El trabajo número 16a

Alicja Weronika Nowakowska

1 Introducción

El uso de los modelos matemáticos resulta crucial para la resolución de los problemas. En este trabajo se usa el modelo estadístico que simplifica la realidad y representa la relación entre las variables explicativas y la variable de respuesta.

2 El problema

Se tiene un conjunto de datos fluoride.txt donde se tiene información sobre un experimento para reducir las caries, realizado sobre 60 niñas. La variable de interés es la diferencia entre el número de dientes faltantes o con empastes antes del tratamiento (B, before) y después (A, after). La variable edad representa la edad de las niñas antes de cada tratamiento (distinguimos tres tipos de tratamiento: SF, APF y un placebo, W, que era agua destilada). El objetivo de este trabajo es ajustar y validar un modelo de la covarianza. Además se evaluará la posible presencia de observaciónes atípicas o influyentes.

3 El modelo de la covarianza

El modelo de la covarianza, también llamado ANCOVA, es un modelo linear general que se usa en los problemas donde existen simultaneamente variables explicativas discretas (que indican el grupo de pertenencia) y continuas. Distinguimos dos tipos: el modelo con interacción y sin ella.

- **El modelo sin interacción** - se asume que el efecto de las variables simplemente se suma. La fórmula para variable j del grupo i es la siguiente:

$$Y_{ij} = \mu + \alpha_i + \gamma z_{ij} + \epsilon_{ij}$$

- **El modelo con interacción** - las variables explicativas afectan una a la otra. La fórmula es la siguiente:

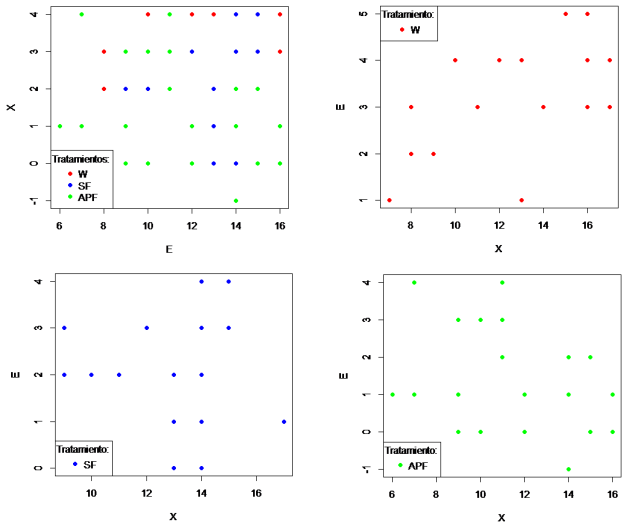
$$Y_{ij} = \mu_i + \gamma_i z_{ij} + \epsilon_{ij}$$

  Donde ϵ_{ij} pertenece a $N(0, \sigma^2)$.

En el trabajo se estudiarán ambos tipos de modelos para evaluar cual funciona mejor. Para cumplir este objetivo se va a usar el paquete R.

4 Ajuste del modelo

Desde ahora la variable de interés (la variable de respuesta) se denota como X, las variables explicativas son la edad - E y el tipo de tratamiento - T. Debajo se presenta la gráfica que presenta los datos obtenidos del experimento considerando diferentes grupos de tratamiento. El grupo del tratamiento "W" tuvo 20 participantes, "SF" tuvo 22 y "APF" 27. Es importante resaltar que algunos datos se repiten para diferentes grupos (por ejemplo existen dos chicas con la misma edad y la misma diferencia aunque pertencen a otros grupos). Debajo se presentan las relaciones entre las variables dependiendo del grupo de pertenencia.



4.1 El modelo sin interacción

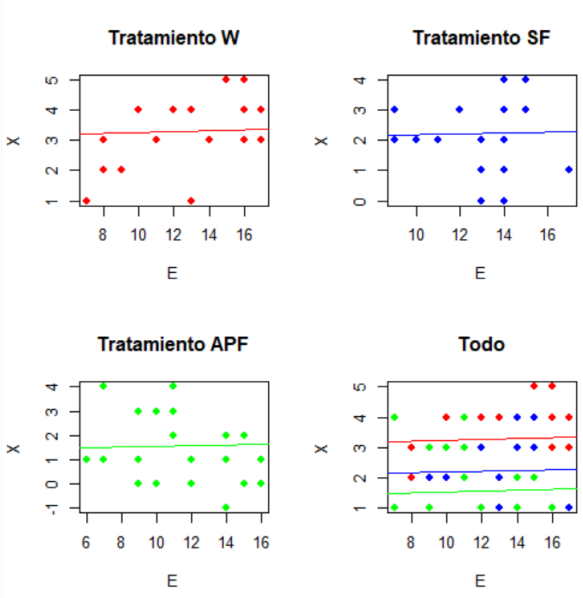
En el modelo sin interacción se crean tres modelos de regresión simple con el mismo coeficiente γ y diferentes valores del intercepto. Usando la función "lm" en el lenguaje R se obtiene:

```
> summary(mod1)

Call:
lm(formula = X ~ E + T)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5926 -0.6212 -0.1746  0.7539  2.5074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.39248    0.69698   1.998   0.0499 *
E            0.01430    0.05715   0.250   0.8033
TSF          0.65350    0.37500   1.743   0.0861 .
TW           1.72096    0.38935   4.420 3.83e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.281 on 65 degrees of freedom
Multiple R-squared:  0.2479,    Adjusted R-squared:  0.2132
F-statistic: 7.141 on 3 and 65 DF,  p-value: 0.0003224
```

Se calcula que $\gamma = 0.0143$. Además el pendiente:

- **Para el tratamiento APF**: $\beta_{APF} = 1.3925$
- **Para el tratamiento W**: $\beta_W = 1.3925 + 1.7210$
- **Para el tratamiento SF**: $\beta_{SF} = 1.3925 + 0.6535$

Lo cual se representa de la siguiente manera:



4.2 El modelo con interacción

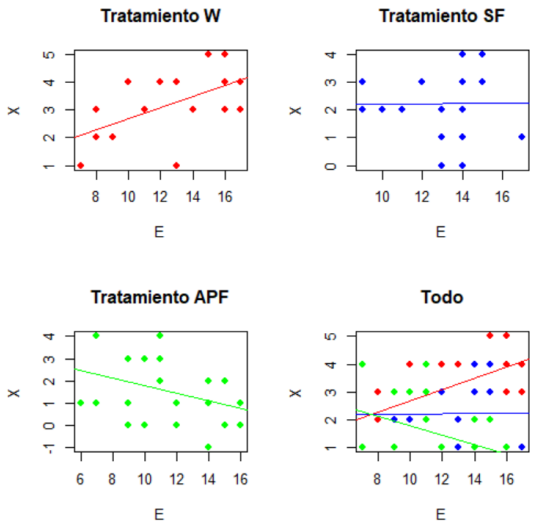
También se crean tres modelos de regresión simple. Sin embargo, el término γ también cambia su valor dependiendo del grupo. Planteando el modelo en el lenguaje R se obtiene:

```
> summary(mod2)

Call:
lm(formula = X ~ E + T + E:T)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2900 -0.9440  0.1091  0.7764  2.3751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.49733    1.02182   3.423  0.00109 **
E           -0.17022    0.08719  -1.952  0.05536 .
TSF         -1.33896    1.81603  -0.737  0.46368
TW          -2.80143    1.54401  -1.814  0.07438 .
E:TSF        0.17565    0.14560   1.206  0.23216
E:TW         0.36977    0.12262   3.016  0.00370 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.216 on 63 degrees of freedom
Multiple R-squared:  0.3428,    Adjusted R-squared:  0.2907
F-statistic: 6.572 on 5 and 63 DF,  p-value: 5.583e-05
```

De esto se tiene que:

- **Para el tratamiento APF**: $\gamma_{APF} = -0.1702, \beta_{APF} = 3.4973$
- **Para el tratamiento W**: $\gamma_W = -0.1702 + 0.3698, \beta_W = 3.4973 - 2.8014$
- **Para el tratamiento SF**: $\gamma_{SF} = -0.1702 + 0.175, \beta_{SF} = 3.4973 - 1.3390$

Los gráficos para este modelo son:



4.3 Elección del modelo

Se considera los siguientes factores para elegir el modelo:

- **El valor de R^2 para modelos de regresión múltiple**: ayuda a evaluar que modelo tiene el mejor ajuste, se apunta a tener el valor mayor posible. En el caso del modelo sin interacción equivale a 0.2132, en el caso del modelo con interacción a 0.3428, lo cual sugiere que el modelo con interacción se ajusta mejor.

- **El resultado de la función ANOVA**: compara los modelos de regresión y calcula si los factores añadidos son significativos.

```
Analysis of Variance Table

Model 1: X ~ E + T
Model 2: X ~ E + T + E:T
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     65 106.628
2     63  93.169  2    13.459 4.5504 0.01426 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  En el caso considerado el p-valor es $0.01426 < 0.05$. Esto indica que se puede rechazar la hipótesis nula que el factor de interracción no es significativo.

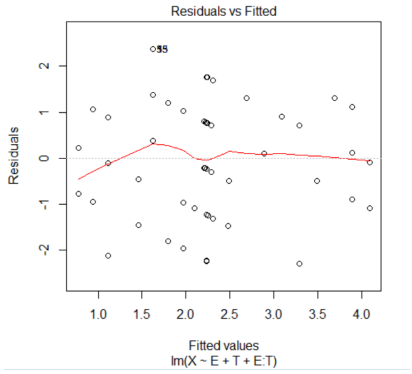
- **El resultado de F-Test**: indica si los resultados obtenidos a través del modelo de regresión son mejores que los aleatorios. Esperamos que el p-valor sea lo más pequeño posible. Para el caso del modelo con interacción el p valor de F-test es $5.583 \cdot 10^{-5}$ y para el modelo sin interacción equivale a 0.0003224. En ambos casos el resultado es aceptable, sin embargo, es mejor para el modelo con interacción.

Teniendo en cuenta todo esto se concluye que los factores de interacción son importantes y se elige el modelo con interacción.

4.4 Validación del modelo

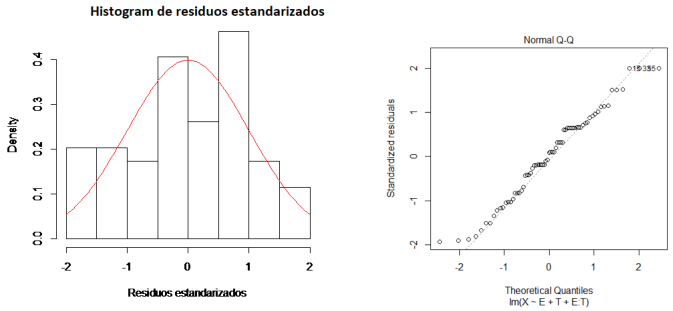
El modelo de la covarianza es un modelo lineal general, por esta razón ha de cumplir las condiciónes particulares:

- **Linealidad del modelo** Analizando la gráfica no se observa ningún



patrón específico. Aunque la curva no sea perfectamente recta se puede asumir que la condición está cumplida.

- **Normalidad de los residuos** En las gráficas siguientes se presentan los residuos estandarizados.

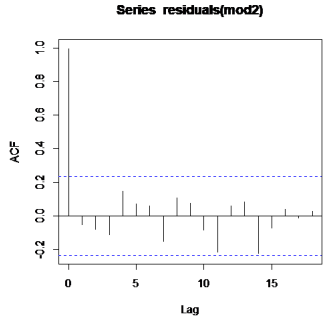


Las observaciones no se diferencian llamativamente del comportamiento de las variables de la distribución normal. Además el resultado del Test de Shapiro-Wilk no permite rechazar la hipotesis nula que los residuos proceden de la distribución normal.

```
> shapiro.test(rest)

        Shapiro-Wilk normality test

data:  rest
W = 0.97619, p-value = 0.2103
```

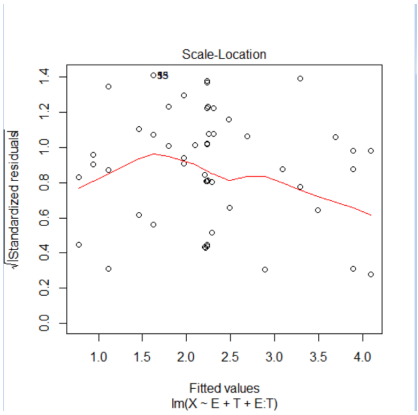
Por todo esto se asume que los residuos cumplen esta condición.

- **Independencia de los residuos**: se comprueba a través de la función acf de R. Ninguna de las rectas supera los limites azules, lo cual significa



que los residuos son independientes.

- **Homocedasticidad**
La varianza de los residuos no puede depender de la variable explicativa.



No parece que la varianza se disminuya o aumente su valor significativamente a lo largo de los valores ajustados, por lo tanto se puede decir que la condición está cumplida.

Teniendo en cuenta este análisis se supone que el modelo es válido.

5 Diagnóstico de la presencia de las observaciones atípicas o influyentes

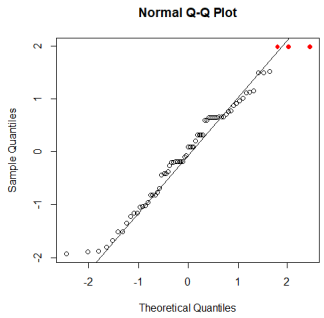
Según la definición, las observaciones atípicas son las que se separan mucho del comportamiento esperable bajo el modelo. Esto puede pasar porque no pertencen al modelo o ha habido un error en su observación o transcripción. Las

observaciones influyentes son aquellas que modifican sustancialmente el ajuste del modelo.

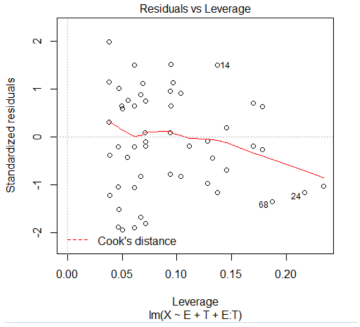
5.1 Presencia de las observaciones atípicas

Ya que todas las observaciónes influyentes son atípicas, primero hay que detectar las observaciónes atípicas. Para eso, se considera lo siguiente:

- **La lista de los residuos estandarizados** (la cual se obtiene usando la función "rstandard" en R): los elementos que sean mayores que 2 o menores que -2 indican las observaciones atípicas. En nuestro caso existen residuos mayores que 1.99 (son observaciones con los números: 33, 15, 55)
- **QQplot**: se puede ver si algunos residuos destacan de los otros, en rojo se marcan los residuos que son mayores que 1.99. La observación roja que es la más destacada y tiene el número 55.



- **Distancia de Cook**: indica la observación atípica si su valor es entre 0.5 y 1. Lo cual no existe para los datos considerados.



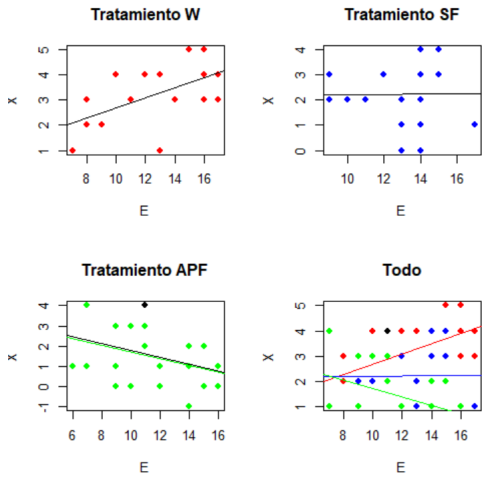
Dado todo esto, se deduce que la observación número 55 tiene el carácter más atípico de todas. Pertenece al grupo "APF".

5.2 Presencia de las observaciones influyentes

Para ver si la observación atípica detectada tiene el carácter influyente hay que eliminarla del conjunto de datos y ver si los cambios en el modelo son relevantes. Debajo se presenta un resumen del modelo sin esta observación.

```
> summary(mod3)

Call:
lm(formula = X ~ E + T + E:T)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29002 -0.87117  0.03643  0.76828  2.46863

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.34692    0.99981   3.348  0.00139 **
E           -0.16505    0.08512  -1.939  0.05705 .
TSF         -1.18854    1.77361  -0.670  0.50526
TW          -2.65102    1.50845  -1.757  0.08378 .
E:TSF        0.17048    0.14210   1.200  0.23479
E:TW         0.36460    0.11968   3.046  0.00340 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.187 on 62 degrees of freedom
Multiple R-squared:  0.3708,    Adjusted R-squared:   0.32
F-statistic: 7.307 on 5 and 62 DF,  p-value: 1.949e-05
```

Se nota que el valor de R^2 se mejoró, en otras palabras, ahora es mayor. Además, el resultado del F-test es menor que antes. En la siguiente gráfica se observa el cambio en las rectas. El color negro indica los ajustes obtenidos con la observación. El punto negro indica la observación eliminada.



Se observa que los cambios en los grupos "W" y "SF" son poco significativos. En el grupo "APF" se nota que dicha observación causa un apalancamiento. Se concluye que la observación con el número 55 influye en el modelo negativamente y es recomendable eliminarla.

6 Conclusión

El modelo ajustado no es perfecto. El R^2 no tiene un valor esperable, aunque lo podemos llamar aceptable. Se observan algunas apostasías que son probablemente causadas por el hecho que el conjunto de datos es bastante pequeño. Además, para mejorar el modelo se puede eliminar el dato 55 que tiene el carácter influyente.
En el caso de la interpretación del modelo se puede decir que:

- El X aumenta más con la edad cuando las niñas solo reciben el tratamiento "W". Los efectos con este tratamiento son peores.
- El tratamiento "SF" no cambia significativamene X para niñas pequeñas y ayuda disminuir X para adolescentes.
- El tratamiento "APF" tiene la mejor influencia para X, es decir, ayuda a disminuirlo significativamente comparado a otros tratamientos. Este tipo del tratamiento es el más recomendable.
- Aunque el modelo no sea perfecto, nos ayuda a interpretar la influencia que tiene cada tratamiento para la salud de las niñas.

7 El código

```
data=read.table("data.txt",header=TRUE,dec=".") #los datos
summary(data)
X=data$A-data$B
E=data$age
T=data$treatmt

#########################################################

X[which(X==0)]=100 #manejar mejor la divisi n por los grupos
#los grupos para hacer gr ficas
w=T=="W"
w=w*1
wx=w*X
we=w*E
wx=wx[which(wx!=0)] #grupo con el tratamiento w
wx[which(wx==100)]=0
we=we[which(we!=0)]

sf=T=="SF"
sf=sf*1
sfx=sf*X
sfe=sf*E
sfx=sfx[which(sfx!=0)] #grupo con el tratamiento sf
sfx[which(sfx==100)]=0
```

```
sfe=sfe[which(sfe!=0)]

apf=T=="APF"
apf=apf*1
apfx=apf*X
apfe=apf*E
apfx=apfx[which(apfx!=0)] #grupo con el tratamiento apf
apfx[which(apfx==100)]=0
apfe=apfe[which(apfe!=0)]

X[which(X==100)]=0

#########################################################

#presntacion de datos

plot(we,wx,col="red")
plot(sfe,sfx,col="blue")
plot(apfe,apfx,col="green",xlab="E",ylab="X")

legend(x="bottomleft", legend=c("W", "SF","APF"), col=c("red", "blue","green"),pch=16
#, text.font=4, bg='lightblue')

points(we,wx,col="red",pch=16)
points(sfe,sfx,col="blue",pch=16)
points(apfe,apfx,col="green",pch=16)

title(main="")

plot(we,wx,col="red",pch=16,xlab="X",ylab="E")
legend(x="topleft", legend=c("W"),col=c("red"),title="Tratamiento:",pch=16)

plot(sfe,sfx,col="blue",pch=16,xlab="X",ylab="E")
legend(x="bottomleft", legend=c("SF"),col=c("blue"),title="Tratamiento:",pch=16)

plot(apfe,apfx,col="green",pch=16,xlab="X",ylab="E")
legend(x="bottomleft", legend=c("APF"),col=c("green"),title="Tratamiento:",pch=16)

#########################################################
mod1=lm(X~E+T) # el modelo sin interaccion
summary(mod1)

#los coeficientes para las rectas

Intercept=1.3925
apf1=Intercept
tsf1=Intercept+0.65350
tw1=Intercept+1.7210
a=0.0143 #es la gamma, igual para todos los grupos
```

```
#los graficos

par(mfrow=c(2,2))
plot(we,wx,col="red",pch=16,xlab="E",ylab="X")
abline(tw1,a,col="red")
title(main="Tratamiento␣W")

plot(sfe,sfx,col="blue",pch=16,xlab="E",ylab="X")
abline(tsf1,a,col="blue")
title(main="Tratamiento␣SF")


plot(apfe,apfx,col="green",pch=16,xlab="E",ylab="X")
abline(apf1,a,col="green")
title(main="Tratamiento␣APF")


plot(we,wx,col="red",pch=16,xlab="E",ylab="X")
points(sfe,sfx,col="blue",pch=16,xlab="E",ylab="X")
points(apfe,apfx,col="green",pch=16,xlab="E",ylab="X")
abline(tw1,a,col="red")
abline(tsf1,a,col="blue")
abline(apf1,a,col="green")
title("Todo")

##########################################################

mod2=lm(X~E+T+E:T)#modelo con interacci n
summary(mod2)

apfb=3.49733 #b-pendiente, a-gamma
apfa=-0.17022
twb=apfb-2.80142
twa=apfa+0.36977
tsfa=apfa+0.17565
tsfb=apfb-1.33896

#gr ficos

par(mfrow=c(2,2))
plot(we,wx,col="red",pch=16,xlab="E",ylab="X")
abline(twb,twa,col="red")
title(main="Tratamiento␣W")

plot(sfe,sfx,col="blue",pch=16,xlab="E",ylab="X")
abline(tsfb,tsfa,col="blue")
title(main="Tratamiento␣SF")


plot(apfe,apfx,col="green",pch=16,xlab="E",ylab="X")
```

```
abline(apfb,apfa,col="green")
title(main="Tratamiento␣APF")


plot(we,wx,col="red",pch=16,xlab="E",ylab="X")
points(sfe,sfx,col="blue",pch=16)
points(apfe,apfx,col="green",pch=16)
abline(twb,twa,col="red")
abline(tsfb,tsfa,col="blue")
abline(apfb,apfa,col="green")
title(main="Todo")

#####################################

#comparaci n de los modelos
anova(mod1,mod2)

#########################################
#validaci n del modelo con interacci n
par(mfrow=c(2,2))
plot(mod2)

rest=rstandard(mod2) #residuos estandarizados
hist(rest,freq=FALSE,xlab="Residuos␣estandarizados")
curve(dnorm,add=TRUE,col="red")
title(main="Histogram␣de␣residuos␣estandarizados")
shapiro.test(rest)

acf(residuals(mod2))

############################################

#detecci n de at picas
rest=rstandard(mod2) #residuos estandarizados
rest #ver los valores -> analizar las observaciones con el n mero 33, 15 y 55

#qqplot

a=qqnorm(rest)
plot(c(a$x[33],a$x[15],a$x[55]),c(rest[33],rest[15],rest[55]),col="red")
points(c(a$x[33],a$x[15],a$x[55]),c(rest[33],rest[15],rest[55]),col="red",pch=16)
qqnorm(rest,plot.it=TRUE)
qqline(rest)
plot(a,pch=16)

############################################

atipicax=X[55]
atipicaedad=E[55]
atipicatrat=T[55]
```

```
#eliminaci n del 55

X=X[-55]
E=E[-55]
T=T[-55]

plot(atipicaedad,atipicax,col="black")

mod3=lm(X~E+T+E:T)
summary(mod3)

#siguientes comandos necesito para comparar los resultados

X[which(X==0)]=100
w=T=="W"
w=w*1
wx=w*X
we=w*E
wx=wx[which(wx!=0)] #grupo con el tratamiento w
wx[which(wx==100)]=0
we=we[which(we!=0)]

sf=T=="SF"
sf=sf*1
sfx=sf*X
sfe=sf*E
sfx=sfx[which(sfx!=0)] #grupo con el tratamiento sf
sfx[which(sfx==100)]=0
sfe=sfe[which(sfe!=0)]

apf=T=="APF"
apf=apf*1
apfx=apf*X
apfe=apf*E
apfx=apfx[which(apfx!=0)] #grupo con el tratamiento sf
apfx[which(apfx==100)]=0
apfe=apfe[which(apfe!=0)]

X[which(X==100)]=0

#coeficientes para el modelo sin dicha observaci n

apfb=3.34692 #pendiente
apfa=-0.16505
twb=apfb-2.65102
twa=apfa+0.36460
tsfa=apfa+0.17048
tsfb=apfb-1.18854
```

```
#coeficientes para el modelo con dicha observaci n

apfb2=3.49733 #pendiente
apfa2=-0.17022
twb2=apfb2-2.80142
twa2=apfa2+0.36977
tsfa2=apfa2+0.17565
tsfb2=apfb2-1.33896

#construcci n de gr ficas

par(mfrow=c(2,2))
plot(we,wx,col="red",pch=16,xlab="E",ylab="X")
abline(twb,twa,col="red")
abline(twb2,twa2,col="black")
title(main="Tratamiento␣W")

plot(sfe,sfx,col="blue",pch=16,xlab="E",ylab="X")
abline(tsfb,tsfa,col="blue")
abline(tsfb2,tsfa2,col="black")
title(main="Tratamiento␣SF")


plot(apfe,apfx,col="green",pch=16,xlab="E",ylab="X")
points(atipicaedad,atipicax,col="black",pch=16)
abline(apfb,apfa,col="green")
abline(apfb2,apfa2,col="black")
title(main="Tratamiento␣APF")

plot(we,wx,col="red",pch=16,xlab="E",ylab="X")
points(sfe,sfx,col="blue",pch=16)
points(apfe,apfx,col="green",pch=16)
points(atipicaedad,atipicax,col="black",pch=16)
abline(twb,twa,col="red")
abline(tsfb,tsfa,col="blue")
abline(apfb,apfa,col="green")
title(main="Todo")
```