MAEA: Entrega final

Alberto Rincón Borreguero

Problema 1

Se clasificó a 177 personas casadas según su estatus de fumador, variable B, con valores de No Fumador, b1, Poco Fumador, b2, ($<$ 6 cigarrillos/día), Fumador Moderado, b3 ($\geq$ 6 y $<$ 15 cigarrillos/día) y Gran Fumador, b4 ($\geq$ 15 cigarrillos/día), y el de su pareja, variable A, con valores No Fumador, a1, Poco Fumador, a2 ($\geq$ 6 cigarrillos/día), Fumador Moderado, a3 ($\geq$ 6 y $<$ 15 cigarrillos/día) y Gran Fumador, a4 ($\geq$ 15 cigarrillos/día). Los resultados aparecen recogidos en la siguiente tabla:

```
X           <- matrix(c(42,12,18,2,18,22,6,8,4,8,10,12,0,2,6,7), ncol = 4)
colnames(X) <- c("No Fumador","Poco Fumador","Fumador Moderado","Gran Fumador")
rownames(X) <- c("Pareja No fumadora","Pareja Poco Fumadora","Pareja Fumadora Moderada","Pareja Gran Fu
X
```

```
##                          No Fumador Poco Fumador Fumador Moderado
## Pareja No fumadora               42           18                4
## Pareja Poco Fumadora             12           22                8
## Pareja Fumadora Moderada         18            6               10
## Pareja Gran Fumadora              2            8               12
##                          Gran Fumador
## Pareja No fumadora                  0
## Pareja Poco Fumadora                2
## Pareja Fumadora Moderada            6
## Pareja Gran Fumadora                7
```

Test de independencia de caracteres χ^2 con hipótesis nula h_0 : *Independecia entre las variables B y A.*

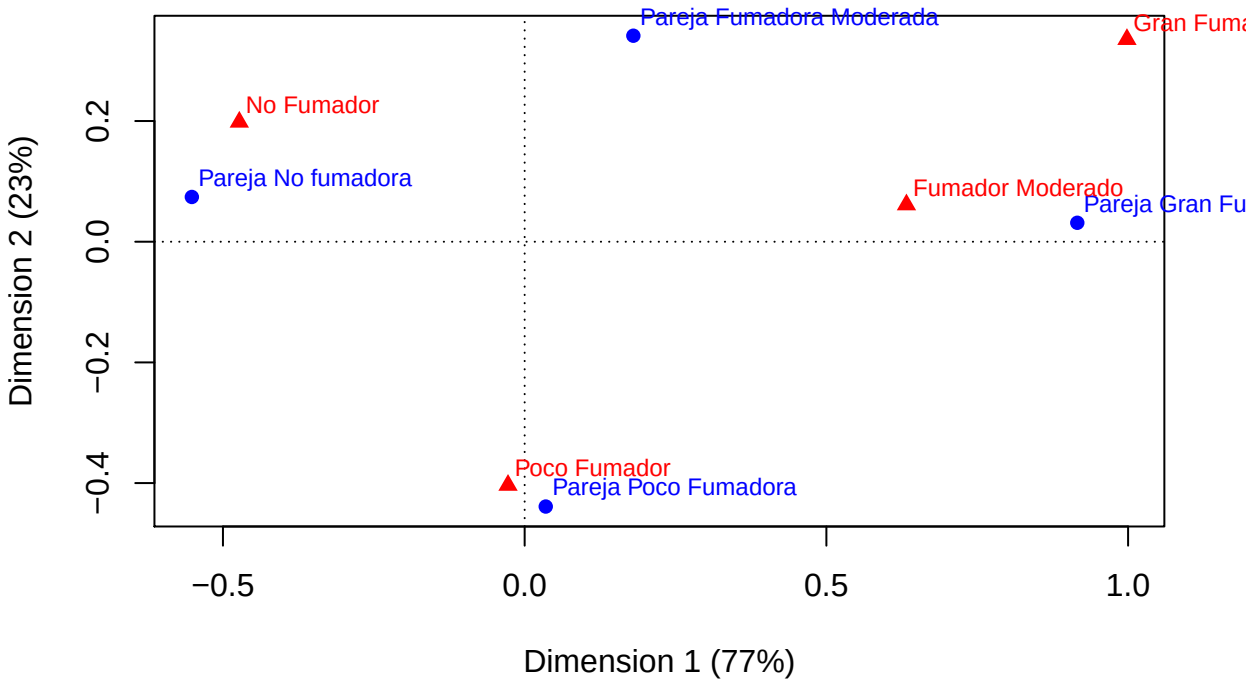
```
chi2 = chisq.test(X)
chi2
```

```
##
##  Pearson's Chi-squared test
##
## data:  X
## X-squared = 58.661, df = 9, p-value = 2.427e-09
```

Dado que el p-valor es prácticamente cero, 0.0000000024, se rechaza la hipótesis nula de independecia entre ambas variables.

Se procede a realizar un análisis de correspondencias para comprobar si exista alguna relación entre los valores observados de las dos variables.

```
library(ca)
correspondencias <- ca(X)
plot(correspondencias)
```



El gráfico bi-dimensional obtenido establece que las personas entrevistadas que se declararon No Fumadoras o Poco Fumadoras tienen parejas con el mismo hábito. Por otra parte, cuando el entrevistado se declara como Fumador Moderado o Gran Fumador, su pareja, por lo general, es Gran Fumadora. Se aprecia que la observación de la variable A, Pareja Fumadora Moderada, no tiene una relación estrecha con ninguna observacion de la variable B, siendo Poco Fumador la más alejada de todas.

Problema 2

```
injerto <- read.csv("datos/examen/injerto.txt", sep=" ")
head(injerto)
```

```
##   pnr rcpage donage type preg index gvhd time dead
## 1   1     27     23    2    0  0.27    0   95    1
## 2   2     13     18    2    0  0.31    0 1385    0
## 3   3     19     19    1    0  0.39    0  465    1
## 4   4     21     22    2    0  0.48    0  810    1
## 5   5     28     38    2    0  0.49    0 1497    0
## 6   6     22     20    2    0  0.50    0 1181    1
```

(a) Dado que las covariables a considerar en un modelo de regresión deben ser independientes, contrastar mediante un test de Spearman de independencia, si pueden considerarse independientes las covariables rcpage y donage.

```
corr <- cor.test(injerto$rcpage, injerto$donage, method = 'spearman')
corr
```

```
##
##  Spearman's rank correlation rho
##
## data:  injerto$rcpage and injerto$donage
## S = 2324.1, p-value = 3.985e-07
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##       rho
## 0.7245021
```

El test de spearman se interpreta de la siguiente forma: Magnitudes de rho cercanas a 1 indican mayor correlación mientras que los valores cercanos a cero indican una menor correlación. En este caso, el valor **0.72** indica que las variables rcpage y donage **no son independientes**.

(b) Analizar mediante una Regresión Logística qué variables son significativas para predecir la probabilidad p de presentar la enfermedad de Injerto contra Huésped, variable gvhd, de entre las 3 covariables siguientes: index, donage, preg.

```
regression <- glm(gvhd~index+donage+preg, data = injerto, family = binomial(link = "logit"))
summary(regression)
```

```
##
## Call:
## glm(formula = gvhd ~ index + donage + preg, family = binomial(link = "logit"),
##     data = injerto)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0716  -0.4978  -0.2732   0.6925   1.9978
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.88275    2.22347  -2.646  0.00815 **
## index        0.88989    0.37068   2.401  0.01637 *
## donage       0.11925    0.06261   1.905  0.05682 .
## preg         1.55904    1.01886   1.530  0.12597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 51.049  on 36  degrees of freedom
## Residual deviance: 29.848  on 33  degrees of freedom
## AIC: 37.848
##
## Number of Fisher Scoring iterations: 5
```

Se observa que la variable *index* con p-valor de 0.01 es la más significativa en la predicción.

(c) Determinar la estimación de p en función de las variables que resulten significativas.

```
regression <- glm(formula = gvhd ~ index, family = binomial(link = "logit"), data = injerto)
summary(regression)
```

```
##
## Call:
## glm(formula = gvhd ~ index, family = binomial(link = "logit"),
##     data = injerto)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9188  -0.7462  -0.5665   0.8256   1.6821
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9885     0.7479  -2.659  0.00784 **
## index         0.7747     0.2921   2.652  0.00799 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 51.049  on 36  degrees of freedom
## Residual deviance: 39.211  on 35  degrees of freedom
## AIC: 43.211
##
## Number of Fisher Scoring iterations: 5
```

(d) ¿Qué probabilidad de presentar la Enfermedad de injerto contra huésped tiene un individuo con índice reacciones de linfocitos igual a $2'5$, cuyo donante que nunca ha estado embarazada y tiene una edad de 55 años?

```
probability <- predict.glm(object = regression, data.frame(index=2.5, preg=0 ,donage=55), type = "respo
```

La probabilidad de presentar la enfermedad es del **48.71%**

Problema 3

Se desea realizar una Regresiónn no Lineal ajustando una función tipo sigmoide a los siguientes pares de datos,

x	y
19	65
25	61
38	56
47	28
53	12
69	10

utilizando la correspondiente función de autoarranque. Determinar la función sigmoide ajustada.

```
pb3.datos <- data.frame(x = c(19,25,38,47,53,69),
                        y = c(65,61,56,28,12,10))
```

Sea la función sigmoide:

$$\eta(x,\theta) = \theta_1 + \frac{\theta_2 - \theta_1}{1 + e^{\theta_3(x-\theta_4)}}$$

Se realiza la regresión no lineal, con su correspondiente función de arranque SSfpl, en la siguiente linea.

```
model <- nls(y~ SSfpl(-x,b1,b2,b3,b4), data=pb3.datos)
```

Obtenemos información acerca del modelo generado mediante la función summary.

```
summary(model)
```

```
##
## Formula: y ~ SSfpl(-x, b1, b2, b3, b4)
```

```
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## b1   9.1077     2.1178   4.301 0.050044 .
## b2  62.9236     1.6828  37.392 0.000714 ***
## b3 -44.7386     0.7732 -57.865 0.000299 ***
## b4   3.3635     0.6024   5.584 0.030610 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.335 on 2 degrees of freedom
##
## Number of iterations to convergence: 0
## Achieved convergence tolerance: 9.216e-06
```

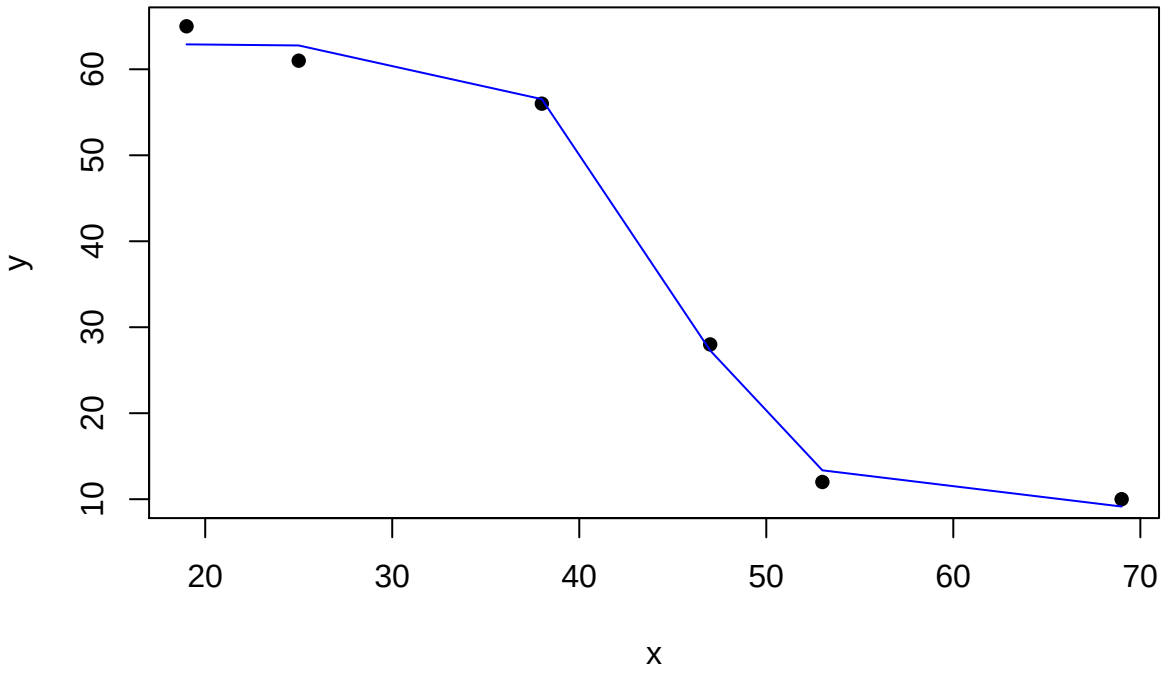
La suma de los errores residuales es muy pequeña, como se aprecia a continuación. Por tanto podemos decir que el modelo se ajusta adecuadamente a los datos.

```
sum(resid(model))
```

```
## [1] -2.664535e-14
```

Por último, mostramos el ajuste del modelo (linea azul) para los datos de la columna y del *dataframe*.

```
plot(pb3.datos, pch=16)
lines(pb3.datos$x, predict(model), col='blue')
```



Problema 4

Se desea estudiar el número de hembras de la mosca tropical americana en una determinada región. Dicha mosca se caracteriza por poner sus huevos en un mosquito, pasando las larvas de la mosca a la piel de la persona cuya sangre ha chupado el mosquito. Examinada la región en cuestión en 10 días elegidos al azar, se obtuvo el siguiente número de moscas hembra de la citada especie: 2,1,3,5,7,2,1,2,3,2 Se pide:

a) Determinar la estimación clásica y cuatro estimaciones robustas del número medio de moscas hembra en la región en estudio. ¿Con qué estimación concluiría?

```
data <- c(2,1,3,5,7,2,1,2,3,2)

# media clasica
mean(data)
```

```
## [1] 2.8
```

```
# media alpha-winsorizada
winsor.mean(data, trim = 0.1)
```

```
## [1] 2.62
```

```
winsor.mean(data, trim = 0.2)
```

```
## [1] 2.44
```

```
# media alpha-recortada muestral
mean(data, trim = 0.1)
```

```
## [1] 2.5
```

```
mean(data, trim = 0.2)
```

```
## [1] 2.333333
```

```
# mediana muestral
median(data)
```

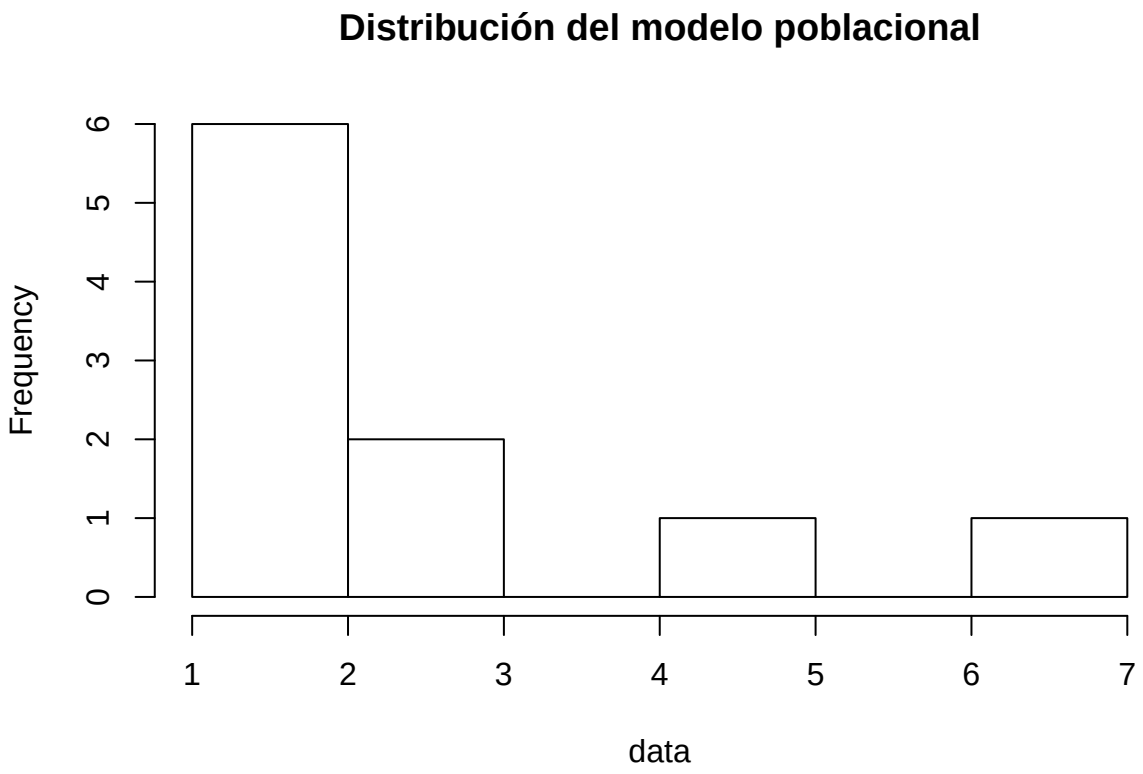
```
## [1] 2
```

```
# estimador de huber
huber(data, 1.28)$mu
```

```
## [1] 2.474431
```

Dada la siguiente distribución de los datos:

```
hist(data, main = 'Distribución del modelo poblacional')
```



Al no ser una distribución normal, se descarta la elección a priori de la estimación clásica que, en caso de normalidad, asegura mínima varianza.

Por tanto, y en base a las recomendaciones de la pagina 76 del libro métodos robustos y de remuestreo, **se elige el estimador huber**.

b) Determinar también la estimación clásica y cuatro estimaciones robustas de la desviación típica del número de moscas hembra en la citada región. ¿Con qué estimación concluiría?

```
# cuasidesivación típica muestral (estimación clásica)
sqrt(var(data))
```

```
## [1] 1.873796
```

```
# Desviación absoluta mediana estandarizada
mad(data)
```

```
## [1] 1.4826
```

```
#Cuasidesviación típica alpha-winsorizada muestral
sqrt(winsor.var(data))
```

```
## [1] 0.6719788
```

```
#Raíz de la varianza media biponderada
sqrt(r.bw(data))[[1]]
```

```
## [1] 1.592355
```

```
#Raíz de la varianza media de porcentaje ajustado
sqrt(pbvar(data))
```

```
## [1] 1.936492
```

De la misma forma que en el apartado anterior, se descarta la elección a priori de la estimación clásica

y también se sigue la recomendación del libro para **elegir como estimador la desviación absoluta mediana estandarizada** (NMAD)

Problema 5

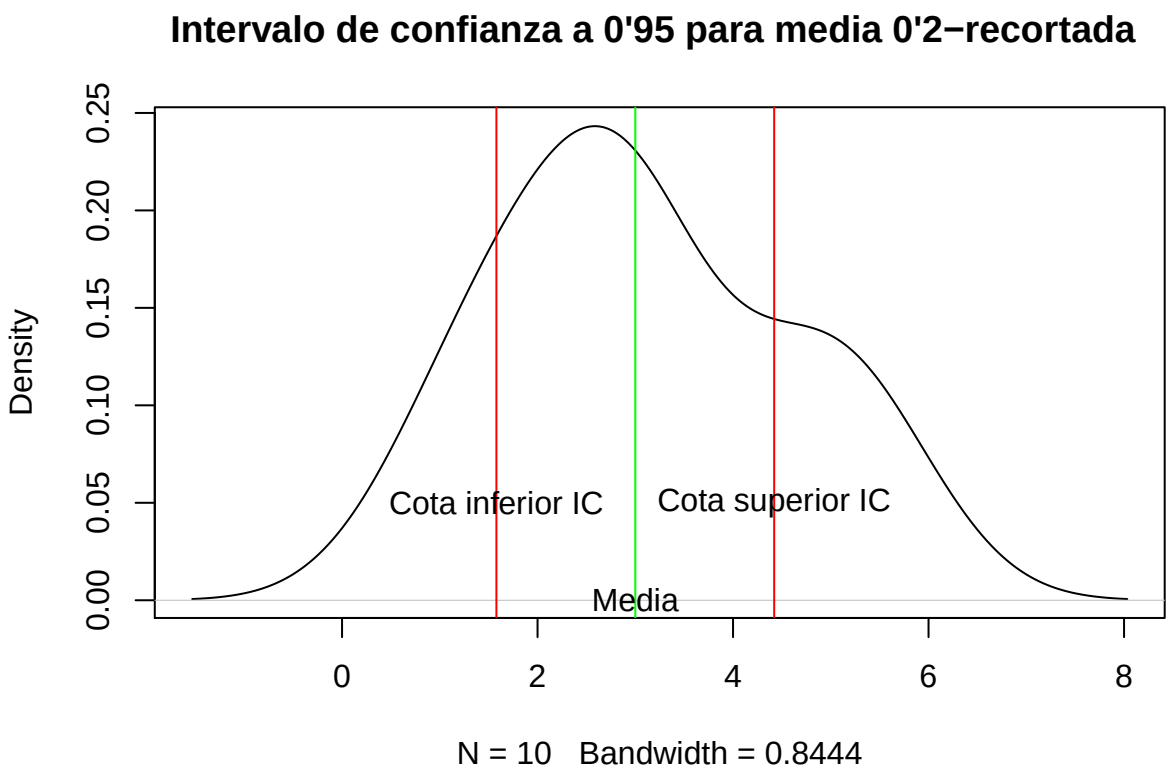
Los tiempos, en minutos, que esperaron, hasta que fueron atendidos en un determinado banco, diez clientes elegidos al azar fueron los siguientes: 1'5, 2, 2'5, 3, 1, 5, 5'5, 4'5, 3, 3. Determinar un intervalo de confianza de coeficiente de confianza 0'95, para la media 0'2-recortada del tiempo de espera y otro intervalo, también de coeficiente de confianza 0'95 para el tiempo mediano de espera.

```
data <- c(1.5, 2, 2.5, 3, 1, 5, 5.5, 4.5, 3, 3)
```

Cálculo del intervalo de confianza con coeficiente 0'95 para media 0'2-recortada.

```
trimCi <- MeanCI(data, conf.level = 0.95 , trim=0.2)

plot(density(data), main = "Intervalo de confianza a 0'95 para media 0'2-recortada ") + abline(v=trimCi
```

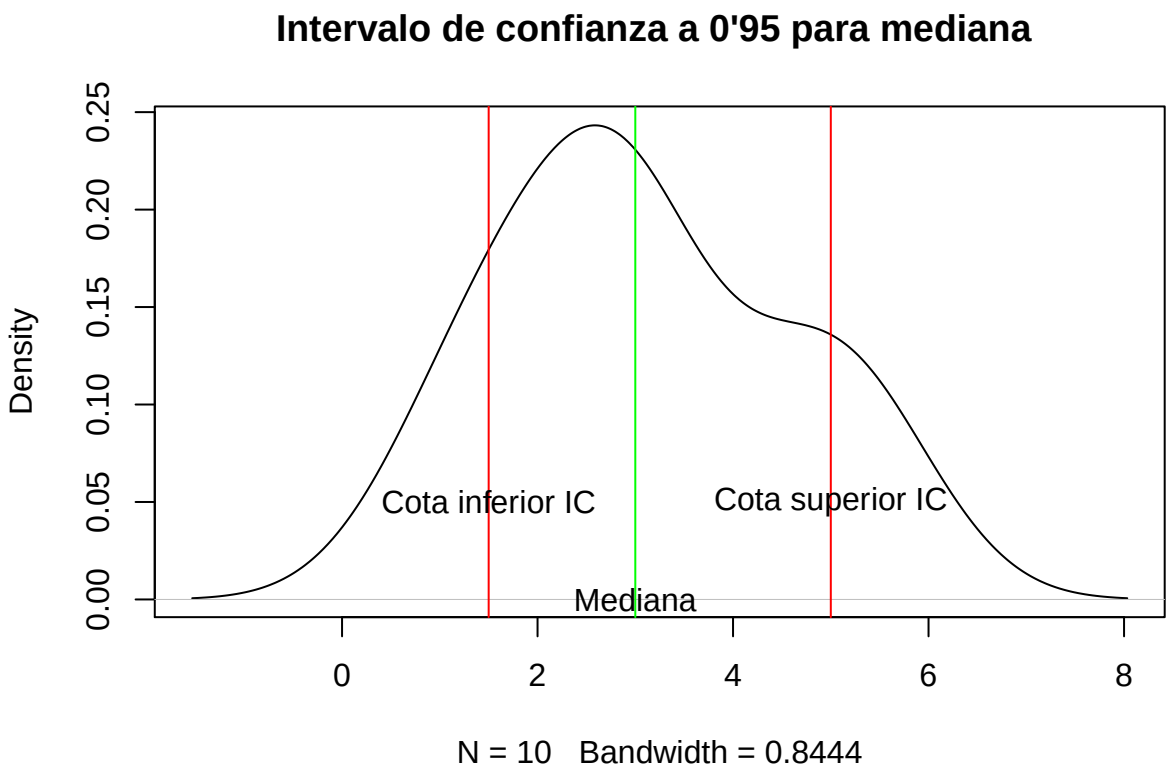



```
## integer(0)
```

Cálculo del intervalo de confianza con coeficiente 0'95 para el tiempo mediano de espera.

```
medianCI <- MedianCI(data, conf.level = 0.95)

plot(density(data), main = "Intervalo de confianza a 0'95 para mediana ") + abline(v=medianCI[1], col='
```




```
## integer(0)
```

Problema 6

Se quiere averiguar si, en promedio, existen diferencias significativas entre los precios de dos restaurantes, A y B. Para ello se eligieron al azar 11 días en los que se anotó el precio del menú del día en el restaurante A y otros 11 días en los que se anotó el precio del menú del día en el restaurante B. Los datos obtenidos fueron los siguientes:

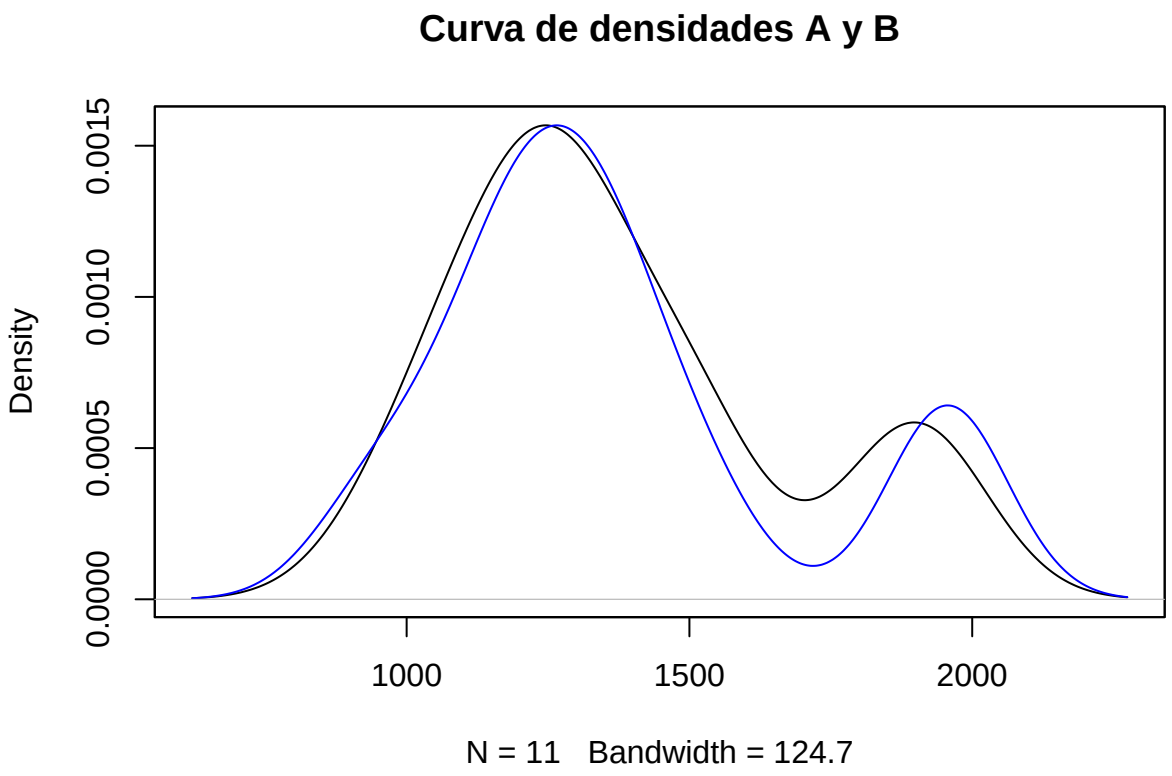
A: 1325 1500 995 1250 1290 1900 1500 1100 1250 1150 1900 B: 1100 1400 1000 1300 1300 1700 1250 1200 1150 1200 1700

```
data_A <- c(1325, 1500, 995, 1250, 1290, 1900, 1500, 1100, 1250, 1150, 1900)
data_B <- c(1100, 1400, 1000, 1300, 1300, 1700, 1250, 1200, 1150, 1200, 1700)
```

¿ Puede afirmarse a partir de dichos datos que existen diferencias significativas entre ambos restaurantes a nivel $\alpha = 0'05$?

Si comparamos las distribuciones de las poblaciones muestrales se comprueba que las diferencias, a simple vista, son escasas, haciéndose notar en la cola derecha (precios más altos). De hecho, se podría decir que cada distribución está compuesta por la suma de dos distribuciones: Una para los precios baratos y medios, y otra para los precios altos.

```
plot(density(data_A), main = "Curva de densidades A y B")
par(new=TRUE)
plot(density(data_B), col='blue' ,xaxt='n', yaxt='n', ann=FALSE, main = "Restaurante B ")
```



a) Utilizando las diferencias de medias 0'2-recortadas muestrales

Aplicando las medias 0'2-recortadas dada la hipótesis nula $H_0 : \mu_{\alpha,1} = \mu_{\alpha,2}$ y la alternativa $H_1 : \mu_{\alpha,1} \neq \mu_{\alpha,2}$.

```
trimCi_A <- mean(data_A, trim = 0.2)
trimCi_B <- mean(data_B, trim = 0.2)
YuenTTest(data_A, data_B, tr = 0.2, conf.level = 0.95)
```

```
##
##  Yuen Two Sample t-test
##
## data:  data_A and data_B
## t = 0.74682, df = 10.587, trim = 0.200, p-value = 0.4714
## alternative hypothesis: true difference in trimmed means is not equal to 0
## 95 percent confidence interval:
##  -130.2806  263.1377
## sample estimates:
## trimmed mean of x trimmed mean of y
##          1323.571          1257.143
```

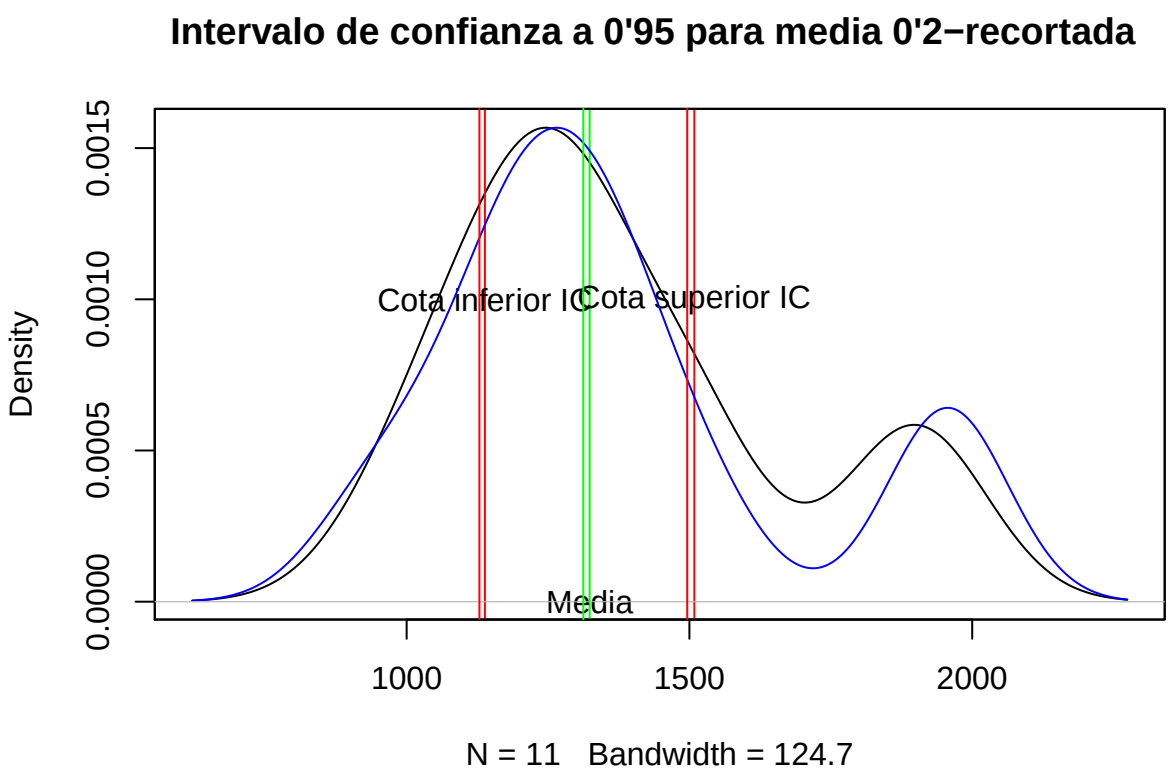
Se acepta H_0 ya que el valor 0 pertenece al intervalo calculado $[-130.2806, 263.1377]$ y el p-valor es 0.47, confirmándose por tanto que **no existen diferencias significativas**.

```
trimCiA <- MeanCI(data_A, conf.level = 0.95 , trim=0.2)
trimCiB <- MeanCI(data_B, conf.level = 0.95 , trim=0.2)

plot(density(data_A), main = "Intervalo de confianza a 0'95 para media 0'2-recortada ") + abline(v=trimC
```

```
## integer(0)
```

```r
par(new=TRUE)
plot(density(data_B), col='blue' ,xaxt='n', yaxt='n', ann=FALSE, main = "Intervalo de confianza a 0'95 p
```




```
## integer(0)
```

b) Utilizando la generalización robusta del test de Wilcoxon-Mann-Whitney

Si consideramos la hipótesis nula $H_0 : M_A = M_B$ y la alternativa $H_1 : M_A \neq M_B$, y dado que $ H_0 : p = 1/2 $, se va a determinar un intervalo de confainza para p, de coeficiente de confianza $0'95$. En caso de que el valor $1/2$ se incluya en el intervalo calculado, entonces se aceptará la hipótesis nula de igualdad de las poblaciones.

Utilizaremos la función wilcox.test, dónde si asignamos el valor del parámetro paired = FALSE, según la documentación estaremos aplicando el test de la suma de rango de Wilcoxon, equivalente al test de Mann-Whitney.

```r
wilcox.test(x = data_A, y = data_B, conf.level = 0.95, conf.int = TRUE, paired = FALSE)
```

```
## Warning in wilcox.test.default(x = data_A, y = data_B, conf.level = 0.95, :
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(x = data_A, y = data_B, conf.level = 0.95, :
## cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_A and data_B
## W = 69, p-value = 0.5982
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -150  250
```

```
## sample estimates:
## difference in location
##                     50
```

Puesto que el valor 0.5 se encuentra dentro del intervalo de confianza y que el p-valor es lo suficientemente grande, confirmamos la hipótesis $H_0 : M_A = M_B$.

Problema 7

Se quiere averiguar si tres fertilizantes, A, B y C presentan diferencias significativas en cuanto a sus efectos sobre el aumento de la cosecha.

Con este propósito se eligieron al azar 15 parcelas al as que se fertilizó aleatoriamente con cada uno de los fertilizantes en cuestión. Los aumentos de cosecha obtenidos fueron los siguientes:

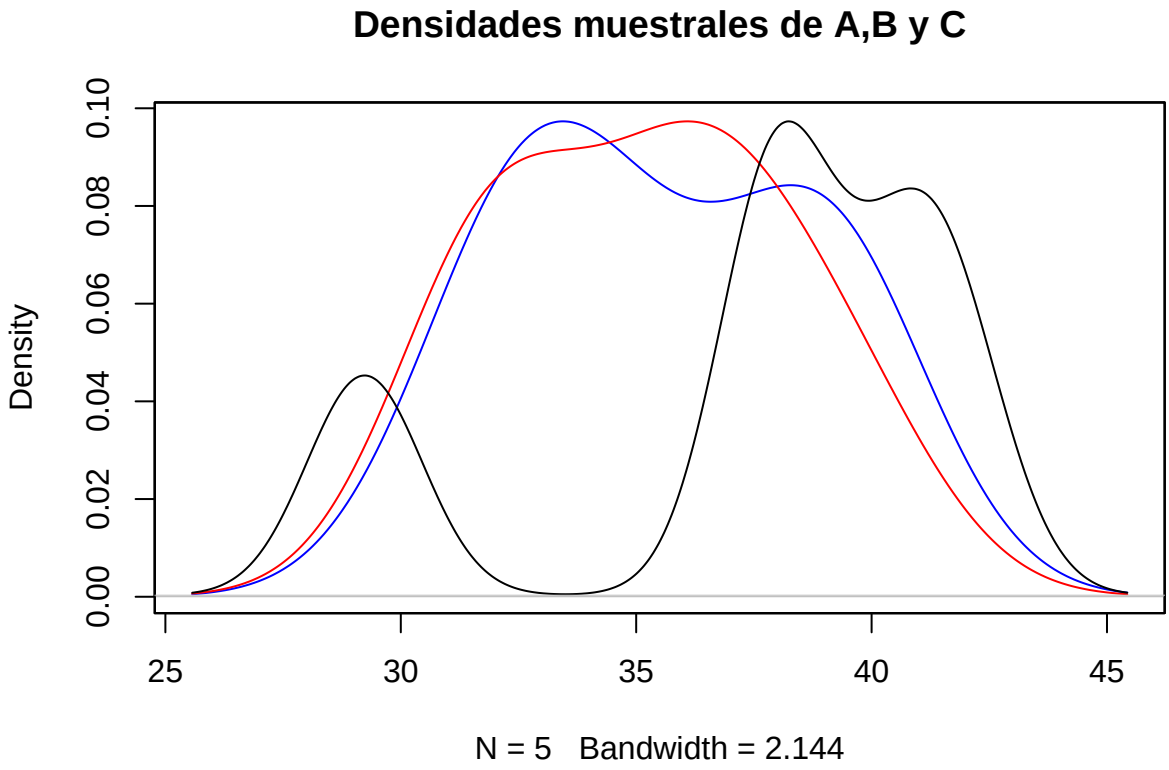
Fertilizante	Aumento de cosecha
A :	39 33 39 35 32
B :	36 40 35 30 29
C :	33 33 36 26 35

A la vista de estos datos y recortando $\alpha = 0'1$, ¿ puede inferirse que existen diferencias significativas entre los tres fertilizantes?

La siguiente gráfica muestra la ausencia de normalidad en todas las distribuciones, siendo la de B (línea color rojo) las más cercana a dicha normalidad. En el caso de C (color negro), se puede ver una composición de dos distribuciones, siendo la primera de ellas (a la izquierda) presumiblemente normal.

```
A <- c(39,33,39,35,32)
B <- c(36,40,35,30,29)
C <- c(33,33,36,26,35)


plot(density(A), col='BLUE', main='Densidades muestrales de A,B y C')
par(new=TRUE)
plot(density(B), col='RED',xaxt='n', yaxt='n', ann=FALSE)
par(new=TRUE)
plot(density(C), xaxt='n', yaxt='n', ann=FALSE)
```



a) Utilizando la generalización robusta del test de Welch.

```
clnames <- list("aumentoCosecha", "fertilizante")
A <- data.frame(c(39,33,39,35,32),factor("A"))
names(A) <- clnames
B <- data.frame(c(36,40,35,30,29), factor("B"))
names(B) <- clnames
C <- data.frame(c(33,33,36,26,35), factor("C"))
names(C) <- clnames

x <- rbind(A,B,C)
welch.test(formula =  aumentoCosecha ~ fertilizante, data = x, rate = 0.1)
```

```
##
##   Welch's Heteroscedastic F Test with Trimmed Means and Winsorized Variances (alpha = 0.05)
## -------------------------------------------------------------------------------------------
##   data : aumentoCosecha and fertilizante
##
##   statistic  : 0.8057529
##   num df     : 2
##   denom df   : 7.864925
##   p.value    : 0.4804701
##
##   Result     : Difference is not statistically significant.
## -------------------------------------------------------------------------------------------
```

b) Utilizando la generalización robusta del test de Box.

Como vemos a continuación, se reafirma que la diferencia no es estadísticamente significativa.

```
box.test(formula =  aumentoCosecha ~ fertilizante, data = x, verbose = TRUE)
```

```
##
##   Box F Test (alpha = 0.05)
## ------------------------------------------------------------
##   data : aumentoCosecha and fertilizante
##
##   statistic  : 0.7253219
##   num df     : 1.936947
##   denom df   : 33.79948
##   p.value    : 0.4873822
##
##   Result     : Difference is not statistically significant.
## ------------------------------------------------------------
```

Problema 8

Se cree que la duración del revestimiento de un estanque depende de la cantidad de cal hidráulica que contiene. Para analizar esta relación se midió, en siete revestimientos, el tiempo, Y, hasta la aparición de filtraciones, teniendo cada uno de los revestimientos diferentes porcentajes de cal hidráulica, X. Los resultados obtenidos fueron los siguientes:

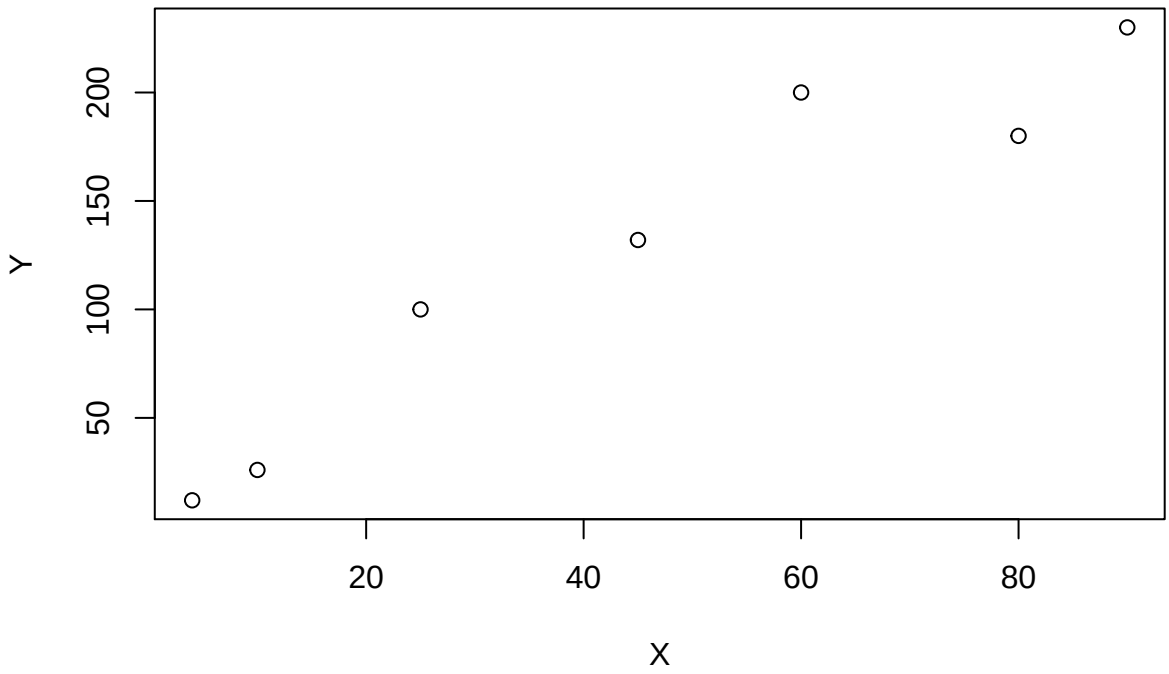
X: 4, 10, 80, 45, 25, 60, 90 Y: 12, 26, 180, 132, 100, 200, 230

```
X <- c(4, 10,  80,  45,  25,  60,  90)
Y <- c(12, 26, 180, 132, 100, 200, 230)
```

Se pide:

a) La recta de M-regresión óptima.

```
plot(X,Y)
```



Se aprecia en la gráfica anterior que no hay ningún punto que pudiera ser considerado outlier.

```
mregresion <- rlm(formula = Y~X, method = 'M', scale.est = 'Huber')$coefficients
mregresion
```

```
## (Intercept)           X
##   17.397885    2.414697
```

Se obtiene que la recta es $ y = 17'39 + 2'41x$ la de mínimos cuadrados, la cual se obtiene ejecutando:

```
ols <- lm(Y~X)$coefficients
ols
```

```
## (Intercept)           X
##   17.397885    2.414697
```

b) La recta de regresión media biponderada.

También puede obtener mediante la función rlm utilizando 'MM' como método.

```
biweighted <- rlm(formula = Y~X, method = 'MM')$coefficients
biweighted
```

```
## (Intercept)           X
##   16.962996    2.416254
```

c) La recta de regresión winsorizada.

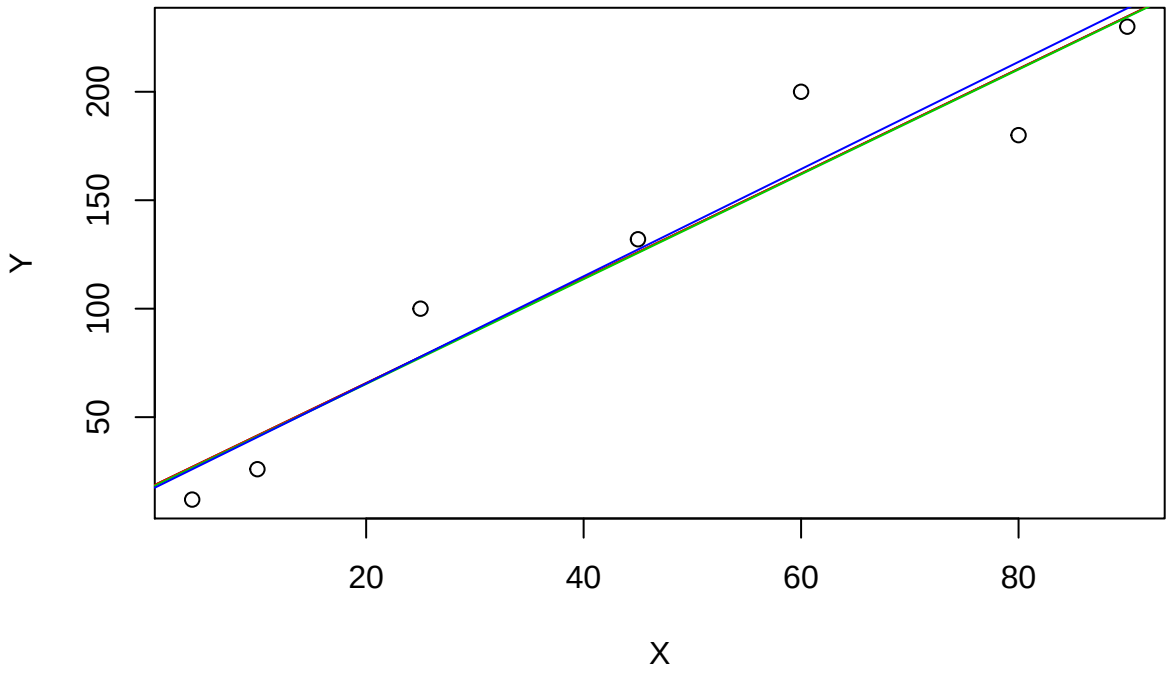
```
data <- data.frame(X, Y)

winsr <- lmWinsor(formula = Y~X, data=data)$coefficients
winsr
```

```
## (Intercept)           X
##   16.076760    2.471158
```

A continuación se dibujan todas las rectas obtenidas en una misma gráfica para su comparación. En dicha gráfica veremos como el resultado obtenido es prácticamente el mismo debido a la ausencia de valores extremos.

```
plot(X,Y)
abline(mregresion)
abline(ols, col=2)
abline(biweighted, col=3)
abline(winsr, col=4)
```


Y
50
100
150
200
20
40
60
80
X