UNIVERSIDAD DE EL SALVADOR
FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE
DEPARTAMENTO DE MATEMÁTICA

Licenciatura en Estadística

Control Estadistico del Paquete R

"UNIDAD SEIS"
Práctica 24 - Análisis de Varianza (ANOVA).

Alumna:
Martha Yoana Medina Sánchez

Fecha de elaboración
Santa Ana - 27 de noviembre de 2015

En muchos casos prácticos existe la necesidad de realizar o de hacer comparaciones entre la media de una característica en diferentes niveles o grupos bajo un nivel de significancia $alfa$ prefijado; en estos casos el **ANÁLISIS DE VARIANZA** es la técnica estadística más adecuada para poder llevar a cabo simultáneamente dichas comparaciones. El Análisisde Varianza, descompone la variabilidad total (VT) de la variable de interés en dos fuentes de variabilidad mutuamente independientes: una debida a los efectos de los grupos o variabilidad explicada por los grupos (VE) y otra debida a los errores (perturbaciones) o variabilidad no explicada (VNE). Es común, en los Diseños de Experimentos llamar a cada uno de esos niveles o grupos con el nombre de "tratamientos". Esta técnica tiene como objetivo identificar la importancia de los diferentes grupos en el estudio y determinar la influencia de ellos sobre la variable de interés.

Si nuestra variable de interés,la cual representaremos por y, es continua una manera muy conveniente de representar las observaciones es por medio de la siguiente ecuación:

$$y_{ij} = \mu_i + u_{ij}$$

Donde:

- y_{ij}: Representa la j-ésima observación correspondiente al i-ésimo grupo.
- μ_i: Representa la media del i-ésimo grupo (o tratamiento)
- u_{ij}: Representa un componente de error aleatorio, llamado perturbaciones, que incorpora todas las demás fuentes de variabilidad del experimento (no incluidas en los grupos o tratamientos).

A la ecuación anterior, se le conoce con el nombre de "modelos en medias". Una forma alternativa y mucho más interesante de escribirlo es considerando el caso en que $t_i = \mu_i$ - μ, por lo que el modelo se convierte en: $y_{ij} = \mu + t_i + e_{ij}$

El cual recibe el nombre de "modelos de efectos"; pues el término t_i representa el efecto del grupo i-ésimo (o tratamiento i-ésimo). Y debe cumplirse que la sumatoria de t_i=0.

Las perturbaciones como ya se mencionó representan la variabilidad intrínseca del experimento y supondremos que verifican las hipótesis siguientes (en caso de duda hay que contrastarlas):

Hipótesis:

- El promedio de las perturbaciones escero, es decir, se cumple que:
  E[u_{ij}] = 0; para todo i,j.
- La varianza de las perturbaciones es constante, es decir, se cumple que:
  var(u_{ij}) = $sigma^2$; para todo i, j.
- La distribución de las perturbaciones debe ser normal, es decir se cumple que:
  u_{ij} = N(0; $sigma^2$) =; para todo i, j.
- Las perturbaciones son independientes, es decir se cumple que:
  cov(u_{ij}; $u_{i' j'}$)=0; para todo i distinto de $i^{'}$, y para todo j distinto de $j^{'}$

Las cuatro hipótesis anteriores sobre las perturbaciones que son las hipótesis básicas del modelo, pueden resumirse en (IID significa que son variables aleatorias independientes e idénticamente distribuidas):

$u_{ij} = NIID$ N(0; $sigma^2$) =; para todo i,j.

Si por ejemplo tenemos una única característica de interés y existen k grupos (o tratamientos) en los cuales se mide ésta, podría estarse interesado en probar la igualdad de las media en cada una de los grupos (tratamientos).

La hipótesis a probar son:
H_o: $\mu_1 = \mu_2 = \mu_3$ ... μ_k
H_1: μ_i distinto μ_j; para al menos un par i distinto de j

Las cuales en términos de los efectos de grupos, son equivalente a las siguientes hipótesis:
H_o: $t_1 = t_2$ ... $t_k = 0$
H_1: t_i distinto 0; para al menos un i

En el caso más general, como nunca podemos estudiara toda la población, sino lo que tenemos es una muestra aleatoria de ella (en realidad es una muestra aleatoria de cada grupo); sucederá que:

- k es el número de grupos de interés (tratamientos).
- n_i es el número de observaciones pertenecientes al grupo i
- N = a la sumatoria de n_i número total de observaciones.

Con dicha muestra debemos contrastarlas hipótesis anteriores y estimar cada uno de los parámetros del modelo. No resulta difícil verificar utilizando el método de máxima verosimilitud que el modelo estimado para los datos es:

$\hat{y}_{ij} = \hat{\mu} + \hat{t}_i + \hat{u}_{ij}$

Donde:

- $\hat{\mu} = \tilde{y}_{..}$
- $\hat{t}_i = \tilde{y}_{i.}$ - $\tilde{y}_{..}$
- $\hat{u}_{ij} = y_{ij}$ - $\tilde{y}_{i.}$

Y se tendrán las siguientes medidas de interés:

- $\tilde{y}_{i.}$ es el promedio de la característica de interés en el grupo i.
- $\tilde{y}_{..}$ es la media general de la característica de interés.

El Análisis de Varianza establece que se debe cumplir la siguiente relación (al ser cada uno de las fuentes ortogonales entre sí):

$VT = VE + VNE$

Donde:

- VT es la variabilidad total del experimento.
- VE es la variabilidad explicada por los grupos o tratamientos.
- VNE es la variabilidad no explicada o residual.

Para poder contrastar simultáneamente la igualdad de las k medias, se hace uso de lo siguiente:

Grupos o Tratamientos:

- Sumas de Cuadrados: VE.
- Grados de Libertad: $K - 1$
- Medias de Cuadrados: $MCE = VE \ / \ K - 1$
- F_o: $F_o = MCE/MCNE$

Error o perturbaciones:

- Sumas de Cuadrados: VNE.
- Grados de Libertad: $N - K$
- Medias de Cuadrados: $MCNE = VNE \ / \ N - 1$

Total:

- Sumas de Cuadrados: $VT = VE + VNE$
- rados de Libertad: $N - 1$

De tal modo que la hipótesis nula se rechaza (a un nivel de confianza del 100(1 - $alfa$) %) si

$F_o > F_{alfa,(K-1),(N-K)}$

- Ejemplo 1:

El Ministerio de Educación está interesado en implementar tres programas de estudio; con el objetivo de medir la habilidad de lectura en los alumnos. Para ello, se eligen alumnos del sexto grado de un Colegio de San Salvador, 27 alumnos fueron asignados al azar, a cada uno de los tres grupos. Se utilizó un programa diferente en cada grupo, se llevó a cabo un examen al inicio y al final de la implementación de los programas, los valores obtenidos representan la diferencia que hay entre la nota del examen que se hizo al inicio y al finalde la implementación del programa. Los datos se muestran en el siguiente cuadro:

```
programa <- c(1, 2, 3)
programa
```

```
## [1] 1 2 3

Observaciones_1 <- c(20, 18, 18, 23, 22, 17, 15, 13, 21)
Observaciones_1

## [1] 20 18 18 23 22 17 15 13 21

Observaciones_2 <- c(15, 20, 13, 12, 16, 17, 21, 15, 13)
Observaciones_2

## [1] 15 20 13 12 16 17 21 15 13

Observaciones_3 <- c(12, 15, 18, 20, 18, 17, 10, 24, 16)
Observaciones_3

## [1] 12 15 18 20 18 17 10 24 16
```

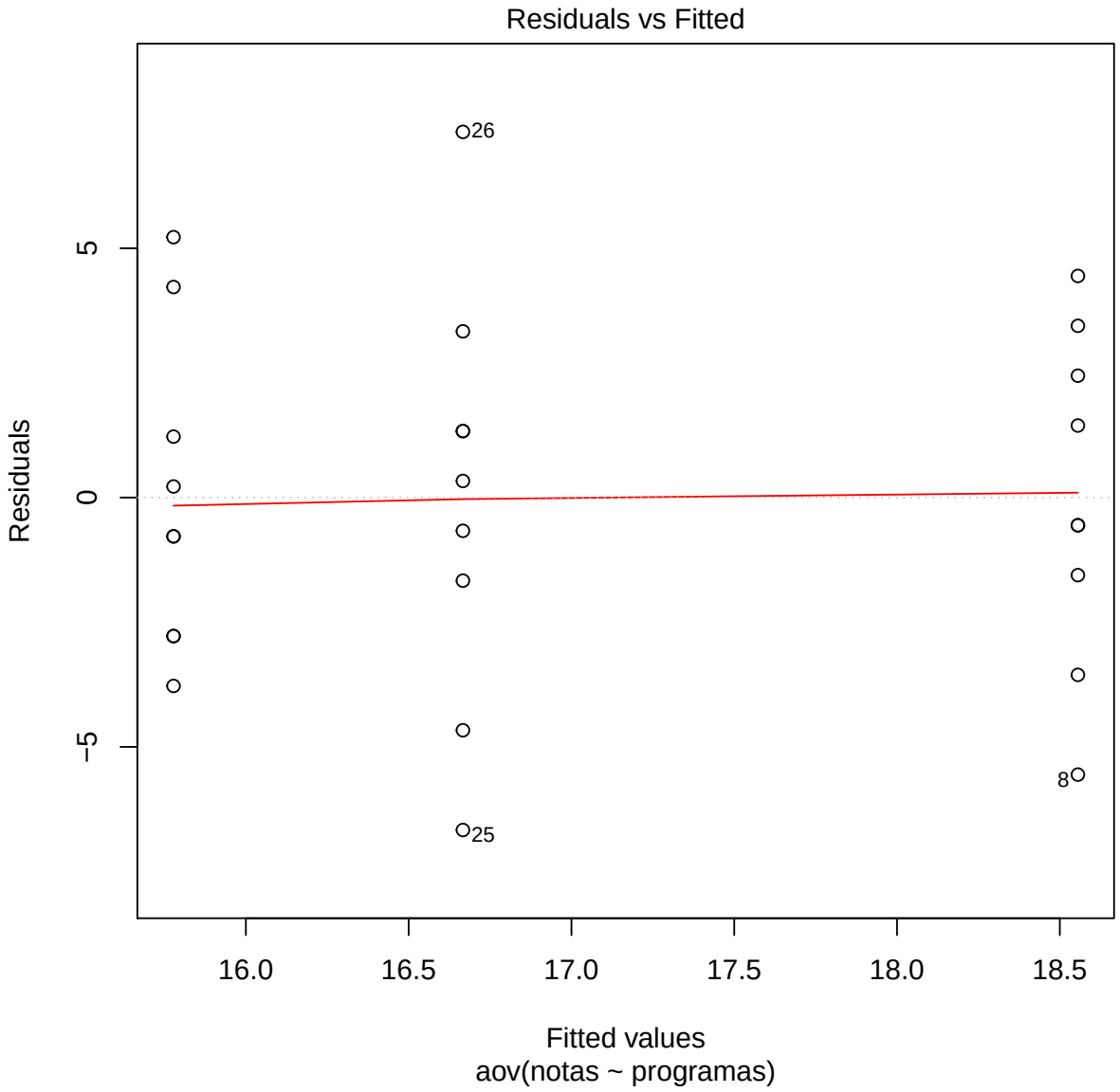
Contraste a un nivel de significancia del 5 % de que los tres métodos de lectura producen el mismo efecto en la habilidad de lectura de los alumnos.

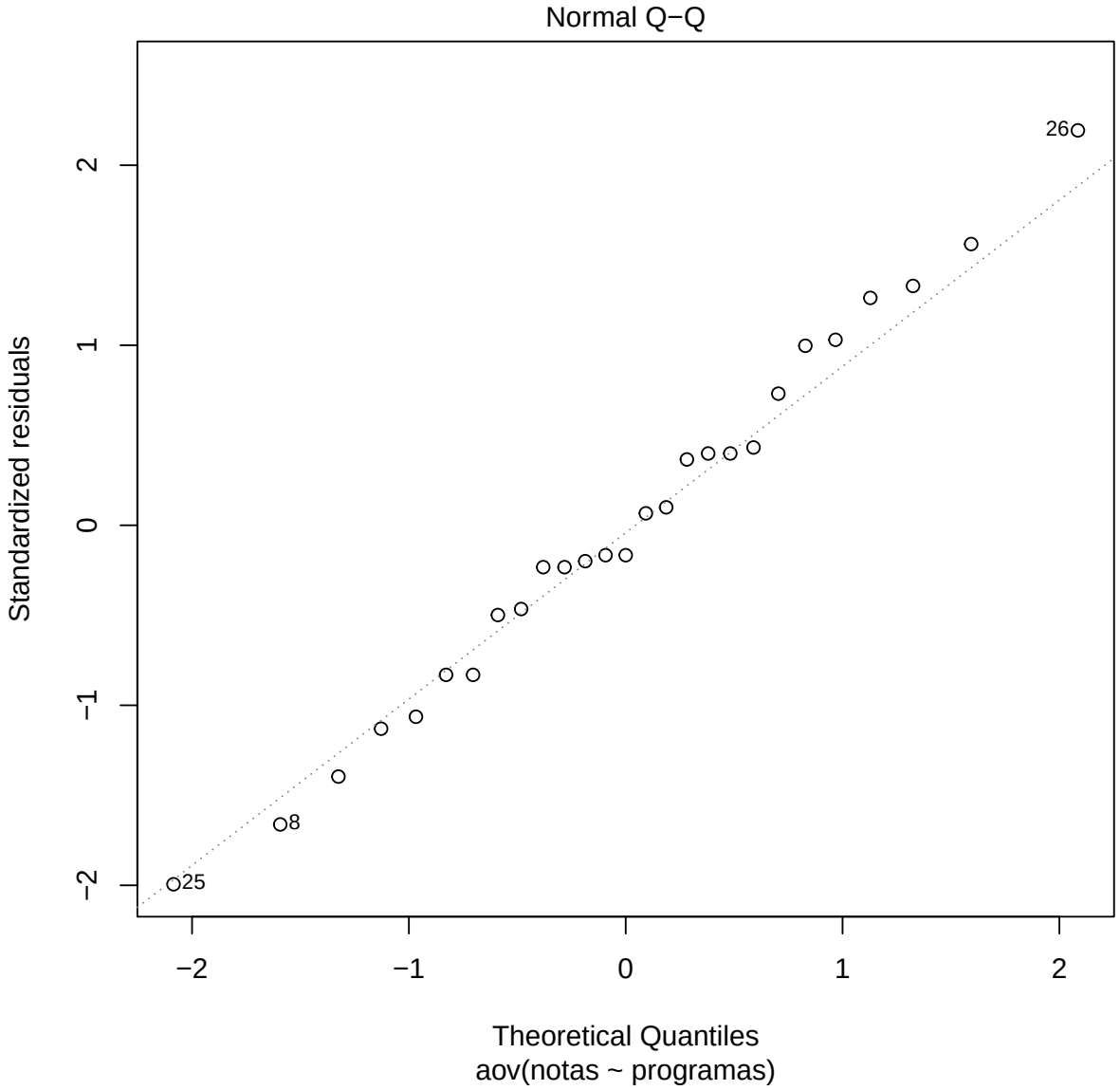
- La variable en estudio es la habilidad de lectura.
- El modelo que genera los datos es el siguiente:
  $\hat{y}_{ij} = \hat{\mu} + \hat{t}_i + \hat{u}_{ij}$
- Las hipótesis son las siguientes:
  H_o: $\mu_1 = \mu_2 = \mu_3$
  H_1: μ_1 distinto μ_2 distinto μ_3
- Ejecutar el script .ªnova1.R"

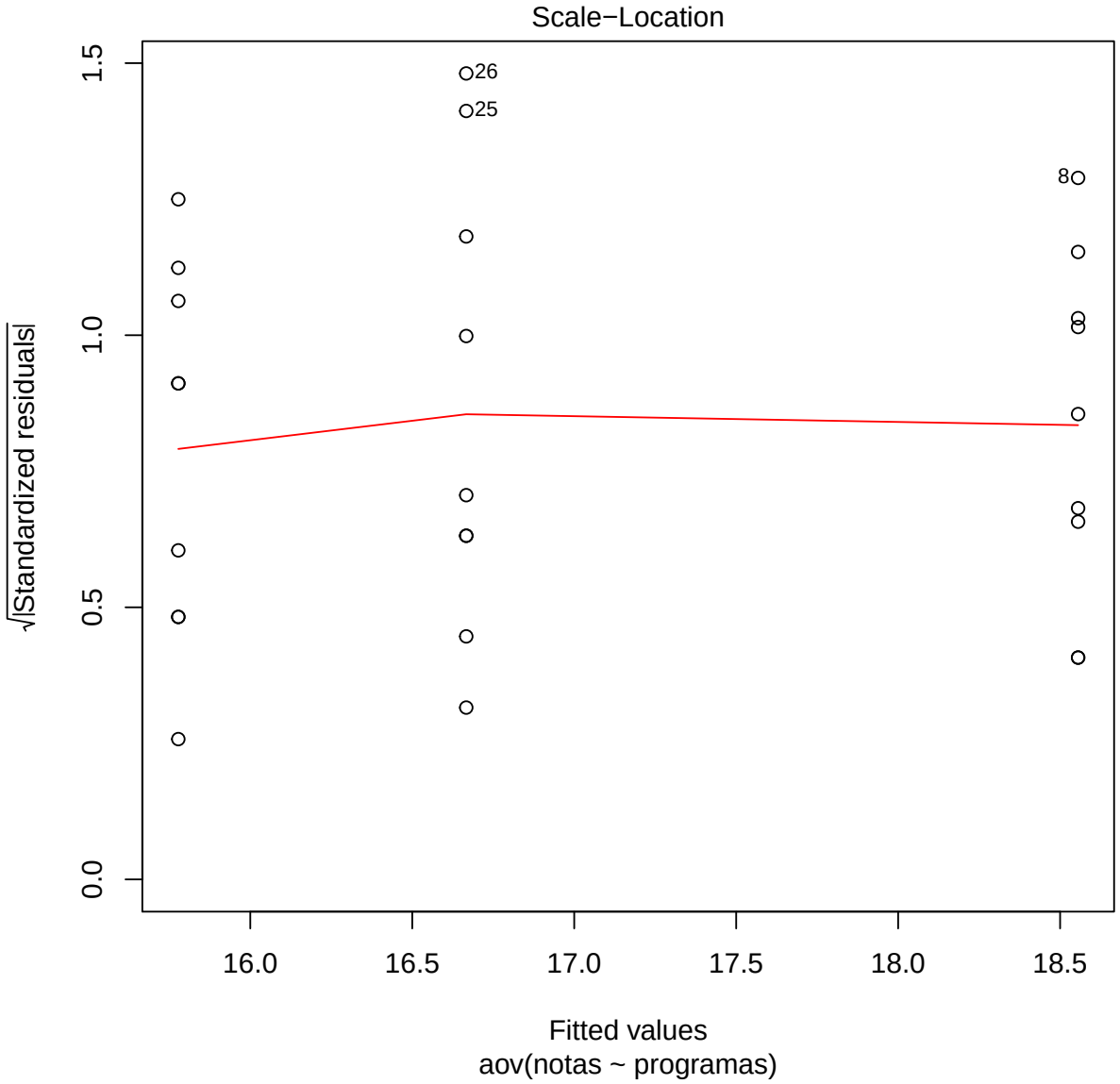
```
# Se digitan las observaciones

notas <- c(20, 18, 18, 23, 22, 17, 15, 13, 21, 15, 20, 13, 12, 16, 17, 21,
           15, 13, 12, 15, 18, 20, 18, 17, 10, 24, 16)

# Se crea un vector de datos en el cual se diferencia cada uno de los
# programas  de estudio los primeros 9 corresponden al primer programa
# de estudio, etiquetado por P1; los siguientes 9 corresponden al segundo
# programa, P2, y lo mismo para el tercero.

programas <- gl(n=3, k=9, labels=c("P1", "P2", "P3"))

#gl genera factores especificados por un patr\'on en sus niveles
# n especifica que se crea una variable factor con tres niveles diferentes
# etiquetados por "P1", "P2" y "P3". La instrucci\'on k=9 indica que a los
# primeros 9 elementos se les asignar\'a el valor de P1; a los siguientes 9
# el valor de P2; y a los \'ultimos 9 el valor de P3.
```

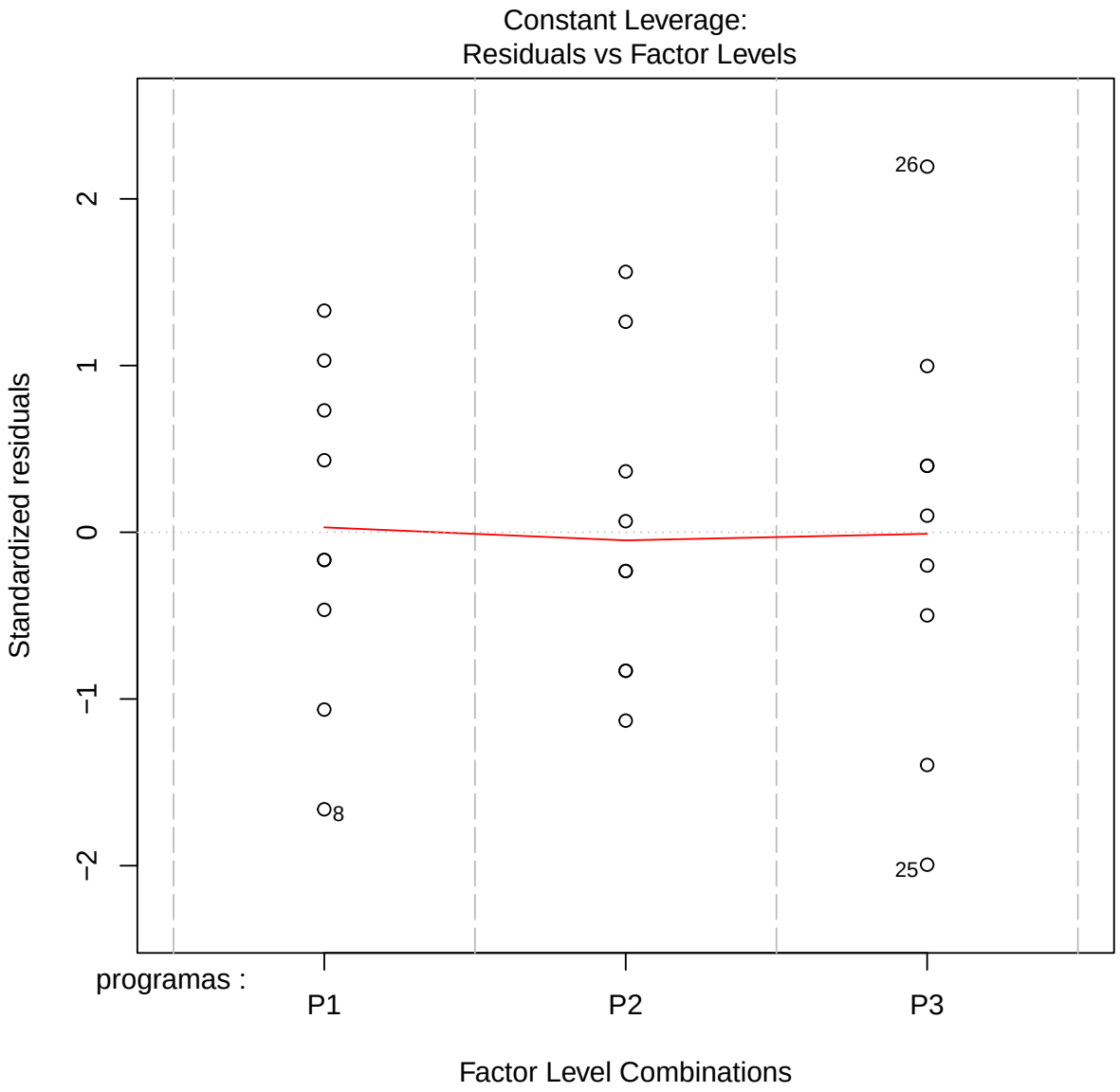
```
#Crea la matriz de datos que contendrÃ¡ la informaci\'on del experimento (es
# necesario que los datos est\'en organizados en una hoja de datos).

datos <- data.frame(notas = notas, programas = programas);datos

##    notas programas
## 1     20        P1
## 2     18        P1
## 3     18        P1
## 4     23        P1
## 5     22        P1
## 6     17        P1
## 7     15        P1
## 8     13        P1
## 9     21        P1
## 10    15        P2
## 11    20        P2
## 12    13        P2
## 13    12        P2
## 14    16        P2
## 15    17        P2
## 16    21        P2
## 17    15        P2
## 18    13        P2
## 19    12        P3
## 20    15        P3
## 21    18        P3
## 22    20        P3
## 23    18        P3
## 24    17        P3
## 25    10        P3
## 26    24        P3
## 27    16        P3

#Aplicando el an\'alisis de varianza

mod1 <- aov(notas ~ programas, data = datos)

# la expresi\'on notas ~ programas indica que se trata de explicar la
# variabilidad de la variable "notas" mediante el conocimiento (o en
# funci\'on de los valores) de la variable "programas", es decir, el
# nombre del factor que distingue a qu\'e grupo pertenece cada observaci\'on.
# Finalmente en "data = datos" se especifica el nombre de la hoja de datos.

# Mostrando la tabla ANOVA
```

```r
summary(mod1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## programas    2  36.22   18.11    1.44  0.257
## Residuals   24 301.78   12.57

# Para hacer un diagn\'ostico de las perturbaciones del modelo

plot(mod1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(notas ~ programas)
26
8
25
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
26
25
8
16.0
16.5
17.0
17.5
18.0
18.5
Fitted values
aov(notas ~ programas)



En la mayoría de los casos los datos a analizar se encontrarán en un archivo ya existente, para esto lo recomendable es que en el archivo tenga la estructura siguiente: una columna en la cual contenga las observaciones de la muestra de nuestra variable dependiente, y una columna adicional con el cual se identifique el grupo de pertenencia de cada una de las observaciones, siendo lo recomendable que sea una variable de tipo carácter; si este fuere el caso únicamente debemos convertir a la variable de tipo carácter en una variable de tipo y factor y realizar el procedimiento descrito en el ejemplo anterior.

En algunos casos, aunque sea muy raro, el archivo contendrá la siguiente estructura: contendrá tantas columnas como grupos se estén considerando, y encada columna se contarán con las observaciones correspondientes a dicho grupos, el número de observaciones no tiene porque ser los mismos por lo que se leerán unos cuantos datos faltantes. Veamos el siguiente ejemplo.

EL EJEMPLO 2, NO SE PUEDE ELABORAR DEBIDO A LA AUSENCIA DE LA BASE DE DATO.

COMENTARIOS FINALES DEL ANOVA.

El ANOVA en su versión paramétrica del test de la F, como todos los procedimientos estadísticos, tiene un cierto grado de robustez frente a un relativo incumplimiento de alguna(s) de sus hipótesis. En concreto, el test de la F soporta mejor las deficiencias respecto a la normalidad que las relacionadas con la homocedasticidad. En todo caso, los test son menos sensibles a las desviaciones de las hipótesis exigidas cuando el número de observaciones de las muestras es aproximadamente el mismo.

Se propone que, cuando se verifiquen todas las hipótesis exigidas la alternativa preferida sea el test de la F. Cuando se dé la normalidad pero no la homocedasticidad, se recomienda una alternativa no paramétrica, como el test de Kruskal Wallis. Si falla, aunque no de forma drástica la normalidad, con valores de p entre 0.01 y 0.05, la robustez del test de la F le hace seguir siendo una buena opción. Por último, si fallara fuertemente la normalidad, serecomienda el uso del test de Kruskal Wallis.