Proyecto Fase 2

Alejandro Campos, Darian Dominguez, Nelson Mendoza

Facultad de Matemática y Computación
Universidad de la Habana
2021

Resumen

En la segunda fase del proyecto correspondiente a la asignatura de Estadística se hace un estudio sobre los datos usados en la fase 1, en el que aplicaremos técnicas de regresión lineal, ANOVA y reducción de dimensión. Para ello se definen las variables pricipales que están contenidas en el set de datos, sobre las que se realizaran las dos primeras técnicas. Además, se hace un análisis de los supuestos de cada modelo a fin de investigar la validez de estos y se construyen gráficas en cada técnica para analizar los resultados obtenidos de forma visual. El objetivo del proyecto es el análisis de los datos usando las técnicas antes mencionadas. Con ayuda del software R y de R studio, se creó la implementación que se puede encontrar en `code/fase_2.R` *de la cual nos auxiliaremos a lo largo de este informe.*

I. Introducción

La regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, y m variables independientes X_i. El análisis de la regresión lineal consiste en encontrar la ecuación de la recta que mejor describe la relación entre las variables, siendo uno de los usos de esta ecuación hacer predicciones. El análisis de la varianza (ANOVA) parte del concepto de regresión lineal, cuya funcionalidad amplía. Es un procedimiento creado por Fisher en 1925 para descomponer la variabilidad de un experimento en componentes independientes que puedan asignarse a causas distintas. Así, un análisis de la varianza permite determinar si diferentes tratamientos muestran diferencias significativas en sus resultados o si, por el contrario, puede suponerse que sus medias poblacionales no difieren. Por lo que, a pesar de su nombre, es una técnica estadística que permite la comparación de las medias de una característica en varias poblaciones. Por último, las técnicas de reducción de dimensión permiten, como su nombre lo indica, reducir el número de variables en una muestra o la cantidad de mediciones realizadas, agrupando en componentes/clústers las variables/mediciones con características similares. Para lograr esto se usa el análisis de componentes principales y técnicas de clasificación como clúster jerárquico, algoritmo k-means y árboles de clasificación.

Todos los métodos antes descritos se abordarán y ejemplificaran a lo largo de este informe, no sin antes explicar en qué consiste

el set de datos que se analizará en cada sección.

II. Set de datos

El set de datos que se analizará a continuación es el conjunto de datos de Delft, utilizado para predecir el rendimiento hidrodinámico de los yates de vela a partir de las dimensiones y velocidad. Esta base de datos muestra el comportamiento de las siguientes variables:

1. Posición longitudinal del centro de flotabilidad, adimensional.
2. Coeficiente prismático, adimensional.
3. Relación longitud-desplazamiento, adimensional.
4. Relación haz-tiro, adimensional.
5. Relación longitud-haz, adimensional.
6. Número de Froude, adimensional.
7. Resistencia residual por unidad de peso de desplazamiento, adimensional.

I. Variables principales

De las variables anteriores se catalogan como principales, por su importancia, el coeficiente prismático, el número de Froude y la resistencia residual. El coeficiente prismático nos da una idea de cómo está diseñado el barco para "penetrar" en el agua, es decir, la facilidad para que el barco se ponga a planear y aumente su velocidad. Indica, además, la relación entre el volumen sumergido y el volumen definido por su manga máxima. Dicho de otra manera, indica el cociente entre el volumen sumergido y el volumen de la pieza a partir de la cual se ha podido "tallar" el casco. Cuanto menor sea este coeficiente más finos serán la popa y proa y, por tanto, mejor afrontarán las olas. El número de Froude relaciona el efecto de las fuerzas de inercia y las fuerzas de gravedad que actúan sobre un fluido, estas fuerzas están presentes en el accionar de las olas causadas por un barco al navegar, por ello, esta variable es de suma importancia para el rendimiento hidrodinámico de un buque. Con el número de Froude se puede predecir la resistencia al avance de los barcos, estimando la resistencia que estos presentan ante las olas, que depende de la resistencia de fricción (debida a la superficie mojada del casco) y la resistencia residual (debida a la formación de olas). Finalmente, la resistencia residual por unidad de peso de desplazamiento es causada por la presión que genera el casco al abrirse paso a través del agua, esta variable es de gran valor para los yates de vela en la etapa de diseño inicial, para evaluar el rendimiento del buque y estimar la potencia propulsora requerida.

III. Regresión lineal

Utilizaremos el método backward para realizar un modelo de regresión lineal que explique el comportamiento de la resistencia residual a partir del resto de las variables principales. Luego nuestro modelo comienza con todas las variables antes mencionadas, analicemos los resultados:

```
Call:
lm(formula = ResiduaryResistance ~ PrismCoeff + FroudeNumber,
    data = yatch_data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.823  -7.586  -1.732   5.922  31.519

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -13.993     12.409  -1.128    0.260
PrismCoeff    -18.597     21.827  -0.852    0.395
FroudeNumber  121.668      5.036  24.159   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.907 on 305 degrees of freedom
Multiple R-squared:  0.6571,    Adjusted R-squared:  0.6548
F-statistic: 292.2 on 2 and 305 DF,  p-value: < 2.2e-16
```

Podemos observar que la variable coeficiente prismático (*PrismCoeff*) ni el intercepto son significativos en el modelo, es decir, sus coeficientes no le aportan nada a este.
Si analizamos la matriz de correlación obtenemos:

```
                          ID  LongPosition    PrismCoeff LengthDispRatio
ID                1.00000000 -6.774668e-02  1.431321e-02   -2.875420e-02
LongPosition     -0.06774668  1.000000e+00 -8.610666e-03   -2.674076e-03
PrismCoeff        0.01431321 -8.610666e-03  1.000000e+00   -4.631607e-02
LengthDispRatio  -0.02875420 -2.674076e-03 -4.631607e-02    1.000000e+00
BeamDraughtRatio -0.05865198  2.928595e-03  3.394618e-01    3.768233e-01
LengthBeamRatio  -0.04870850 -3.369351e-03 -8.669450e-02    6.763646e-01
FroudeNumber      0.04533868 -1.457676e-20  3.462115e-20    3.456373e-21
ResiduaryResistance  0.05232963  1.930617e-02 -2.856912e-02   -2.967365e-03
                  BeamDraughtRatio LengthBeamRatio  FroudeNumber  ResiduaryResist
ID                    -5.865198e-02   -4.870850e-02  4.533868e-02    0.052329630
LongPosition           2.928595e-03   -3.369351e-03 -1.457676e-20    0.019306170
PrismCoeff             3.394618e-01   -8.669450e-02  3.462115e-20   -0.028569120
LengthDispRatio        3.768233e-01    6.763646e-01  3.456373e-21   -0.002967365
BeamDraughtRatio       1.000000e+00   -3.802223e-01 -1.396091e-20   -0.012421130
LengthBeamRatio       -3.802223e-01    1.000000e+00 -4.408595e-21   -0.001025470
FroudeNumber          -1.396091e-20   -4.408595e-21  1.000000e+00    0.810092224
ResiduaryResistance   -1.242113e-02   -1.025470e-03  8.100922e-01    1.000000000
```

Al analizar la matriz es fácil darse cuenta de que las variables coeficiente prismático y resistencia residual no están correlacionadas, de hecho, ninguna de las variables están correlacionadas entre sí, a excepción del número de Froude y resistenia residual. Como, en el caso de este modelo en particular, la variable dependiente *ResiduaryResistance* no está correlacionada con la variable independente *PrismCoeff*, debemos eliminarla. Por lo tanto tenemos un nuevo modelo sin la variable *PrismCoeff* que, además habíamos visto, no era significativa.

Luego llegamos a un modelo que tiene a *ResiduaryResistance* como variable dependiente y a *FroudeNumber* como variable independiente. Los resultados obtenidos con este modelo son los siguientes:

```
Call:
lm(formula = ResiduaryResistance ~ FroudeNumber, data = yatch_data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.240   -7.669   -1.726    6.404   32.154

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -24.484      1.534  -15.96   <2e-16 ***
FroudeNumber  121.668      5.034   24.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.903 on 306 degrees of freedom
Multiple R-squared:  0.6562,    Adjusted R-squared:  0.6551
F-statistic: 584.2 on 1 and 306 DF,  p-value: < 2.2e-16
```

Ahora tenemos que $\Pr(> |t|)$ de la variable independiente y del intercepto son menores que 0.05. El p-value también es menor que 0.05. Por lo tanto, podemos proceder a hacer un análisis de la presición del modelo.
El modelo resultante es:

$$\hat{ResiduaryResistance} = -24.484 + 121.668 * FroudeNumber$$

Se observa que el coeficiente del intercepto es mucho menor que el coeficiente del número de Froude, por lo tanto podemos decir que la mayor parte de la resistencia residual de los barcos está explicada a partir del número de Froude. El coeficiente del número de Froude es significativo al 0% y el del intercepto también es significativo al 0%. Analizando los valores de estos coeficientes, se puede afirmar que por cada aumento unitario en el número de Froude debemos esperar que la resistencia residual aumente en 121.668.

Pasando al análisis de los residuos tenemos que el R-cuadrado ajustado es 0.66, menor que 0.70, por lo que podemos decir que es un modelo bastante malo, no obstante, sigamos con el análisis. El error estándar es 8.9, no es tan pequeño, no es lo ideal. El p-value del estadígrafo F, como se había dicho, es menor que 0.05, lo que quiere decir que existe al menos una variable significativamente diferente a cero en el modelo.

Veamos ahora si el modelo cumple los supuestos.
Recordemos que los supuestos son:

1. Existe una relación lineal entre las variables dependientes e independientes.
2. Los errores $(e_1, ..., e_n)$ son independientes.
3. El valor esperado del error aleatorio e_i es cero $(E(e_i) = 0)$
4. La Varianza del error aleatorio es constante $(V(e_i) = \theta^2)$. Homocedasticidad.
5. Los errores además de ser independientes son idénticamente distribuidos y siguen distribución normal con media cero y varianza constante $(e_i\ N(0, \theta^2))$
6. Las variables independientes del modelo no están correlacionadas.

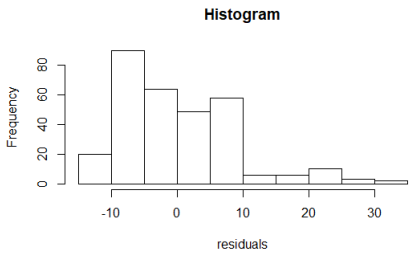
Los supuestos 1 y 6 se cumplen en el modelo en cuestión, pues el número de Froude esta correlacionado con la resistencia residual (ver matriz de correlación) y no hay más

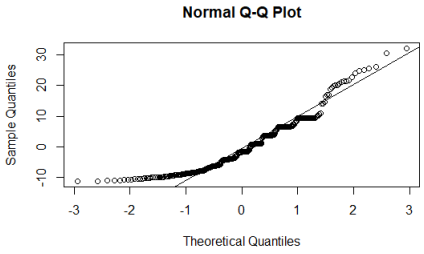
variables independientes que el mismo número de Froude.
Realicemos un análisis de los residuos para verificar si el resto de los supuestos se cumple:

Para analizar el cumplimiento del supuesto 3 debemos verificar si la media de los errores es cero y la suma de los errores es cero. Con ayuda de los comandos de R, **mean** y **sum**, se llega a que, en efecto, este supuesto se cumple.

```
> mean(residuals)
[1] -7.139401e-17
> sum(residuals)
[1] -2.187139e-14
```

Para verificar el cumplimiento del supuesto 5 debemos comprobar si los errores están normalmente distribuidos. El histograma de residuos y el gráfico QQ-plot son formas de evaluar visualmente si los residuos siguen una distribución normal. Por tanto, buscamos que el histograma tenga forma de campana y en el QQ-plot que la mayoría de los puntos de los residuos se encuentren sobre la recta o muy cercana a ella. Con ayuda de R construimos los gráficos antes mencionados para este modelo:





Parece que los residuos no siguen una distribución normal, comprobémoslo con el test de Shapiro-Wilk, con ayuda de R.

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.90797, p-value = 8.997e-13
```

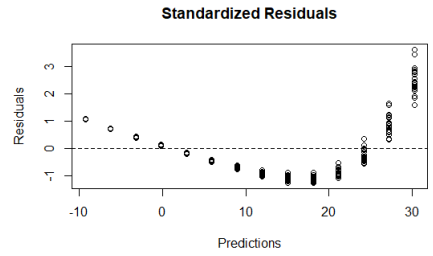
Como se observa, el p-valor del test de Shapiro-wilk es menor que 0.05, luego se rechaza la hipótesis nula por lo que los errores no siguen una distribución normal. Ya habíamos visto que este modelo era bastante malo, y ahora no cumple uno de los supuestos, por lo tanto ya podemos desecharlo. Sin embargo, analicemos el resto de los supuestos.

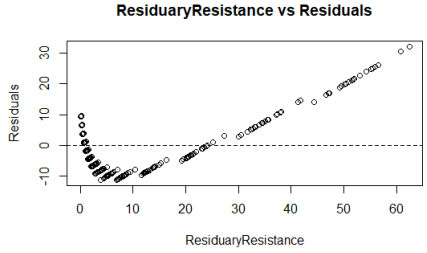
La prueba Durbin-Watson se usa para probar si los residuos son independientes. La hipótesis nula de esta prueba es que los errores son independientes.

```
        Durbin-Watson test

data:  backward
DW = 0.50842, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Como el p-valor de esta prueba es menor que 0.05 se rechaza la hipótesis nula, por lo que podemos afirmar que tampoco se cumple el supuesto 2.

Para probar el supuesto 4 de la Homocedasticidad podemos graficar los residuos como se muestra a continuación:





Con estos gráficos se comprueba que estos puntos no siguen una franja, por lo que es muy probable que este supuesto tampoco se cumpla. Utilicemos la prueba de Breusch-Pagan, que se utiliza para determinar la heterocedasticidad en un modelo de regresión lineal, para verificar lo anterior.

```
        studentized Breusch-Pagan test

data:  backward
BP = 62.16, df = 1, p-value = 3.166e-15
```

Como el p-valor de esta prueba es menor que 0.05 se rechaza la hipótesis nula por lo que podemos afirmar que se cumple la heterocedasticidad. Por lo que el supuesto de Homocedasticidad no se cumple.

Se concluye que para estos datos no existe un modelo de regresión lineal que se ajuste a ellos.

No podemos dejar de mencionar que, dado el análisis de la matriz de correlación realizado en esta sección, no es posible utilizar otra combinación de variables para realizar otro modelo de regresión, ya que las variables no tienen correlación entre sí, a excepción de las que se analizaron. El único modelo que se podría hacer con estos datos fue el que analizamos anteriormente, ya que el resto de combinaciones de modelos posibles no cumple, a priori, el supuesto 1.

El análisis de correlación también se aborda de una forma mejor explicativa en la sección V, subsección I del presente informe.

IV. ANOVA

Realicemos un análisis de ANOVA a fin de investigar si el número de Froude afecta la resistencia residual de los barcos veleros. Además, debemos considerar como factor secundario el coeficiente prismático de estos. Es decir, tenemos la variable de estudio *ResiduaryResistance*, y es claro que el número de Froude se puede ver como tratamiento y el coeficiente prismático como bloque.

Luego el problema que acabamos de plantear responde al siguiente modelo estadístico de bloques:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Donde Y_{ij} es la medición que corresponde al tratamiento i y al bloque j, en este caso la resistencia residual de los barcos; μ es la media global poblacional; α_i es el efecto debido al tratamiento i, en este caso los distintos números de Froude; β_j es el efecto debido al bloque j, en este caso los distintos coeficientes prismáticos, y e_{ij} es el error aleatorio atribuible a la medición Y_{ij}.

La hipótesis que debemos formular es la siguiente:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_{14} = \mu$$
$$H_1 : \mu_i \neq \mu_j \quad \text{para algún} \quad i \neq j$$

La cual se puede reescribir de forma equivalente como:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_{14} = 0$$
$$H_1 : \alpha_i \neq 0 \quad \text{para algún} \quad i$$

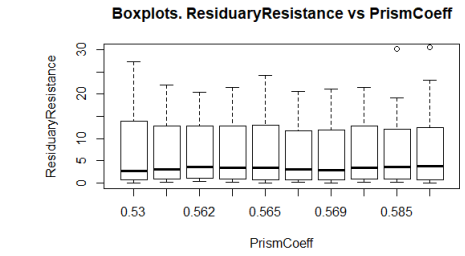
Primero deberíamos acomodar los datos para poder trabajar con ellos. Lo que buscamos es tenerlos de la siguiente forma:

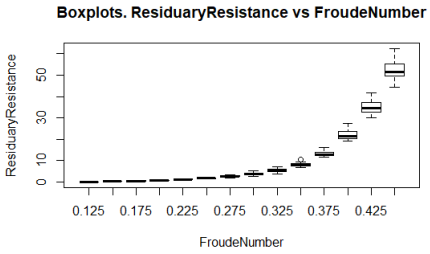
FroudeNumber	PrismCoeff	ResiduaryRes
f1	p1	r1
f1	p2	r2
f1	p3	r3
⋮	⋮	⋮
f1	p10	r10
f2	p1	r11
⋮	⋮	⋮
f14	p10	r308

Como tenemos 14 números de Froude distintos y 10 coeficientes prismáticos distintos, entonces ponemos en la primera columna secuencialmente 10 veces el primer número de Froude, luego el segundo y así sucesivamente,

para poder listar en la segunda columna los valores de los coeficientes prismáticos en cada caso y, por último, listar la resistencia residual de cada barco.

Luego necesitamos comparar las medias de los 14 niveles del factor y las medias de los 10 niveles del bloque. Para esto realizamos gráficos de cajas con las medias de cada uno.





Podemos observar que las medias de la resistencia residual de los coeficientes prismáticos son bastante cercanas, oscilando entre 2 y 3 aproximadamente, por lo que es posible que el coeficiente prismático no tenga efecto sobre la resistencia residual de los veleros.

Por otro lado, el gráfico de la resistencia residual y el número de Froude muestra mucha diferencia en las medias, podemos apreciar que empieza en valores muy cercanos a 0 y, a medida que el número de Froude aumenta, obtenemos valores de resistencia residual promedio hasta de 50. Por lo tanto, es muy probable que el número de Froude tenga efecto sobre la resistencia residual de los veleros.

Si realizamos un análisis de ANOVA, obetenemos:

```
             Df Sum Sq Mean Sq F value Pr(>F)
PrismCoeff    1     58      58   0.726  0.395
FroudeNumber  1  46306   46306 583.657 <2e-16 ***
Residuals   305  24198      79
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notamos que en el caso del coeficiente prismático el p-valor es mayor que 0.05, luego no podemos rechazar H_0 por lo que se acepta que este no influye en la resistencia residual de los barcos.

En el caso del número de Froude, como el p-valor es menor que la significación prefijada $\alpha = 0.05$, entonces se rechaza H_0 y se acepta que al menos un par de números de Froude tienen una resistencia promedio diferente, es decir, influyen sobre la resistencia residual de los barcos.

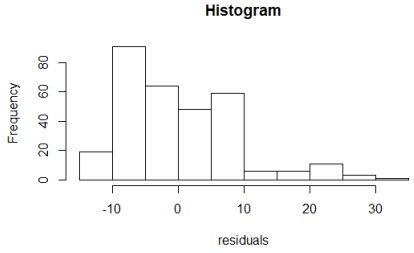
Verifiquemos si el modelo cumple los supuestos. Recordemos que estos son:

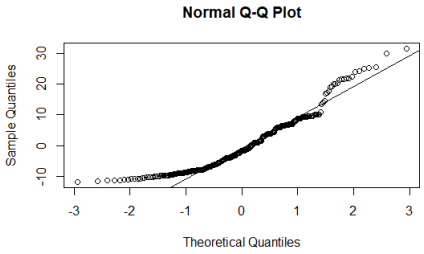
1. Los e_{ij} siguen una distribución normal con media cero.

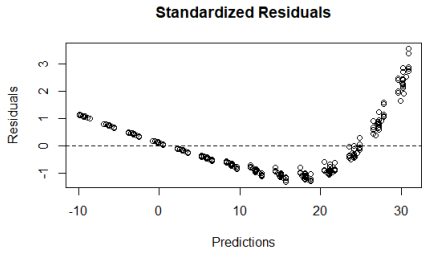
2. Los e_{ij} son independientes entre sí.

3. Los residuos de cada tratamiento tienen la misma varianza θ^2.

Para verificar el cumplimiento del supuesto 1 debemos comprobar si los errores están normalmente distribuidos. El histograma de residuos y el gráfico QQ-plot son formas de evaluar visualmente si los residuos siguen una distribución normal. Por tanto, buscamos que el histograma tenga forma de campana y en el QQ-plot que la mayoría de los puntos de los residuos se encuentren sobre la recta o muy cercana a ella. Con ayuda de R construimos los gráficos antes mencionados para este modelo:







Parece que los residuos no siguen una distribución normal, utilizaremos el test de Shapiro-Wilk, con ayuda de R, para comprobarlo.

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.91326, p-value = 2.424e-12
```

Como se observa, el p-valor del test de Shapiro-wilk es menor que 0.05, luego se rechaza la hipótesis nula por lo que los errores no siguen una distribución normal. Podemos desechar este modelo, no obstante, analicemos el resto de los supuestos.

La prueba Durbin-Watson se usa para probar si los residuos son independientes. La hipótesis nula de esta prueba es que los errores son independientes.

```
        Durbin-Watson test

data:  anova
DW = 0.50717, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Como el p-valor de esta prueba es menor que 0.05 se rechaza la hipótesis nula, por lo que podemos afirmar que tampoco se cumple el supuesto 2.

Por último, para probar el supuesto 3 de la Homocedasticidad podemos graficar los residuos como se muestra a continuación:

Los puntos no forman una franja, parece ser que no tienen varianza constante. Utilicemos la prueba de Bartlett para confirmarlo.

```
        Bartlett test of homogeneity of variances

data:  residuals and FroudeNumber
Bartlett's K-squared = 329.23, df = 13, p-value < 2.2e-16
```

Como el p-valor de esta prueba es menor que 0.05 se rechaza la hipótesis nula por lo que el supuesto de Homocedasticidad no se cumple.

De forma análoga a como se desarrolló en esta sección, se realizaron varios análisis de ANOVA considerando otras variables, pero por desgracia ninguno resultó válido. Solo se refleja en este informe el realizado con las variables principales definidas previamente, cuyo modelo analizamos en esta sección.

V. Reducción de dimensión

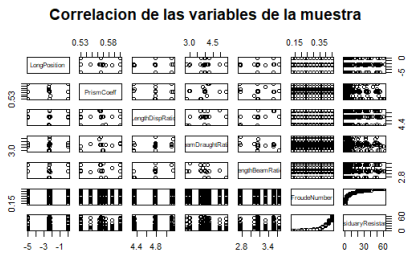
Realizaremos en esta sección varias técnicas de reducción de dimensión a nuestro set de datos, ya sea en cuanto la cantidad de variables como en el tamaño de la muestra.

I. Análisis de componentes principales

El objetivo del ACP es reducir dimensión agrupando variables en componentes principales incorrelacionadas entre sí. En el

caso de nuestro set de datos, como ya vimos en la sección II, tenemos 7 variables.

Lo primero que debemos estudiar es la correlación de nuestra muestra. Para ello graficamos los datos para ver si existe algún tipo de correlación entre ellos.



Son demasiadas variables para analizar de forma visual. Debemos acudir entonces a la matriz de correlación, analizada en la sección III, pero esta vez trabajaremos con esta matriz en forma gráfica.

```
                   LP  P LD  B LB  F  R
LongPosition        1
PrismCoeff             1
LengthDispRatio          1
BeamDraughtRatio         .  .  1
LengthBeamRatio          ,  .  1
FroudeNumber                      1
ResiduaryResistance               + 1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

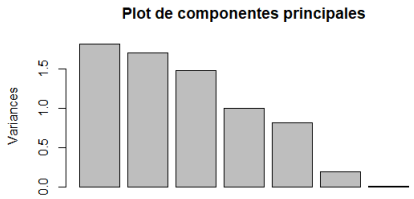
Como se puede apreciar no es una matriz altamente correlacionada, y solo tiene una marcada correlación el número de Froude y la resistencia residual, denotado con el símbolo de $+$.

En este caso ya las variables son independientes, por lo que este análisis solo serviría para reducir dimensión. Así que podemos proseguir a realizar el ACP, obteniendo los siguientes resultados:

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7
Standard deviation     1.3459 1.3027 1.2171 1.0002 0.9018 0.43494 0.08665
Proportion of Variance 0.2588 0.2424 0.2116 0.1429 0.1162 0.02702 0.00107
Cumulative Proportion  0.2588 0.5012 0.7128 0.8557 0.9719 0.99893 1.00000
```

Ahora debemos escoger nuestras componentes principales. Según el criterio de Kaiser nos quedaríamos con las cuatro primeras componentes, explicando estas el 86% de la muestra. Pero no hacemos mucho reduciendo a cuatro componentes dado que tenemos 7 variables. Por tanto, como el valor propio de la cuarta componente está a solo 0.0002 por encima de 1, y como las tres primeras componentes explican el 71% de la muestra, entonces nos quedaremos con PC1, PC2 y PC3.

Podemos, además, ver las componentes de forma visual para ratificar nuestra decisión. Si graficamos estas componentes obtenemos:



Se puede observar que a partir de la cuarta componente hay una caída más pronunciada. Con las tres primeras se explica una buena parte, sería un 71%, que si bien no es lo ideal, no está mal.

Ahora analizamos la matriz de valores propios y así sabremos qué variable es importante para cada componente y en qué medida.

```
                          PC1          PC2          PC3
LongPosition       -0.0170306151  0.00462287 -0.005086103
PrismCoeff          0.0440094523  0.20587786  0.498048809
LengthDispRatio    -0.0002153088 -0.63677403  0.429514872
BeamDraughtRatio    0.0395797935  0.15206298  0.747360320
LengthBeamRatio    -0.0224614444 -0.72662849 -0.076559704
FroudeNumber       -0.7049519612  0.02396851  0.047321822
ResiduaryResistance -0.7062186175  0.02061973  0.028112249
```

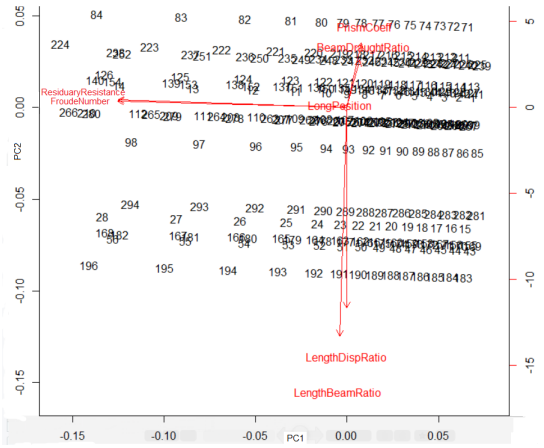
Comencemos por la primera componente. Tomamos el mayor valor propio modularmente, 0.71, y lo dividimos entre 2, dando como resultado 0.36. Luego todo valor propio cuyo módulo esté por encima de 0.36 en la columna de la PC1 nos dará las variables que conforman esta componente. Por tanto, la interpretación sería que la primera componente está caracterizada por barcos que tienen bajo número de Froude y baja resistencia residual, o sea, son yates que pueden ignorar casi por completo la resistencia por olas en su navegación.

Siguiendo el mismo análisis para la

segunda componente tenemos que el mayor valor abosluto de valores propios es 0.73, dividido entre 2 sería 0.37, por tanto, en esta componente los barcos tienen baja relación longitud - desplazamiento y también una baja relación longitud - haz, o sea, son yates que están hechos para "planear en el agua", es decir, alcanzan una gran velocidad.

En la tercera componente tenemos que el mayor valor propio modular es 0.75, dividido entre 2 es 0.38, entonces esta componente está representada por veleros que tienen un alto coeficiente prismático, una elevada relación longitud - desplazamiento y una elevada relación haz - tiro, por lo que son yates veleros que no están diseñados para alcanzar una velocidad muy alta.

Por último, podemos ver un biplot del ACP.



Primero se puede ver que las variables que representan la PC1 son número de Froude y resistencia residual de forma negativa pues tienen un ángulo obtuso. La PC2 tiene como variables a la relación longitud desplazamiento y la relación longitud - haz de forma negativa. Mientras más larga sea la recta que define a una variable más representada está la variable en la componente y más importante será dentro de la misma. También podemos observar algunos yates que están cerca de cada componente.

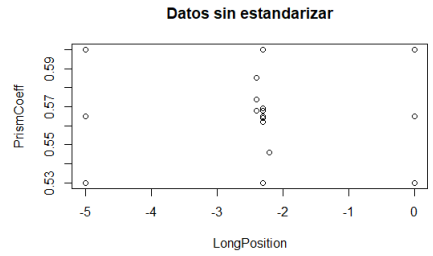
II. Clúster jerárquico

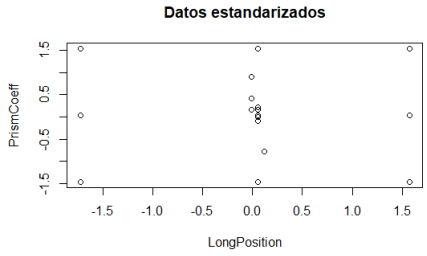
Entre las técnicas para la reducción de dimensión se encuentran las técnicas de clasificación, que tienen como objetivo agrupar individuos de grandes poblaciones para reducir la dimensión de la muestra.

A partir de esta subsección nos centraremos en aplicar técnicas para reducir el tamaño de la muestra.

El primer paso, independientemente del procedimiento a seguir, será estandarizar los datos para evitar errores en la clasificación por cuestiones de variabilidad en las unidades de medidas. Para esto tipificaremos cada una de las 7 variables. Obteniendo $Z_i = \frac{X_i - \mu_i}{\sigma_i}$.

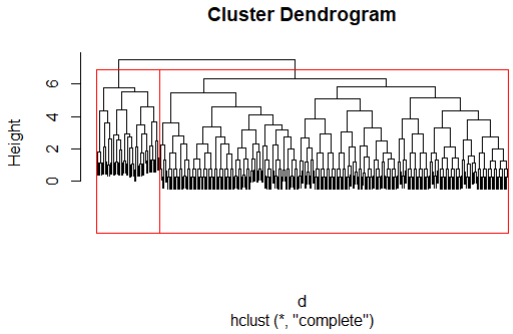
Para estandarizar usamos la función de R **scale** y, como se observa a continuación, los datos estandarizados tienen el mismo comportamiento que los que no fueron estandarizados, la diferencia está en la escala de las mediciones que es similar.



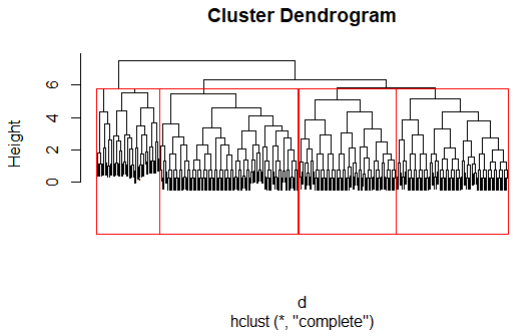


Una vez escaladas las mediciones construimos la matriz de distancias que será simétrica, para la cual utilizamos la distancia euclidiana. Luego realizamos un clúster jerárquico con el método de ajuste completo y la matriz de distancias. Si graficamos el ajuste obtendremos el Dendograma del clúster

jerárquico.

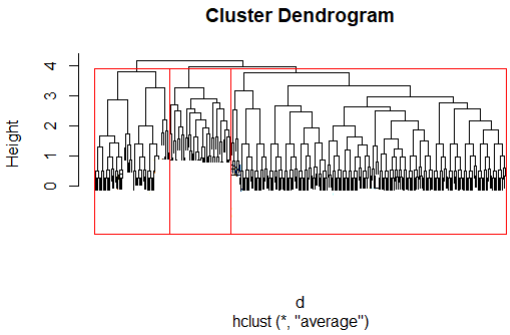


A partir de este gráfico podemos determinar la cantidad de clústers haciendo cortes horizontales a alturas determinadas. Por ejemplo, si cortamos a altura poco mayor que 6, obtendremos dos clústers, como pudimos observar en la figura, uno que contiene a un pequeño grupo de yates y otro que contiene al resto. Pero no tiene sentido, a pesar de que un pequeño grupo de veleros se comporte de manera distinta, analizar el resto de los barcos como uno solo, pues si bajamos la altura vemos que ese gran clúster se divide en clústers más pequeños. Tampoco tendría sentido realizar un análisis por debajo de la altura 4 pues tendríamos muchos clústers de elementos muy parecidos. En este caso decidimos quedarnos con 4 clústers como se muestra a continuación.



Por desgracia los clústers son demasiado grandes y no quedan legibles los barcos que forman parte de cada uno. No obstante, podemos hacernos una idea de ese pequeño grupo de yates que se comporta de manera distinta. Si analizamos los datos, nos percatamos de que existen algunos veleros que tienen una resistencia residual bastante alta y una posición longitudinal muy baja, motivo por el cual se comportan de forma distinta a la mayoría de los yates. Estos barcos podrían formar parte del primer clúster.

Todas las técnicas de clúster jerárquico no dan los mismos resultados, veamos qué sucede si, en vez de utilizar un ajuste completo, ajustamos por las medias. En este caso los ajustes son las medidas de asociación entre las variables.



Podemos notar que, a diferencia del ajuste completo, si hacemos un corte muy elevado obtenemos tres clústers de yates que se comportan de manera diferente, dos relativamente pequeños y uno que contiene a la mayoría de los yates.

III. K-means

Habiendo realizado la clasificación por el método jerárquico, utilicemos ahora el algoritmo k-means. El problema más grande al aplicar este algoritmo es que necesitamos tener una idea de cuántos clústers tendremos, en este caso partiremos del número de clústers encontrados en el algoritmo anterior. Es decir, comenzaremos con 4 clústers.

```
K-means clustering with 4 clusters of sizes 26, 158, 65, 59

Cluster means:
  LongPosition  PrismCoeff LengthDispRatio BeamDraughtRatio LengthBeamRatio FroudeNumber
1  0.021026821 -0.64990813      -1.7728664      -1.90045871     -0.2291071   -0.1238332
2  0.001787090 -0.23195299       0.4978478      -0.02883204      0.4599542   -0.4263625
3 -0.004898556 -0.09040764       0.1841261       0.05239082      0.1009199    1.3278731
4 -0.008655108  1.00716408      -0.7548070       0.85698281     -1.2419622   -0.2665562
  ResiduaryResistance
1          -0.2146249
2          -0.4953359
3           1.6215041
4          -0.3653297

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 2
 [44] 2 2 2 2 2 2 2 2 2 2 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 3 2 2 2
 [87] 2 2 2 2 2 2 2 2 2 2 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2
[130] 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 2 2 2 2
[173] 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 2 2 2 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 1 1 1 1 1
[216] 1 1 1 1 1 1 1 1 3 4 4 4 4 4 4 4 4 4 4 4 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 3 2 2 2 2 2 2 2
[259] 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 4 4 4 4 4 4 4 4
[302] 4 4 4 4 4 4 3

Within cluster sum of squares by cluster:
[1]  53.75841 626.02373 338.54746 263.42216
 (between_SS / total_SS =  40.4 %)
```

Al aplicar la función k-means y comprobar su resultado obtenemos un vector de clústers que nos dice en qué clúster está cada barco y cuántos elementos hay. En este caso tenemos 4 clústers con 26, 158, 65 y 69 barcos respectivamente. Notemos que sigue habiendo un clúster con un pequeño grupo de barcos que se comportan de forma distinta.

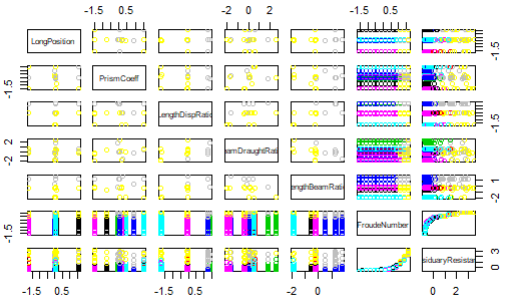
También tenemos las medias de los elementos de cada variable por cada clúster y, una de las informaciones más importantes que podemos obtener, es la medida de similitud entre los elementos de cada clúster, en este caso sería un 40.4%, que no es bueno. Debemos, al menos, duplicar la cantidad de clústers para elevar un poco ese porcentaje.

Repetimos el algoritmo esta vez con 8 clústers.

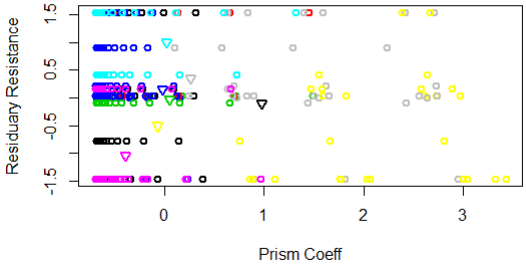
```
K-means clustering with 8 clusters of sizes 55, 25, 25, 74, 25, 46, 27, 31

Cluster means:
  LongPosition  PrismCoeff LengthDispRatio BeamDraughtRatio LengthBeamRatio FroudeNumber
1   0.97924911 -0.10031615      -0.04203148       0.15173820      -0.2774945   -0.3714996
2  -1.73020701  0.81853294      -0.05309618       0.33415570      -0.2291071   -0.1832732
3   0.05406897 -0.02474725       0.64872183       2.22764938      -1.0081444   -0.1832732
4  -0.01469442  0.15196699       0.87742786      -0.27127320       1.2088927   -0.4217023
5   0.02234851  1.00402023      -1.73493033       0.29840187      -1.8646016   -0.1832732
6  -0.38984510 -1.04007363      -0.96019527      -1.18506485      -0.2194646   -0.3068910
7  -0.06341422 -0.50678522      -0.49369364      -0.20971871      -0.4919523    1.4447209
8   0.26511236  0.35010550       0.75365800       0.01286552       0.8622112    1.3062407
  ResiduaryResistance
1          -0.4506266
2          -0.2972039
3          -0.2848033
4          -0.4936074
5          -0.3035889
6          -0.4262029
7           2.1458198
8           1.4554682

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 8 7 7 4 4 4 4 4 4 4 4 4 4 4 4 8 8 8 3 3 3 3 3 3 3 3 3 3 3 3 7 4
 [44] 4 4 4 4 4 4 4 4 4 4 8 8 8 5 5 5 5 5 5 5 5 5 5 5 5 7 7 6 6 6 6 6 6 6 6 6 6 6 6 7 7 3 3
 [87] 3 3 3 3 3 3 3 3 3 3 8 8 8 4 4 4 4 4 4 4 4 4 4 4 4 8 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 7 7 7 1 1 1
[130] 1 1 1 1 1 1 1 1 8 7 7 2 2 2 2 2 2 2 2 2 2 2 7 7 4 4 4 4 4 4 4 4 4 4 4 4 8 8 8 8 4 4 4 4 4
[173] 4 4 4 4 4 4 4 4 8 8 8 4 4 4 4 4 4 4 4 4 4 4 8 8 8 8 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 6 6 6 6 6
[216] 6 6 6 6 6 6 6 7 7 1 1 1 1 1 1 1 1 1 1 1 1 8 8 8 2 2 2 2 2 2 2 2 2 2 2 2 2 2 7 1 1 1 1 1 1 1
[259] 1 1 1 1 1 1 7 7 7 6 6 6 6 6 6 6 6 6 6 7 7 7 4 4 4 4 4 4 4 4 4 4 4 8 8 8 8 5 5 5 5 5 5 5 5
[302] 5 5 5 5 5 5 5 7

Within cluster sum of squares by cluster:
[1] 129.45501  42.38666  60.20443 219.75298  35.92640 146.46868 116.78173 116.49170
 (between_SS / total_SS =  59.6 %)
```

Esta vez obtuvimos un porcentaje del 60% que no está mal. Y tampoco es exagerada la cantidad de clústers, teniendo en cuenta que estamos trabajando una muestra de cardinalidad igual a 308.

Si graficamos el resultado de k-means obtendremos una matriz de gráficos.



La cual resulta imposible de analizar, por lo que si queremos analizar las relaciones de forma visual entre variables debemos analizarlas dos a dos, por ejemplo, resistencia residual y coeficiente prismático.



Podemos observar los distintos clústers representados con colores distintos. Además se han representado los centros de cada uno con triángulos.

IV. Árboles de clasificación

En esta sección construiremos un árbol de clasificación para nuestros datos. Los árboles de clasificación son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.

Hay distintas maneras de obtener estos árboles, la que usaremos en esta ocasión es conocida como CART: Classification And Regression Trees. Esta es una técnica

de aprendizaje supervisado. Tenemos una variable objetivo (dependiente) y nuestra meta es obtener una función que nos permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.

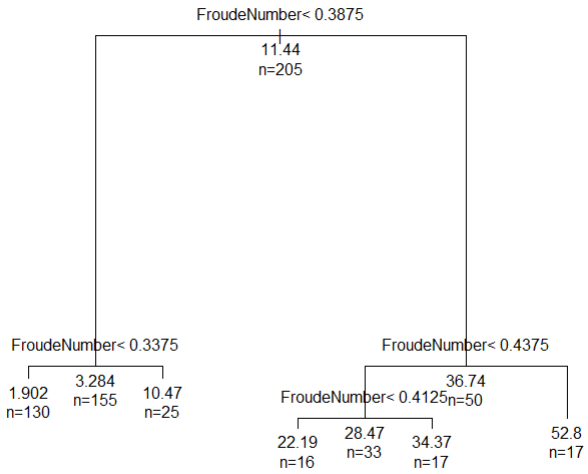
En este caso la idea es ser capaz de predecir la resistencia residual de un yate velero basado en el resto de las variables que se brindan. Como nuestros datos son continuos nuestro árbol de clasificación se convierte en un árbol de regresión. La implementación particular de CART que usaremos es conocida como Recursive Partitioning and Regression Trees o RPART. De ahí el nombre del paquete que utilizaremos. De manera general, lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo.

Lo primero es escoger los conjuntos de entrenamiento y de prueba, que serán conjuntos disjuntos, en este caso se toman dos tercios de la población para el conjunto de entrenamiento y el tercio restante servirá para probar el árbol y calcular el error de clasificación.

Usamos la función de R **rpart** de la librería *rpart* para entrenar nuestro modelo y obtenemos el siguiente resultado.

```
n= 205

node), split, n, deviance, yval
      * denotes terminal node

 1) root 205 45190.69000 10.536680
   2) FroudeNumber< 0.3875 160  2400.19700  3.456812
     4) FroudeNumber< 0.3375 132   425.53920  1.930000 *
     5) FroudeNumber>=0.3375 28   216.30010 10.654640 *
   3) FroudeNumber>=0.3875 45  6255.25900 35.709560
     6) FroudeNumber< 0.4375 33  1606.49700 29.765760
      12) FroudeNumber< 0.4125 14    43.76509 22.170710 *
      13) FroudeNumber>=0.4125 19   160.08370 35.362110 *
     7) FroudeNumber>=0.4375 12   276.83050 52.055000 *
```

Lo anterior muestra el esquema de nuestro árbol. Cada inciso nos indica un nodo y la regla de clasificación que le corresponde. Siguiendo estos nodos, podemos llegar a las hojas del árbol, que corresponde a la clasificación de nuestros datos. Todo lo anterior resulta mucho más claro si lo visualizamos, así que creamos la siguiente gráfica usando nuestro modelo.



En este gráfico, cada una de las intersecciones representa un nodo de nuestro árbol, con su regla de clasificación. Tengamos en cuenta que si bajamos por su rama izquierda estamos asumiendo que la regla se cumple. De esta forma si tenemos un velero con un número de Froude menor que 0.34 nuestro árbol predice que la resistencia residual de este será de 1.9. Si el número de Froude es menor que 0.38 y mayor que 0.34 ($0.34 \leq FroudeNumber < 0.39$) la resistencia residual del barco será de 10.47. Por otro lado, si el número de Froude es menor que 0.41 tendríamos una resistencia residual de 22.19. Si $0.41 \leq FroudeNumber < 0.44$ entonces la resistencia residual sería igual a 34.37. Por último, si $FroudeNumber \geq 0.44$ implica que la resistencia residual del velero tendría un valor de 52.8.

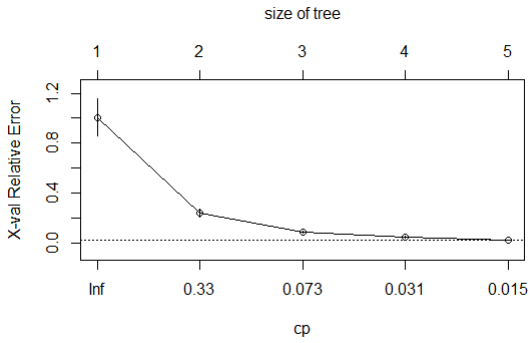
Posteriormente podaremos el árbol de regresión y, para ello, debemos encontrar el valor óptimo a utilizar para *cp*, el parámetro de complejidad que le pasamos a la función **rpart**, que conduce al error de prueba más bajo. Debemos tener en cuenta que el valor óptimo de *cp* es el que conduce al *xerror* más bajo en la imagen siguiente, que representa el error en las observaciones de los datos de validación

cruzada.

```
Regression tree:
rpart(formula = ResiduaryResistance ~ ., data = data[sub, ],
    cp = 0.01, maxdepth = 3)

Variables actually used in tree construction:
[1] FroudeNumber

Root node error: 45191/205 = 220.44

n= 205

        CP nsplit rel error   xerror      xstd
1 0.808468      0  1.000000 1.015327 0.1376423
2 0.096744      1  0.191532 0.205050 0.0306190
3 0.038910      2  0.094788 0.101651 0.0108200
4 0.031038      3  0.055878 0.070522 0.0095121
5 0.010000      4  0.024840 0.027615 0.0039308
```

Podemos notar que el valor que corresponde al *xerror* más pequeño es $cp = 0.01$, el mismo que utilizamos para crear el árbol inicial.

Otra forma de ver el *cp* óptimo es graficar el conjunto de posibles podas de costo-complejidad de un árbol. Para las medias geométricas de los intervalos de valores *cp* para los que una poda es óptima, se ha realizado una validación cruzada en la construcción inicial por **rpart**.



Una buena opción *cp* para podar es, a menudo, el valor más a la izquierda para el que la media se encuentra por debajo de la línea horizontal. Se puede observar que esto ocurre alrededor del mismo valor de *cp* utilizado.

Ahora bien, necesitamos ser más sistemáticos para indagar qué tan bien hace predicciones nuestro modelo. Para ello generamos un vector de predicciones, utilizando nuestro conjunto de prueba, que contendrá los valores predichos por el modelo que hemos entrenado. Posteriormente cruzamos la predicción con los datos reales de nuestro conjunto de prueba para generar una matriz de confusión a partir de la cual se calculará el error.

```
> error
[1] 0.9902913
```

Este valor de error es bastante elevado, lo que quiere decir que nuestro árbol no es efectivo. Este resultado es consistente con el análisis de regresión que hicimos en la sección III, donde no obtuvimos ningún modelo de regresión válido.

VI. Conclusiones

A lo largo de este informe se analizaron conjuntos de datos, pudiendo caracterizar su comportamiento gracias a las técnicas de regresión lineal, ANOVA y reducción de dimensión.