuvo una puntuación BIC de 1494.935 con liga identidad y 1510.475 para logarítmica.

Para realizar la modificación del método backward se consideran las variables *age, abdom* y *wrist*. Y se obtiene mediante la siguiente función en R.

Para este se obtuvo una puntuación BIC de 1494.935, para el modelo ajustado con función liga *log* se obtuvo el siguiente valor BIC: 1509.475.

Para realizar la modificación del método lasso se consideran las variables *age, height, abdom* y *wrist*.

El valor de BIC que se obtuvo es: 1497.058, para el modelo ajustado usando la función liga *log* se obtuvo una puntuación de 1513.417.

Inciso ii)

Para la transformación del metodo backward del segundo inciso se van a considerar las siguientes variables: *hip, height*hip, neck*abdom, neck*hip*, obtenidas del método backward con interacciones. Se obtuvo la siguiente puntuación del criterio BIC: 1498.046. Para el modelo ajustado con función liga *log* se obtuvo el siguiente valor BIC: 1510.402. Para el modelo ajustado con función liga *log* se obtuvo el siguiente valor BIC: 1510.339.

El método Lasso no arrojó resultados por la incompatibilidad mencionada.

Tablas

En la primera de las tablas presenta los modelos obtenidos en los incisos i. y ii. del problema con su respectivo valor BIC para cada uno de los modelos presentados. En la segunda tabla se presenta los modelos obtenidos en el inciso iii. del problema. Dada la longitud de la tabla esta se incluye en el en una sección al final de este reporte.

Interpretación de los resultados.

Para los modelos obtenidos en los incisos i) y ii) (*Cuadro 1*) se puede observar lo siguiente:

Modelo simple	**BIC**
Mejor subconjunto	1412.142
Forward	1412.255
Backward	1415.872
Lasso	1413.107

Modelo con interacciones	**BIC**
Backward	1416.311
Forward	1405.596

- La variable que aparece con más frecuencia en los modelos es la variable abdom con 7 apariciones, seguida por la variable wrist con 5 apariciones. Por lo tanto podemos afirmar que la variable abdom, la cual corresponde a la medición de la circunferencia del abdomen, y la variable wrist, correspodiente a la medición de la muñeca, son variables significativas en la predicción de la grasa corporal en los hombres.
- La variable age también aparece al menos 3 veces, sin embargo dos de los modelos en los que aparece tienen asociado un valor BIC alto por lo tanto no podemos afirmar que sea una variable significativa en la predicción del promedio de grasa corporal.
- El modelo con el valor BIC más alto fue el
- El modelo con el valor BIC más bajo corresponde al penúltimo modelo, con valor BIC de 1405.596, que es el modelo ajustado cuyas variables fueron seleccionadas por el método Lasso. En este caso, dado que este modelo fue el mejor de todos los modelos, entonces podemos afirmar que la variable abdom, y las interacciones height*wrist y chest*hip son significativas en predecir el valor de grasa corporal promedio en hombres.

Para los modelos lineales generalizados ajustados con una distribución Gamma y funciones liga identidad y log se puede observar lo siguiente:

Modelo Gamma	**BIC**
Mejor subconjunto, simple, liga identidad	1493.01
Mejor subconjunto, simple, liga log	1512.765
Forward, simple, liga identidad	1494.935
Forward, simple, liga log	1510.475
Backward, simple, liga identidad	1494.935
Backward, simple, ligalog	1509.475
Lasso, simple, liga identidad	1497.058
Lasso, simple, liga log	1513.417
Backward, con interacciones, liga identidad	1498.046
Backward, con interacciones, liga log	1510.339
Forward, con interacciones, liga identidad	1510.402
Forward, con interacciones, liga log	1510.339

- La variable más común en estos modelos es abdom, la cual aparece 13 veces, seguida de la variable wrist la cual aparece 12 veces. Además los coeficientes asociados a estas variables son positivos para todos los modelos, por lo cual podemos afirmar, basándonos en nuestro análisis, que las variables más significativas en predecir el valor promedio de grasa corporal en hombres son las variables abdom y wrist.
- Las variables que no aparecen en el modelo son las siguientes: adipos, thigh, knee, ankle, biceps, forearm. Por lo tanto, de acuerdo a los modelos del Cuadro 2, ninguna de las variables anteriormente mencionadas tiene algún efecto sobre el valor de grasa promedio en hombres.
- Las interacciones encontradas en las variables fueron las siguientes: hip:height, hip:neck, chest:hip, neck:abdom, height:wrist, sin embargo estas interacciones aparecieron no más de dos veces, algunas únicamente una vez, y además sus coeficientes asociados son valores pequeños por lo que no existe sustento para considerar a estas interacciones como determinantes en la predicción del promedio de grasa corporal en hombres.
- El primer modelo correspondiente a la modificación del método de mejor subconjunto, es el modelo con el valor BIC más bajo, por lo tanto, de entre todos los modelos **modificados** con distribución Gamma es el modelo que mejor predice el valor promedio de la grasa corporal en hombres.

3. Componentes principales y análisis factorial exploratorio

El objetivo principal es describir las observaciones con un número reducido de variables y hallar proyecciones para coadyuvar a una posible reinterpretación.

i. Componentes principales

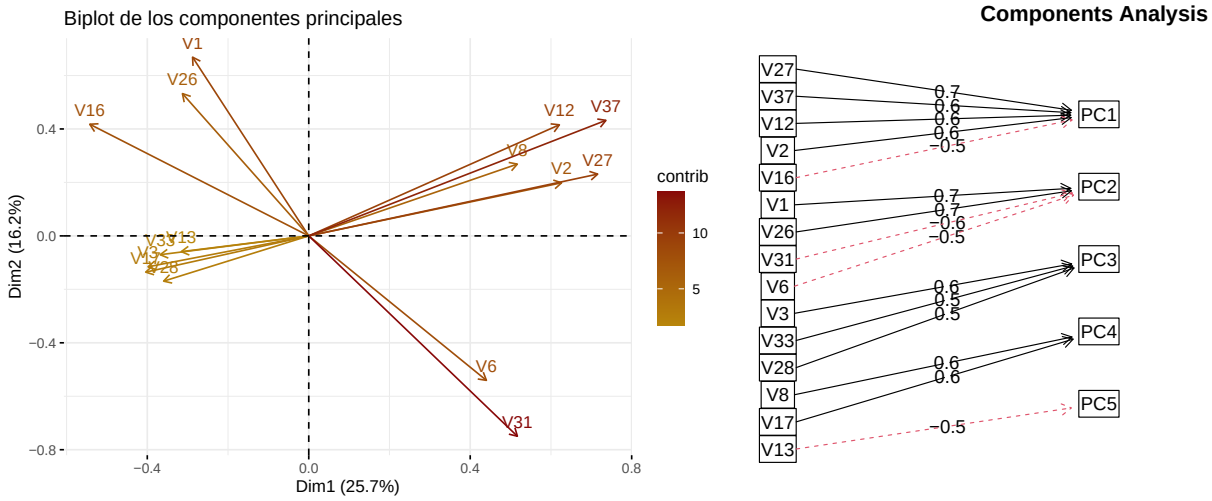
Escala original

La manera más sencilla es analizar sin transformar los datos.

De aquí parecen haber asociaciones según la influencia gráfica de las primeras dos componentes principales. Se puede decir que la primera representa la percepción negativa generada en los demás. Se espera que no se tenga un buen desarrollo social a valores positivos, mientras que negativos serían de alguien agradable y diligente con sus responsabilidades. La segunda se asocia a la extroversión o introversión del individuo.

Escala transformada

Para hallar un mejor contraste se puede usar una escala distinta y realizar el análisis de componentes principales con el método de maximización de varianza explicada sin el requisito de correlación cero.



Biplot de las CP para datos sin escala (←) y análisis de CP para datos con escala logarítmica (→)

Al aplicarle logaritmo natural a los datos se hallaron asociaciones similares.

- Con la escala logarítmica, las preguntas de alta correlación para la componente uno son las preguntas 2, 12, 27 y 37, mientras que la 16 es de correlación negativa; el valor positivo de esta componente se asocia al desenvolvimiento negativo con otras personas.
- Para la segunda componente, las de mayor influencia positiva son la 1 y 26: características usualmente asociadas a la extroversión. Por otro lado, las preguntas 6 y 31 tienen correlación significativamente negativa y tratan más sobre ser introvertido. Con estos resultados, se puede decir que esta componente revisa la forma en que uno responde ante la socialización.
- La componente 3 es influenciada positivamente por la variable 3, 28 y 33. Podría decirse que define la ética de trabajo individual y la responsabilidad.
- Hay correlación notoria de las preguntas 8 y 17 con la cuarta componente. Puede ser la asociada a la naturaleza relajada prominente en los hippies.

- La quinta componente se afecta negativamente por la pregunta 13, de donde se infiere que apunta a la desconfianza generada por la persona.

Aunque se explica el 65 % de la varianza, la interpretación se dificulta con más componentes. Sin embargo, estos pueden ser de utilidad para formar un panorama general de los pacientes de un terapeuta.

ii. Análisis exploratorio factorial

Se seguirá el mismo plan de acción para esta sección.

Escala original

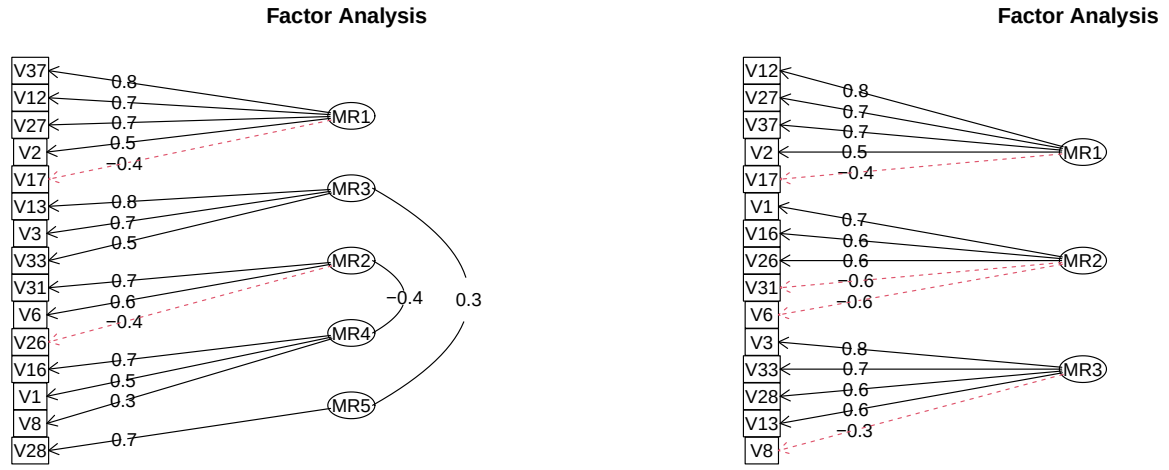
Con cinco factores se encontró una interpretabilidad adecuada con una representación aceptable para todas las variables. El primer factor explica las preguntas asociadas con los estímulos extrapersonales negativos y la incapacidad de perdonar en menor medida. El tercero se asocia a características de un buen trabajador y está relacionado con el quinto factor, que modela la tendencia a perseverar hasta acabar las tareas. El segundo trata a la personalidad sumisa, tanto introvertida como insegura, y está relacionado negativamente con el cuarto, que ve el carácter abierto y amigable. Estos factores podrían servir para que el encargado de recursos humanos de una institución optimice su clasificación de candidatos.

Escala transformada

Para este caso se estandarizaron de la forma usual a los datos:

$$z = \frac{x - \mu}{\sigma},$$

mediante la función *scale* de R base.



Análisis factorial para los datos sin transformar (izquierda) y estandarizados (derecha)

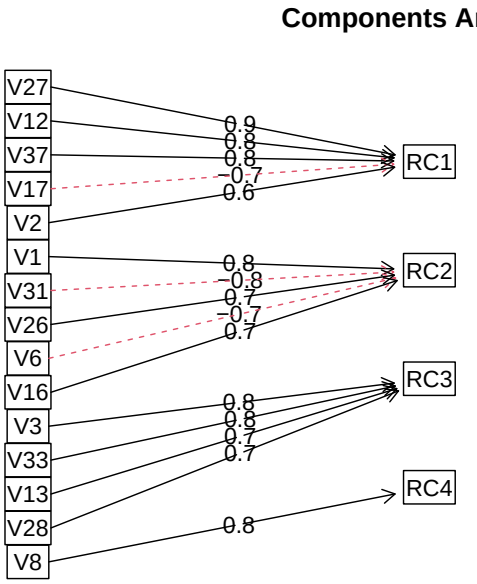
La interpretación se modifica levemente para tres factores, aunque el modelo pierde capacidad de explicación de las preguntas 8 y 17. El primer factor queda exactamente igual al caso anterior, pero el segundo trata a la extroversión personal. El tercero se asocia más a la eficiencia y responsabilidad laboral. Este análisis podría ser de utilidad para que una filtrar pretendientes amorosos, por lo que un sitio de citas como Tinder haría buen uso de él.

iii. Modificaciones

Se procede a experimentar con la categorización ordinal y rotaciones. Por lo tanto se calculó la matriz de correlaciones por el método policórico para los datos ordinales mediante la paquetería *polycor*.

Dado el mayor porcentaje de varianza explicada (0.66) y correlaciones más extremas —además de una interpretación más sensata— el mejor modelo es el de cuatro componentes y rotación *cluster*. A continuación están las representaciones de las componentes principales halladas:

1. Con las preguntas 2, 12, 27 y 37 con alta y la 17 con baja correlación, se tiene que esta componente involucra a la incapacidad de formar buenas relaciones con los demás.
2. Presenta a la extroversión y carisma dado que las preguntas 1, 16 y 26 tienen alta influencia positiva y las 6 y 31 negativa.
3. Son las señales de un buen trabajador: centrado, responsable y conclusivo, descritas por la buena relación con las preguntas 3, 13, 28 y 33.
4. La última componente acapara sólo a la pregunta 8; se puede decir que trató a esta variable "atípica" al resumir el potencial de valemadrismo. Si acaso esta variable podría relacionarse de forma negativa con la tercera componente, los datos no parecen haberlo sugerido. Esto podría ser consecuencia de un mal planteo de la pregunta, un posible sesgo en los encuestados, o quizá simplemente no hay correlación entre estas preguntas.



Componentes principales de datos ordinales con rotación *cluster*

4. Análisis de conglomerados

Considere los datos en el archivo Dat4ExA.csv, sólo los casos con respuesta en todas las variables. Estos datos corresponden a una encuesta realizada por la compañía Oddjob Airways con la intención de conocer las expectativas de sus clientes sobre ciertos aspectos del servicio de la compañía. El objetivo es analizar si se pueden identificar grupos de clientes que en un futuro se puedan usar para focalizar la publicidad de la empresa. Las respuestas van de 1 a 100, donde 100 es que la persona considera que ese aspecto es crucial en el servicio, mientras que 1 corresponde a que no lo es.

- a) Asumiendo que las variables son continuas, obtenga algunos grupos considerando el método k-means. Explore el uso de los datos en la escala original y con alguna escala tranformada.

Resultados.

El análisis que se expone a continuación se enfoca en responder a la siguientes preguntas: ¿Es posible identificar grupos de clientes que en un futuro se puedan usar para focalizar la publicidad de la empresa?, ¿Que afirmaciones podemos acerca de esos grupos de clientes?.

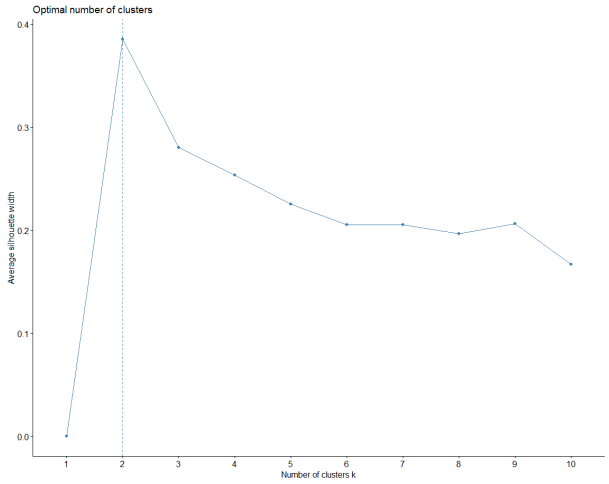
Al realizar el análisis de los datos se obtuvieron dos grupos o **clusters**. El análisis se implemento en el lenguaje de programación R y para llevarlo acabo se usaron las siguientes librerías: GGally, psych, factoextra, Nbclust, tidyverse.

Antes de proceder a realizar el análisis fue necesario completar ciertos datos faltantes pues dentro del archivo *Dat4ExA* había valores faltantes representados por las iniciales *NA* que corresponden a preguntas de la encuesta no respondidas por ciertos clientes. Para poder proceder con el análisis estos valores se rellenaron usando el promedio de los valores para cada columna.

Para poder realizar un análisis efectivo de los datos se decidió primeramente efectuar un análisis de componentes principales, las razones de esta decisión son variadas y se exponen a continuación:

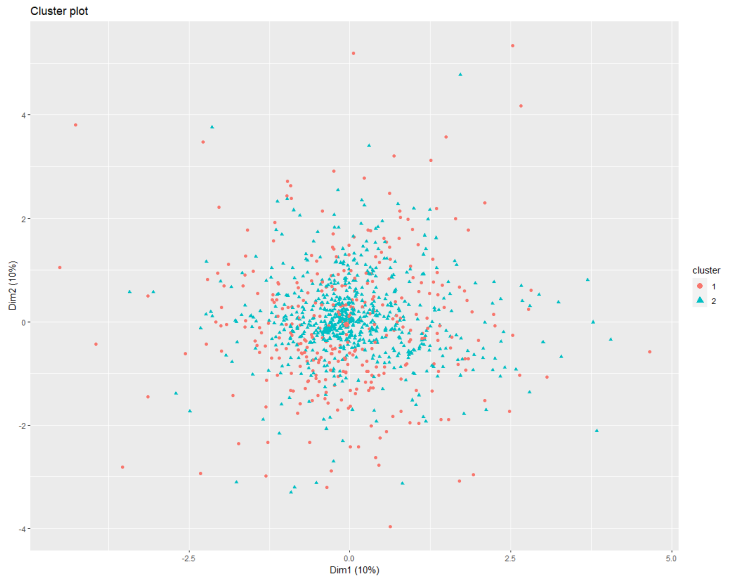
1. Reducir la dimensionalidad: La encuesta de la compañia tiene como propósito evualuar la expectativas de los clientes con los servicios de la compañia a tráves de 10 aspectos, esto se traduce a 10 variables a considerar en el conjunto de datos por lo que existe una cantidad moderada de dimensiones, al realizar un análisis de componentes principales antes de aplicar el algoritmo k-means se logra reducir la dimensionalidad de los datos facilitando el proceso de agrupación posterior.
2. Visualización de los resultados: Reducir la dimensionalidad a componentes pricipales permite visualizar los resultados obtenidos con mayor facilidad lo cual facilita la exposición de los resultados y la comprensión de los **clusters** obtenidos.
3. Correlación entre las variables: Dado el número de aspectos evaluados en la encuesta es posible que algunas variables esten correlacionadas lo cual dificulta el análisis de conglomerados, al realizar un análisis de componentes principales ayuda a eliminar la multicolinealidad de las variables al crear componentes que son combinaciones lineales de las variables originales lo cual facilita la agrupación de los datos.

Para realizar el análisis de componentes principales en R se utilizó la función *prcomp()*. Posteriormente a la creación del objeto **pca**, se obtuvo, mediante la función *fviz_nbclust()*, la siguiente gráfica que indica el número óptimo de clusters de acuerdo al método *silhoutte* el cual es una técnica para evaluar la calidad de un clustering de un conjunto de datos.



Número óptimo de clusters mediante el método Silhoutte

En el método Silhoutte se elige el número de clusters que refleje la puntuación de silhoutte más alta, en la gráfica podemos observar que la puntuación de Silhoutte más alta se alcanza cuando el número de clusters es 2. Una vez que conocemos el número de grupos en los que se van a dividir los datos podemos utilizar la función *fviz_cluster* para obtener una gráfica donde se muestren ambos clusters obtenidos. A continuación se presenta la gráfica corrrespondiente:



Grupos obtenidos. Dim1 vs Dim2

Discusión de los resultados.

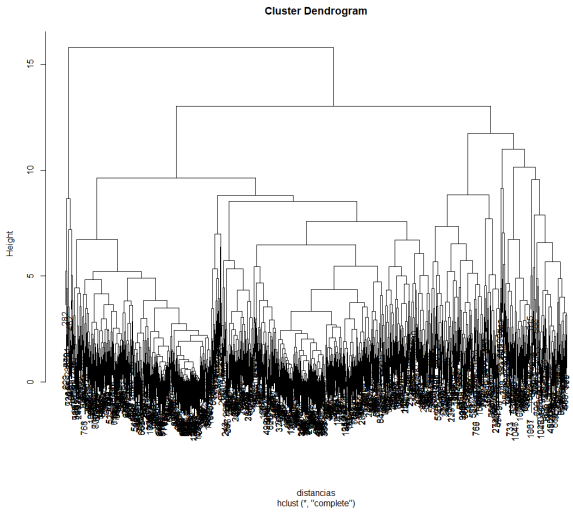
El eje x de la gráfica corresponde a la primera dimensión del análisis de componentes principales, mientras que el eje y de la gráfica corresponde a la segunda dimensión del análisis. Las leyendas de 10 % indican que cada una de estas leyendas explican el 10 % de la varianza total en los datos. Recordemos que al

realizar un análisis de componentes principales cada componente principal corresponde a una combinación lineal de las variables originales de manera tal que cada dimensión no se encuentra correlacionada con las otras dimensiones, por lo tanto, en este caso, cada dimensión explica la dirección de las componentes principales donde se encuentran las mayores diferencias en respuestas entre los clientes encuestados. En este caso las dimensiones Dim1 y Dim2 representan aspectos de las expectativas de los clientes respecto a los servicios ofrecidos por la compañia Oddjob Airways.

Además se puede observar que la mayoría de las observaciones en ambos clusters 1(rojo) y 2(azul) se encuentran centrados respecto al punto (0,0), la forma en que están dispersos los datos sugiere que podríamos trazar imaginariamente dos elipses concéntricas, la más pequeña formada por el cluster 2 correspondiente a los puntos azules y la más grande formada por el cluster 1, correspondientes a puntos rojos, esto indica que la mayoría de los encuestados tienen opiniones similares respecto a los aspectos evaluados en la encuesta, es decir, no existe mucha variabilidad en las respuestas de los clientes para los aspectos que la encuesta busca evaluar. Dado lo que hemos obtenido hasta ahorita podemos concluir que hay dos grupos de clientes, para el primer grupo correspondiente al cluster de los puntos en color rojo de la gráfica existe una variabilidad mucho mayor en la puntuación otorgada a las preguntas y por lo tanto en cuanto lo que piensan acerca de los servicios ofrecidos por la compañia dentro de los aspectos evaluados, para el segundo grupo hay una puntuación más concentrada de las respuestas dadas y por lo mismo una mayor similaridad en cuanto su opinión en los aspectos evaluados por la encuesta. Esto sugiere que la compañia Oddjob Airways podría beneficiarse al enfocar su publicidad y su campaña de marketing en estos dos grupos de clientes, adaptando su estrategia dependiendo de si los clientes pertenecen al primer grupo o al segundo.

- Asumiendo que las variables son continuas, obtenga algunos grupos considerando el método de conglomerados jerárquico aglomerativo. Explore el uso de los datos en la escala original y con alguna escala transformada, así como varias disimilaridades (entre clientes y clusters).

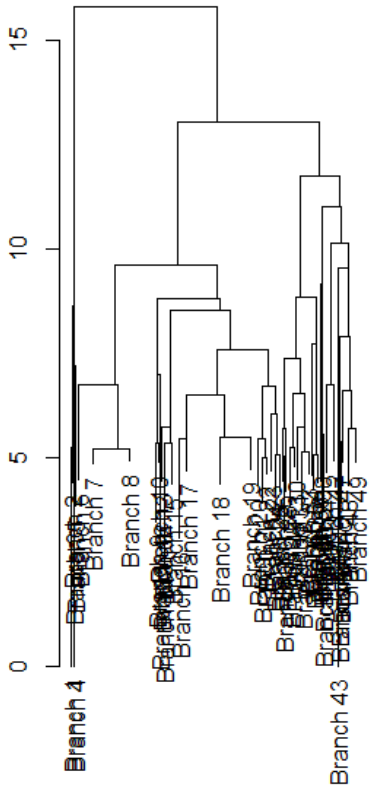
Para realizar el análisis de clusters aglomerativo se utilizo la función hclust() para crear el dendograma, el cual se muestra a continuación:



Dendograma completo

Dada la imposibilidad de leer este dendograma posterior a su obtención tuvimos que limitar la altura del dendoograma a 5. De esta manera obtuvimos el siguiente dendograma que es mucho más sencillo de visualizar.

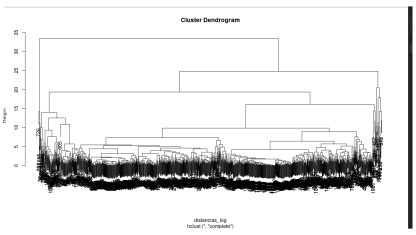
Sin embargo, aún reduciendo la altura del dendograma, encontramos que existen muchas ramas en el mismo, esto se puede deber a múltiples causas, una de las cuales es que tenemos un gran número de observaciones individuales, en el caso de la encuesta realizada, estas observaciones se deben al número



Dendograma cortado en altura 5

total de clientes entrevistados, dada las opciones de la encuesta la variabilidad en las respuestas es tal que dificulta su observación en un dendograma así como su interpretación.

Escala transformada. Escala logaritmica.

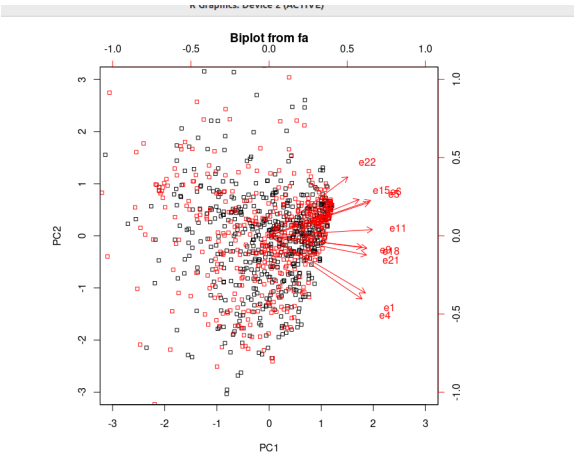


Dendograma tras transformar los datos a escala logaritmica

En el dendograma anterior podemos observar que tenemos una situación similar a cuando obtuvimos el dendograma de los datos sin transformar. Dada la cantidad de observaciones no es imposible determinar clusters específicos a partir del dendograma en esta escala.

Tercer inciso.

Analizamos los conglomerados generando un gráfico de componentes principales.



Gráfica de componentes principales

Interpretación: Se obtuvo una gráfica de dos componentes principales. En la gráfica se pueden observar dos grupos representados por puntos rojos y negros, además se observa que la mayor parte de las observaciones se concentran a la derecha del centro, así mismo, las flechas representan la dirección hacia la cual las variables originales aportan a los componentes. Esto quiere decir que de acuerdo a los aspectos que evaluan las variables e1,....,e22 estos influyen mucho en el primer componente principal.

Considerando a todos los modelos obtenidos para este problema, el modelo más útil para dividir a los clientes en categorías de acuerdo a los aspectos evaluados por la encuesta es el modelo que se obtuvo en el inciso i) puesto que es el modelo más claro en cuanto a resultados y en cuanto a implementación en el lenguaje de programación R, gracias a que se realizo un preprocesamiento de los datos mediante un análisis de componentes principales entonces se pudieron identificar dos clusters de clientes en los cuales la empresa puede focalizar su publicidad, sin embargo, es necesario mencionar que, de acuerdo a este modelo, los clusters de clientes no presentan opiniones tan variadas de los aspectos evaluados por la encuesta.