Examen 2. Seminario de Estadística

Integrantes:

- Bonilla Alarcón Alejandro (420004750)
- Mejía Nájera Angel Josué (420003674)

Versión del examen: B

1	2
3	4

Tarea Examen 2

Versión B

Alejandro Bonilla Alarcón & Angel Josué Mejía Nájera

18-05-2023

Índice

1. Ejercicio 1

1.1. Problema

Sea $X_1, ..., X_n$ una muestra aleatoria de la distribución $Poisson(\theta)$. Es de interés estimar el parámetro $\tau(\theta) = e^{-\theta} = P(X = 0)$. Se puede verificar que $\tau\hat{(\theta)} = (\frac{n-1}{n})^{\sum_{i=1}^{n} X_i}$ es el UMVUE de $\tau(\theta) = e^{-\theta}$. Para poder estimar el parámetro de interés se consideran el método de Monte Carlo y el método de Bootstrap.

1.2. Método Monte Carlo

En general, con este método

$$E(g(Z)) \approx \frac{\sum_{b=1}^{B} g(Z_b)}{B}$$

donde $Z_1, ..., Z_B$ son números aleatorios de la distribución de la variable aleatoria Z.

De esta forma, se generan B muestras aleatorias de tamaño n de una distribución $Poisson(\theta)$ y se calculan B estimaciones del parámetro de interés, $\hat{\tau}_1, ..., \hat{\tau}_B$.

De manera específica, se supone que conocemos la distribución de las variables $X_1, ..., X_n$ y los paramétros para hacer el método Monte Carlo son $\theta = 2$, $n = 2$, $B = 10,000$.

1.3. Método Bootstrap

Para este método se considera que se tiene una muestra de las variables $X_1, ..., X_n$, sobre la cual se realizó un remuestreo para poder estimar la distribución.

De manera específica, se considera que se cuenta con una muestra $X_1, ..., X_{20}$ de una distribución $Poisson(2)$.

A partir de esta muestra, se realizan 10,000 remuestreos con reemplazo y se calcula $\hat{\tau}$ para cada muestra.

1.4. Resultados

A continuación se muestra el histograma de las estimaciones obtenidas con el método Monte Carlo y el método de Bootstrap, así como un cuadro para comprar los resultados de las aproximaciones de $E[\hat{\tau}]$ y $V(\hat{\tau})$ para cada método.

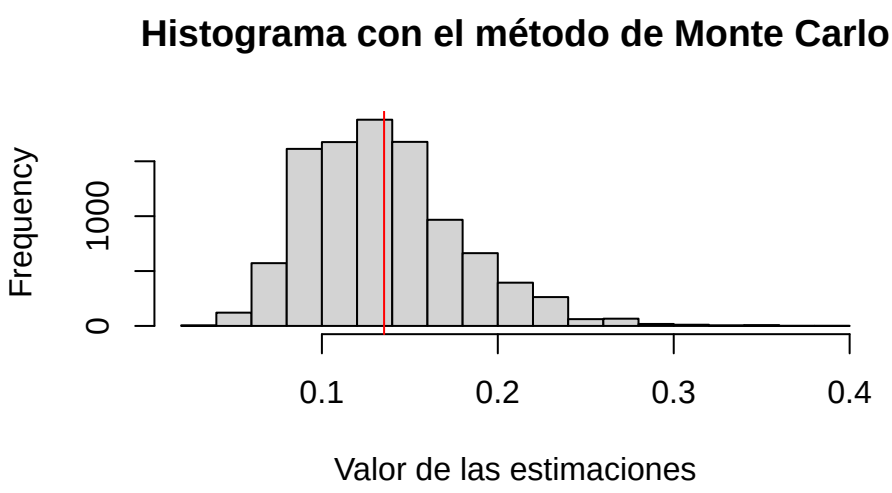


Figura 1: Histograma de las estimaciones con el método de Monte Carlo

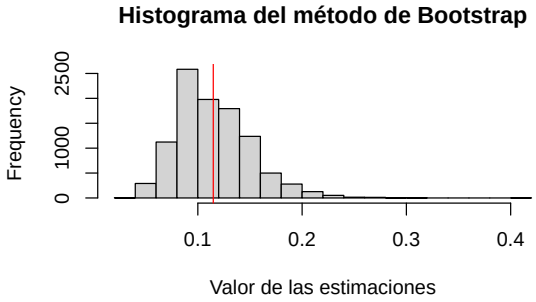


Figura 2: Histograma de las estimaciones con el método de Bootstrap

Cuadro 1: Aproximación de las estadísticas de la distribución del parámetro con el método de Monte Carlo y Bootstrap

	Media	Varianza
Monte Carlo	0.1354	0.0019
Bootstrap	0.1148	0.0012

Con los parámetros conocidos se sabe que $\tau(\theta) = 0.13533$.

A partir del Cuadro 1 observamos que el método Monte Carlo aproxima mejor a τ, pero tambien es el método que arroja más variabilidad en las estimaciones. Sin embargo, el método Bootstrap también se acercó mucho al valor real de τ y cuenta con la ventaja de que no es necesario conocer la distribución de los datos para generar distintas muestras.

2. Ejercicio 2

2.1. Datos

La base de datos *PimaIndiansDiabetes2* del paquete *mlbench* contiene información sobre 768 pacientes por parte del National Institute of Diabetes and Digestive and Kidney Diseases, donde los datos insuales con valores 0 fueron corregidos para ser NAs. El objetivo es realizar una selección óptima y con base en el criterio BIC, pues se desea obtener el menor número posible de variables independientes, de las 8 variables clínicas observadas. De tal forma que, ya sea adicional o como sustituto de la variable *glucose*, ayuden a mejorar la modelación de la probabilidad de presentar o no diabetes (var *diabetes*).

Para esta selección, se tomó en consideración una selección a través del Mejor Subconjunto, métodos Stepwise y penalizaciones Lasso, probando diferentes ligas (*logit*, *probit* y *cloglog*), así como tomando en cuenta Efectos Principales, Interacciones y tranformaciones cuadráticas de las variables.

2.2. Análisis Descriptivo

Variable	Descripción
`pregnant`	Número de veces embarazada(Entero).
`glucose`	Concentración de glucosa en plasma(prueba de tolerancia a la glucosa)(Entero).
`pressure`	Tensión arterial diastólica(mm Hg)(Entero).
`triceps`	Espesor del pliegue cutáneo del tríceps(mm)(Entero).
`insulin`	Insulina sérica de 2 horas(mu U/ml)(Contínua).
`mass`	Indice de masa corporal (peso en kg/(altura en m)^2)(Contínua).
`pedigree`	Función pedigrí de la diabetes(Contínua).
`age`	Edad en años(Int).
`diabetes`	Variable binaria de dos niveles: (2 = pos = Paciente con Diabetes, 1 = neg = Paciente sin Diabetes).

Cuadro 2: Variables originales del DataFrame `PimaIndiansDiabetes2` y su notación dentro del modelo de RLM.

Se comprobó la existencia de NAs en el DataFrame PimaIndiansDiabetes2, por lo que se consideró la eliminación de todas aquellas observaciones que poseyeran al menos una variable con registro NA, quedando en total 392 observaciones de las 768 originales.

A partir de la Figura 3 se observa que:

- El número de no diabéticos observados es casi el doble que el de diabéticos.
- Existen valores atípicos en el número de embarazos, llegando a presentar un paciente con 15 embarazos.
- Las personas con mayor edad presentan diabetes, mientras que los más jovenes no.
- Las personas que poseen diabetes concentran un mayor nivel de glucosa e insulina.

```
## pdf
##   2
```

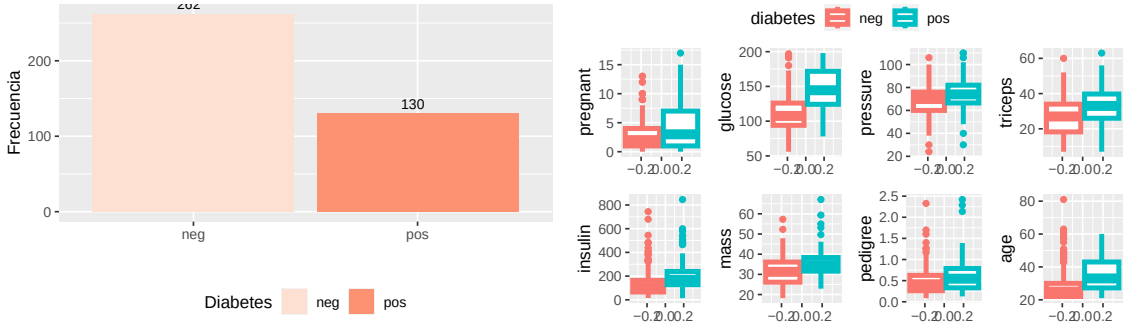


Figura 3: (Izquierda) Gráfico de conteo para la variable diabetes. (Derecha) Conjunto de Boxplots por cada variable independiente y diferenciados según el color si el paciente tuvo diabetes o no.

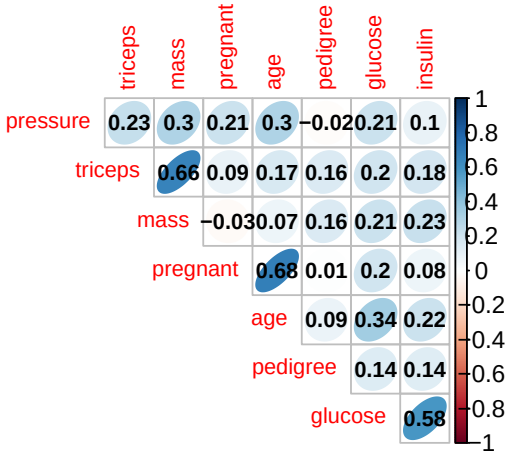


Figura 4: Gráfico de calor y correlaciones de las 9 covariables clínicas.

De la Figura 4, se observan algunas correlaciones significativas, principalmente entre las variables *pregnant* y *age*, *triceps* y *mass* e *insulina* con *glucosa*. Y, en general, se observan correlaciones positivas. Dado que no existen correlaciones mayores a 0.9, se espera no haya problemas de colinealidad.

2.2.1. Estadísticas básicas

En la tabla 3 se aprecia que, en promedio, los pacientes con diabetes presentan mayores valores para todas las covariables que aquellos pacientes que no poseen diabetes. Además, con base en la desviación estándar, se puede inferir que las observaciones de pacientes con diabetes son más dispersas.

Cuadro 3: Tabla comparativa de medias y Desv. Estándar para cada Covariable

Mean(SD)		
Diabetes		
Positivo	Negativo	Variables
4.47[3.92]	2.72[2.62]	pregnant
145.19[29.84]	111.43[24.64]	glucose
74.08[13.02]	68.97[11.89]	pressure
32.96[9.64]	27.25[10.43]	triceps
206.85[132.7]	130.85[102.63]	insulin
35.78[6.73]	31.75[6.79]	mass
0.63[0.41]	0.47[0.3]	pedigree
35.94[10.63]	28.35[8.99]	age

2.3. Modelos Ajustados (Modelo binomial)

En la tabla 4 se consideró lo siguiente:

- Las estimaciones de los coeficientes Betas se realizó a través de Máxima-Versimilitud.
- Dado que el DataFrame solo posee covariables contínuas o enteras, solo se consideró usar el método de penalización Lasso simple en su configuración relax.
- En las estimaciones de los coeficientes a través de penalización Lasso, se estandarizaron las variables (media 0, varianza 1).
- El BIC se ajustó para ser comparable en todos los modelos.
- El cumplimiento de los supuestos se verificó visualmente a través de los residuales Dharma.
- Los modelos con prefijo *log* son aquellos que incluyeron transformación logaritmo sobre las covariables.
- Algunos modelos poseen variables no significativas, sin embargo, se optó por no eliminar dichas variables pues era usual que fuese la variable glucosa o insulina la variable no significativa, lo cual no va en concordancia con lo que dicta la teoría que se tiene sobre la relación entre diabetes y dichas variables.

Por otro lado, en la tabla 4 se observa lo siguiente:

- Las covariables que más se presenten son, en orden de importancia, son: "glucose", "mass", .ªgeƶ "pedigree". Es decir, aparte de la glucosa, para modelar la probabilidad de diabetes, es de importancia considerar el IMC, la edad y la probabilidad de diabetes según los antecedentes familiares (pedigree).
- La covariable ïnsuline"no se incluye con frecuencia en los modelos ajustados, lo cual puede atribuirse a la fuerte correlación que existe entre esta variable y "glucose", así como al posible efecto explicativo mayor que tiene esta última.
- Es notable la reducción sobre el BIC que tienen los modelos que incluyen interacciones y términos cuadráticos, siendo que los 3 mejores modelos incluyen estos efectos adicionales.
- Los modelos con liga logit tuvieron el mejor desempeño, luego los modelos con liga probit y finalmente, con liga cloglog, quienes inclusive tuvieron problemas a la hora de cumplir los supuestos de uniformidad (Dharma).
- Los modelos con transformación logaritmo tuvieron mejor desempeño en cuanto al BIC.
- Se puede concluir que usar interacciones, términos cuadráticos y un preprocesamiento a los datos, mejoró notablemente el desempeño del modelado, según la métrica BIC.

Cuadro 4: Comparación de Modelos seleccionados y ordenados con base en el BIC. Tomando en cuenta distintas ligas, interacciones y/o términos cuadráticos de las 8 covariables

Modelo	Ajuste	Liga	Supuestos	BIC
pedigree, I(glucose * mass), I(glucose * age)	Log_Lasso_Reducción	logit	Cumple	365.7
I(glucose^2), age, mass, pedigree, I(age^2)	log_Forward	logit	Cumple	368.0
glucose, mass, pedigree, age, I(age^2), I(insulin * pedigree), I(insulin * age)	Lasso	logit	Cumple	368.8
glucose, mass, pedigree, age	log_Bestglm	logit	Cumple	369.9
glucose, age, mass, pedigree	log_Forward	logit	Cumple	369.9
glucose, mass, pedigree, age	log_Lasso	logit	Cumple	369.9
pedigree, I(glucose^2), glucose:mass, glucose:age	log_Lasso	logit	Cumple	370.3
glucose, age, mass, I(age^2), pedigree	Forward	logit	Cumple	370.6
glucose, age, mass, I(age^2)	Lasso	probit	Cumple	371.2
glucose, insulin, mass, pedigree, age, I(age^2), insulin:pedigree, insulin:age	Backward	logit	Cumple	371.9
glucose, insulin, mass, pedigree, age, I(age^2), insulin:pedigree, insulin:age	Both	logit	Cumple	371.9
glucose, age, mass, pedigree	Forward	logit	Cumple	377.1
glucose, mass, pedigree, age	Bestglm	logit	Cumple	377.1
glucose, mass, pedigree, age	Backward	logit	Cumple	377.1
glucose, mass, pedigree, age	Both	logit	Cumple	377.1
glucose, mass, pedigree, age	Lasso	logit	Cumple	377.1
glucose, mass, age	Bestglm	probit	Cumple	378.1
glucose, age, mass	Forward	probit	Cumple	378.1
glucose, mass, age	Forward	probit	Cumple	378.1
I(glucose^2), glucose:mass, glucose:age	Lasso	probit	Cumple	382.0
pregnant, glucose, mass	Bestglm	cloglog	Cumple	385.7
glucose, I(pregnant^2), mass	Lasso	cloglog	Cumple	386.1
I(glucose^2), glucose:triceps, glucose:mass, glucose:age	Both_Reducción	logit	Cumple	386.3
glucose, mass, age	Forward	cloglog	No Cumple	386.7
glucose, triceps, age	Forward	cloglog	Cumple	387.9
I(glucose^2), glucose:mass, glucose:age	Lasso	cloglog	No Cumple	394.1

Cuadro 5: Coeficientes del mejor modelo

	x
(Intercept)	-19.7115
pedigree	0.6380
I(glucose * mass)	0.6685
I(glucose * age)	0.4979

Finalmente, nótese que el mejor modelo resultó de una reducción manual del modelo obtenido a través de penalización Lasso, tomando en cuenta interacciones y usando transformación logaritmo y liga logit. Dicho modelo posee solo 3 variables, de las cuales 2 resultan ser interacciones entre la variable *glucose* con las variables *mass* y *age*.

Con base en los coeficientes presentados en la tabla 5, y bajo la consideración de que la función liga es creciente, podemos interpretar que, a mayor edad y/o mayor concentración de glucosa y/o mayor IMC, así como mayor pedigree, podriamos espererar un aumento en la probabilidad de que el paciente presente diabetes.

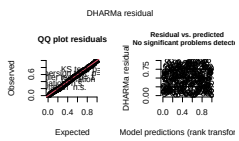


Figura 5: Verificación de supuestos a través de residuales Dharma.

3. Ejercicio 3

3.1. Datos

La base de datos del archivo Dat3Ex.csv contiene los datos de una encuesta que intenta analizar la personalidad de un grupo de 228 alumnos de licenciatura de una universidad de Estados Unidos. Las respuestas van del 1 al 5 donde:

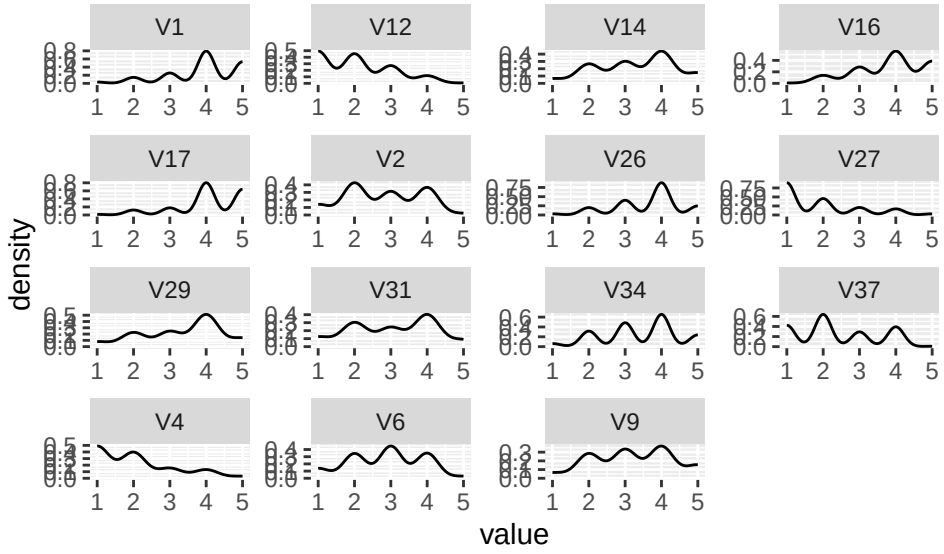
- 1-Disagree strongly
- 2-Disagree a little
- 3-Neither agree nor disagree
- 4-Agree a little
- 5-Agree strongly

y sólo se consideraran las siguientes variables:

- V1-Is talkative, V2-Tends to find fault with others
- V4-Is depressed, blue, V6-Is reserved
- V9-Is relaxed,handles stress well, V12-Starts quarrels with others
- V14-Can be tense V16-Generates a lot of enthusiasm
- V17-Has a forgiving nature V26-Has an assertive personality
- V27-Can be cold and aloof V29-Can be moody
- V31-Is considerate and kind to almost everyone V34-Remains calm in tense situations
- V37-Is sometimes rude to others

3.2. Análisis de datos

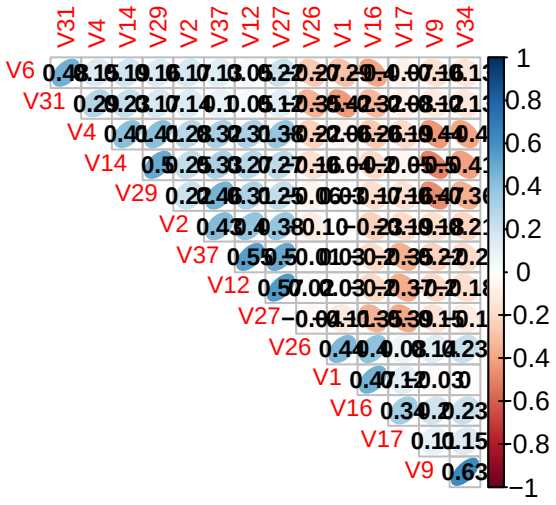
A través de la gráfica de densidades **??**, es posible notar una mayor ponderancia en contestar la opcion 4 (ligeramente de acuerdo) sobre todo en las preguntas de naturaleza positiva salvo algunas negativas como V29 y V14, donde hay mayor preponderancia en contestar la opción 1 es con preguntas de índole negativa.



Cuadro 6: Tabla de Frecuencias con Moda

Pregunta	Moda	1	2	3	4	5
V1	4	4	19	33	103	69
V12	1	84	77	46	20	1
V14	4	12	50	56	83	27
V16	4	1	23	47	93	64
V17	4	2	16	23	104	83
V2	2	25	77	56	66	4
V26	4	4	26	52	114	32
V27	1	116	59	27	22	4
V29	4	15	43	46	98	26
V31	4	24	60	47	80	17
V34	4	8	41	63	85	31
V37	2	55	83	38	51	1
V4	1	93	75	29	25	6
V6	3	24	60	78	61	5
V9	4	12	54	63	70	29

Analizando el correlograma de calor **??**,se observan algunas correlaciones significativas, por ejemplo, la correlacion positiva más fuerte la presenta la pregunta V9 con la pregunta V34 (0.63), lo que se podría interpretar como que hay cierto sesgo a que, si la persona está de acuerdo con ser una persona relajada tiende a ser calmada en situaciones tensas. Además, sólo existe una correlacion negativa <= -0.5, dada entre las preguntas V9 y V14, la cual se podría intepretar como que, si una persona está de acuerdo con ser relajada, tendrá cierta tendencia a no estar de acuerdo con ser una persona tensa.



Con base en los coeficientes presentados en la tabla 6, observamos una mayor preponderancia en contestar la opción 4, la cual se puede interpretar como que los estudiantes suelen estar ligeramente de acuerdo con las cosas.

3.3. Componentes principales

Usando los datos originales y los datos transformados con la función logaritmo se obtienen los componentes principales. En las siguientes Figuras 6 y 7 se pueden observar el desagregado de los componentes principales para ambos tipos de datos.

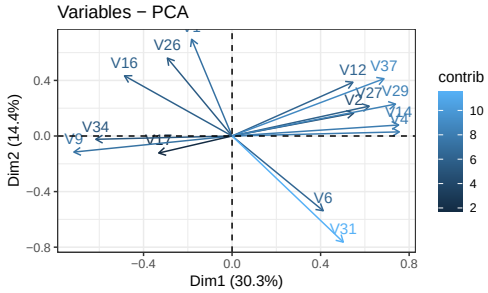


Figura 6: Componentes principales de los datos originales

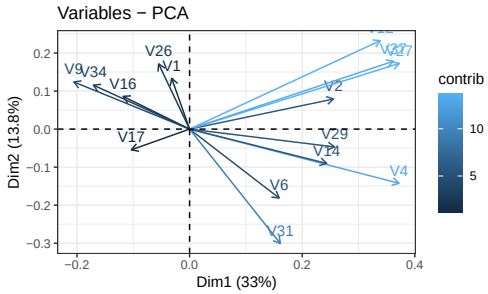


Figura 7: Componentes principales de los datos transformados

3.4. Análisis exploratorio

Usando los datos originales y los datos transformados se obtiene el análisis exploratorio con 3 factores como se observa en las Figuras 8 y 9.

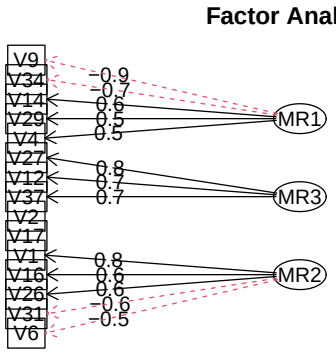


Figura 8: Diagrama usando tres factores con los datos originales

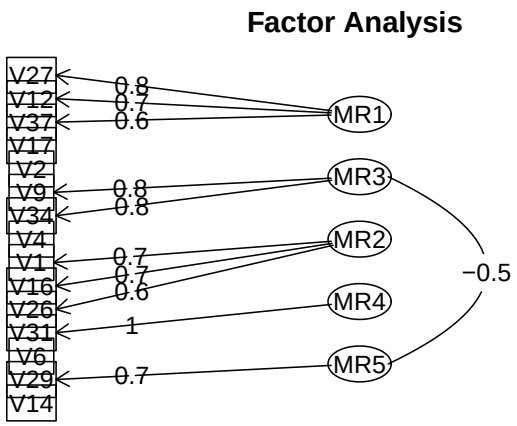


Figura 9: Diagrama usando tres factores con los datos transformados

##Mejor candidato

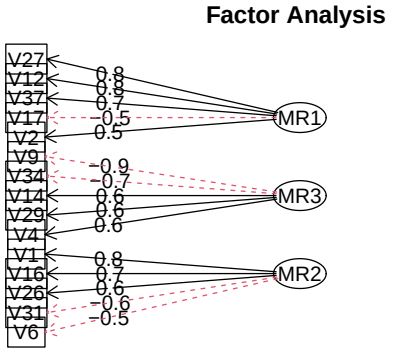


Figura 10: Diagrama usando tres factores con matrix de correlaciones preparada para datos ordinales

Notamos que en el diagrama de factores 10, los 3 factores se agrupan bien con todas las variables propuestas. De tal forma que se podría interpretar que:

4. Ejercicio 4

4.1. Datos

La base de datos Dat4ExB.csv contiene los resultados de una encuesta realizada por la compañia Oddjob Airways con la intención de conocer las expectativas de sus clientes sobre ciertos aspectos del servicio de la compañia. 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. La descripción de los aspectos que se consideran es:

- e1 ". . . with Oddjob Airways you will arrive on time."
- e2 ". . . the entire journey with Oddjob Airways will occur as booked."
- e5 ". . . Oddjob Airways provides you with a very pleasant travel experience."
- e8 ". . . Oddjob Airways offers a comfortable on-board experience."
- e9 ". . . Oddjob Airways gives you a sense of safety."
- e10 ". . . the condition of Oddjob Airways's aircraft is immaculate."
- e16 ". . . Oddjob Airways offers you a variety of foods and beverages that fits your personal needs."
- e17 ". . . all of Oddjob Airways's personnel are always hospitable and welcoming."
- e21 ". . . Oddjob Airways makes traveling uncomplicated."
- e22 ". . . Oddjob Airways provides you with interesting on-board entertainment, service, and information sources."

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.

4.2. Estrategias consideradas

Para poder identificar los datos se consideraron los métodos k-means y el de conglomerados jerárquicos aglomerativo aplicados a los datos en escala original y a los datos estandarizados. Asímismo, se aplicó el método de componentes principales para transformar las variables.

Los datos se estandarizaron para que tuvieran media 0 y varianza 1.

En el método k-means se consideró $k = 2$, tanto para los datos originales como para los datos estandarizados, ya que permitia una mejor interpretación y fue el indicado por el índice Silhouette.

Para el método de conglomerados jerárquicos aglomerativos se provaron varias disimilaridades entre observaciones (Camberra y Euclidian) y entre clusters (complete y ward.D). De igual forma, se consideró $k = 2$ dado que proporcionaba una mejor interpretación al diferenciar bien los grupos, y fue el indicado por el índice Silhouette.

Por último, se aplicó el método de componentes principales y se realizaron los mismos procedimientos descritos anteriormente.

4.3. Resultados

Priorizando la interpretabilidad de los resultados se eligio el método de conglomerados jerárquicos aglomerativo aplicado a los componentes principales de los datos estandarizados.

Como se observa en la Figura 11, el primer componente principal esta correlacionado positivamente con todas las caracteristicas del servicio, mientras que el componente principal 2 esta negativamente correlacionado con los servicios e1 y e2, y positivamente correlacionado con los servicios e16 y e22. Por lo que un alto valor del componente principal 1 representa que todos los servicios son cruciales y un alto valor del compoente principal 2 considera que llegar a tiempo y que el vuelo ocurra como estaba planeado no son aspectos cruciales mientras que la comida, bebidas y entreteniento son aspectos cruciales.

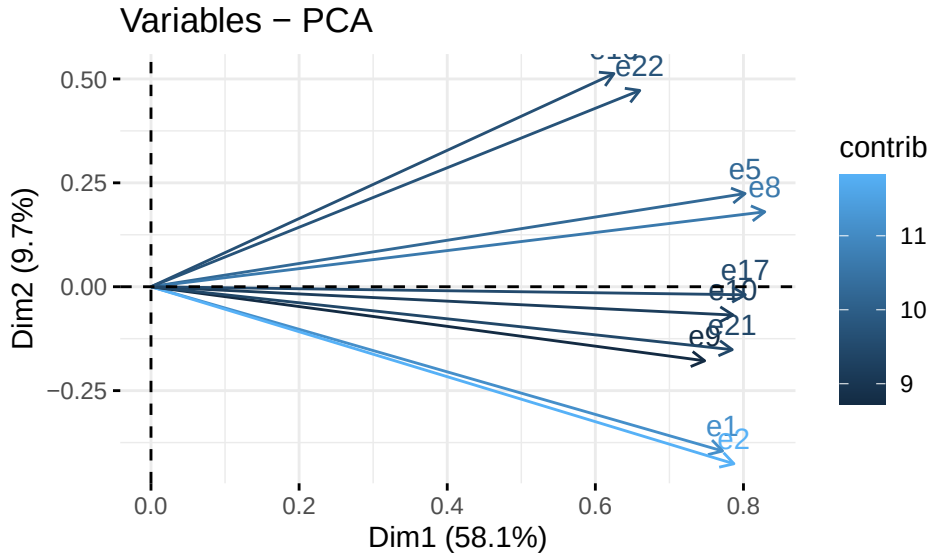


Figura 11: Componentes principales de los datos estandarizados

A los componentes principales de los datos se les aplicó el método de conglomerados jerárquicos aglomerativo usando la disimilaridad de Euclidian entre observaciones y la disimilaridad complete entre clusters para luego diferenciar solo entre dos grupos.

En la Figura 12 se pueden observar los dos grupos considerados.

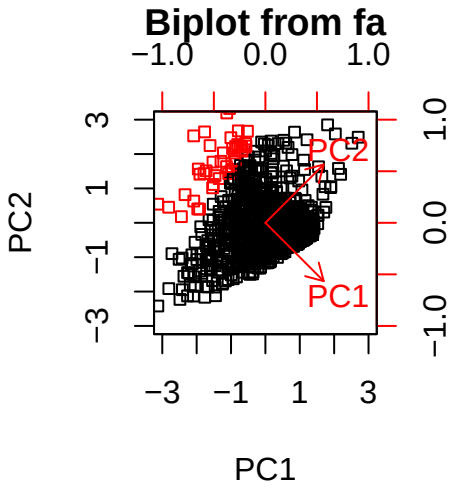


Figura 12: Proyección de los datos sobre los componentes principales

Mientras que en la siguiente tabla se pueden observar los valores de los componentes principales para cada grupo.

```
##      item group1 vars   n     mean     sd  median  trimmed    mad     min
## PC11    1      1    1 886  0.37138 1.8414  0.6958  0.50003 2.0894  -4.384
## PC12    2      2    1  53 -6.20843 2.3310 -5.3794 -5.75110 1.2343 -12.913
## PC21    3      1    2 886 -0.02699 0.9725  0.1372  0.03261 0.8611  -4.063
## PC22    4      2    2  53  0.45121 1.0555  0.6080  0.52370 0.9607  -2.491
## c21     5      1    3 886  1.00000 0.0000  1.0000  1.00000 0.0000   1.000
## c22     6      2    3  53  2.00000 0.0000  2.0000  2.00000 0.0000   2.000
##         max range    skew kurtosis      se
## PC11  2.897 7.281 -0.5046  -0.7302 0.06186
## PC12 -3.945 8.968 -1.5786   1.5115 0.32019
## PC21  3.607 7.671 -0.4935   1.4311 0.03267
## PC22  2.668 5.159 -0.6390   0.1701 0.14498
## c21   1.000 0.000     NaN      NaN 0.00000
## c22   2.000 0.000     NaN      NaN 0.00000
```

Como se puede observar el grupo 1 tiene mayores valores que el grupo 2 en el componente principal 1. En cuanto al componente principal 2, el grupo 1 tiene valores negativos pero cercanos a 0 mientras que el grupo 2 tiene valores postivos.

Así que el grupo 1 son personas que tienen una consideración mayor al promedio de que todos los aspectos son cruciales en el servicio de la aerolínea. El grupo 2 son personas que, con respecto al promedio, no consideran cruciales los servicios generales, pero sí aquellos aspectos que estan fuera del negocio principal de la aerolínea, como la comida, bebidas y el entretenimiento que ofrece.