Examen

Maria Jose Corea

2022-10-20

1. La base contenida en carbohydrate.csv tiene observaciones con porcentaje de calor´ıas totales obtenidas de carbohidratos complejos (carbohydrate), edad (age), peso relativo (weight) y porcentaje de calor´ıas en prote´ınas (protein) para 20 pacientes de sexo masculino diabeticos y dependientes de insulina.

a) Considere como variable dependiente carbohydrate y el resto de variables como posibles covariables. Analice la posible asociacion entre variables de manera exploratoria

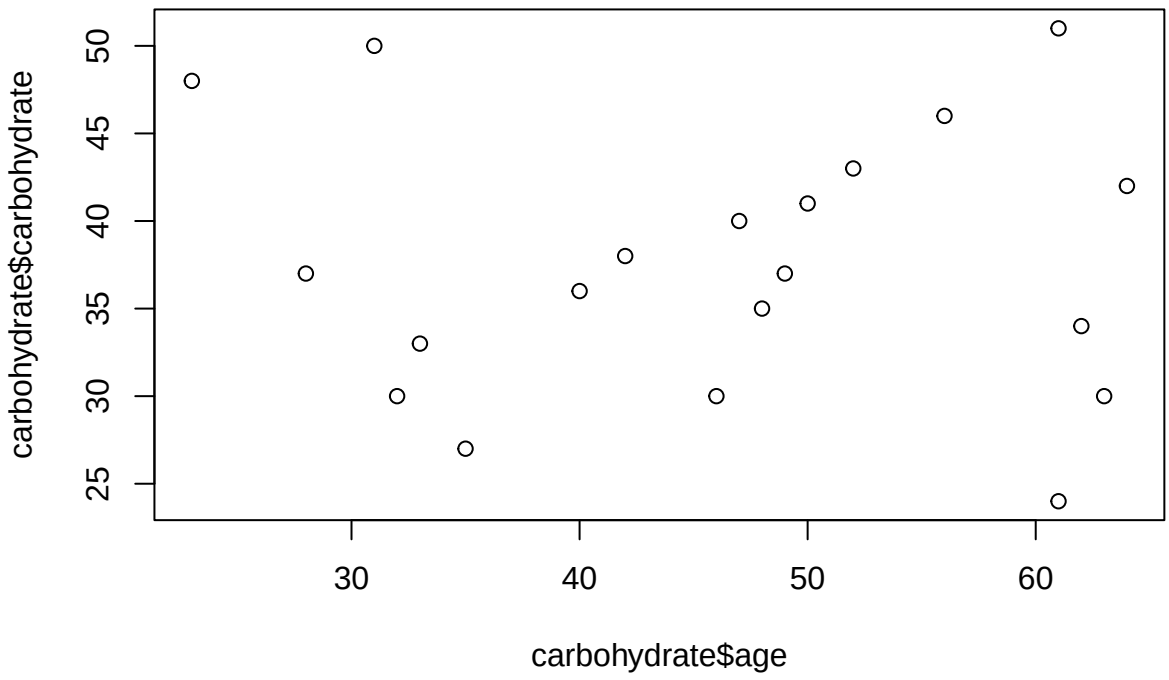
```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

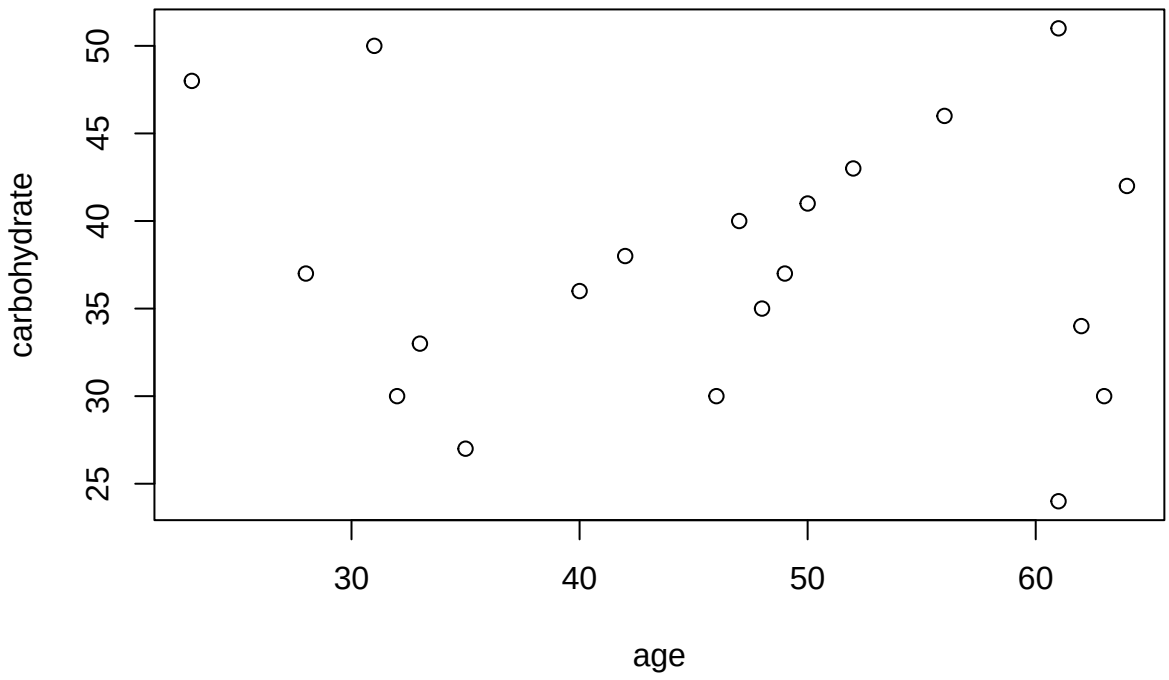
```
carbohydrate <- read_csv("carbohydrate.csv")
```

```
## Rows: 20 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): carbohydrate, age, weight, protein
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

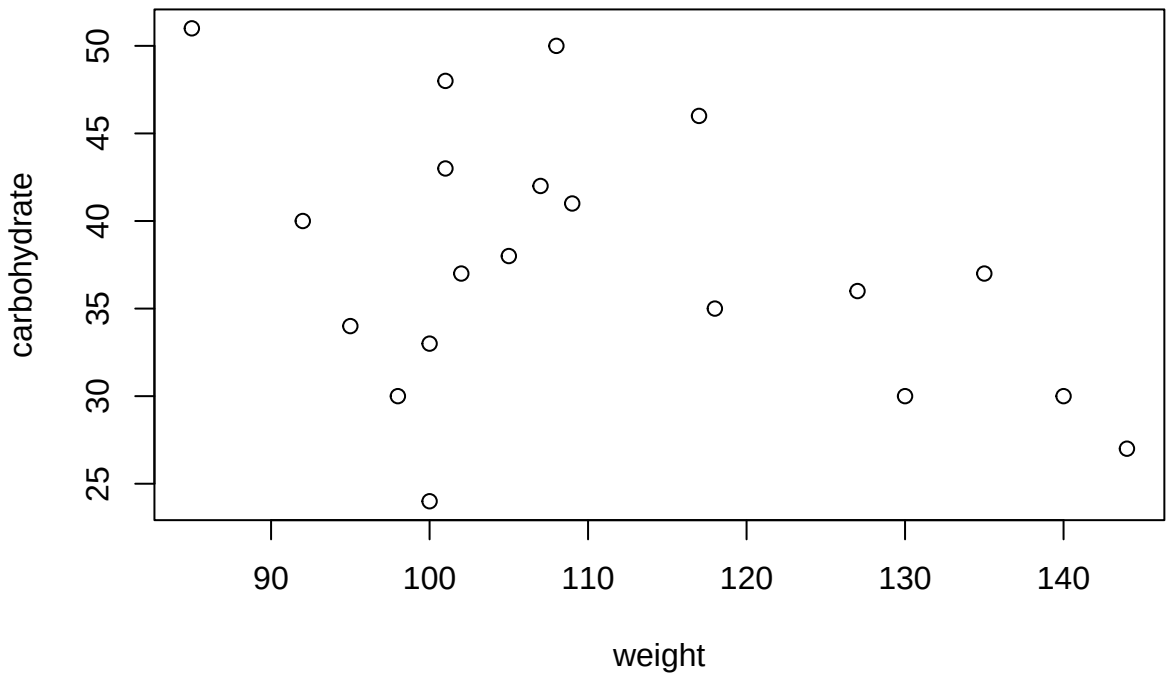
```
plot(carbohydrate$carbohydrate~carbohydrate$age)
```



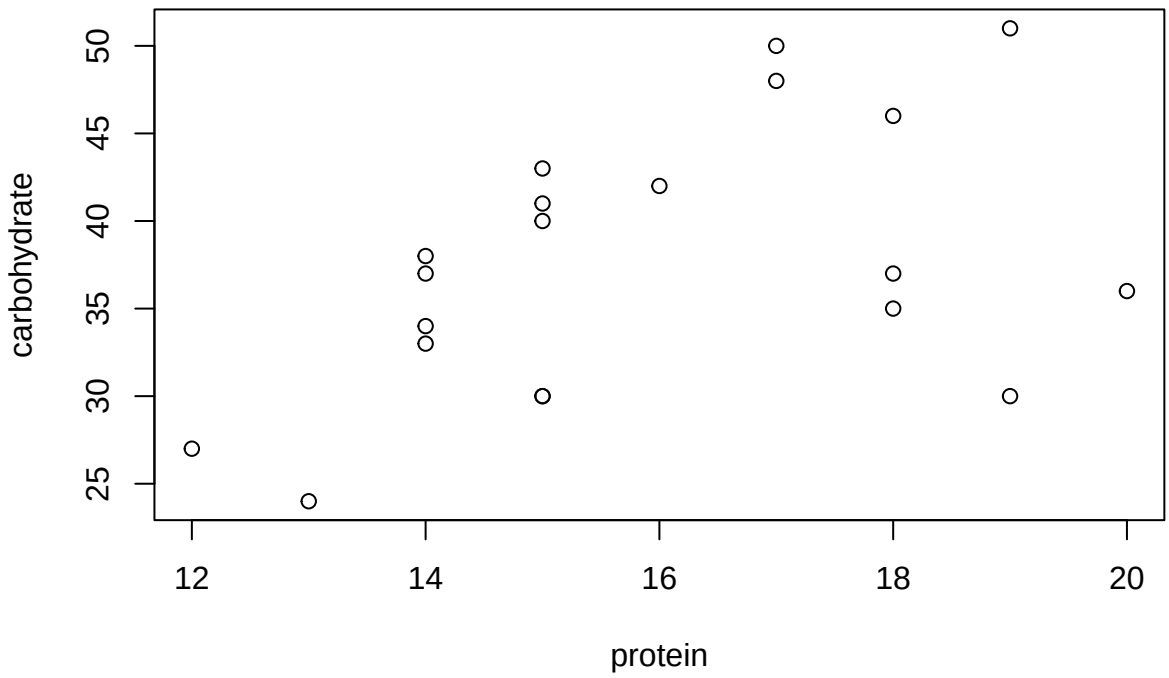

```
plot( carbohydrate~ age, data = carbohydrate)
```




```
plot(carbohydrate ~ weight, data = carbohydrate)
```




```
plot(carbohydrate ~ protein, data = carbohydrate)
```




```
summary(carbohydrate)
```

```
##   carbohydrate        age            weight         protein
##  Min.   :24.00   Min.   :23.00   Min.   : 85.0   Min.   :12.0
##  1st Qu.:32.25   1st Qu.:34.50   1st Qu.:100.0   1st Qu.:14.0
##  Median :37.00   Median :47.50   Median :106.0   Median :15.0
##  Mean   :37.60   Mean   :46.15   Mean   :110.7   Mean   :15.9
##  3rd Qu.:42.25   3rd Qu.:57.25   3rd Qu.:120.2   3rd Qu.:18.0
##  Max.   :51.00   Max.   :64.00   Max.   :144.0   Max.   :20.0
```

```
var(carbohydrate$age)
```

```
## [1] 163.1868
```

```
var(carbohydrate$carbohydrate)
```

```
## [1] 57.51579
```

```
var(carbohydrate$weight)
```

```
## [1] 276.6421
```

```
var(carbohydrate$protein)
```

```
## [1] 4.936842
```

Se ve una Mayor varianza en age y weight, los valores mínimos de carbohidratos son 24 y el maximo 51, con una media de 37.

Min. :24.00 Min. :23.00 Min. : 85.0 Min. :12.0
1st Qu.:32.25 1st Qu.:34.50 1st Qu.:100.0 1st Qu.:14.0
Median :37.00 Median :47.50 Median :106.0 Median :15.0
Mean :37.60 Mean :46.15 Mean :110.7 Mean :15.9
3rd Qu.:42.25 3rd Qu.:57.25 3rd Qu.:120.2 3rd Qu.:18.0
Max. :51.00 Max. :64.00 Max. :144.0 Max. :20.0
[1] 163.1868 [1] 57.51579 [1] 276.6421 [1] 4.936842

b) Usando AIC y seleccion para adelante, encuentre un modelo ´ optimo. Interprete todos ´ los estimadores de los coeficientes en el modelo lineal ajustado.

```
 min.model <- lm(carbohydrate ~ age +weight+protein, data=carbohydrate)
 max.model <- lm( carbohydrate ~ (protein+weight+age)^2, data=carbohydrate)
  auto.forward <- step( min.model, direction="forward",
scope=list(lower=min.model, upper=max.model) )
```

```
## Start:  AIC=74.92
## carbohydrate ~ age + weight + protein
##
##                   Df Sum of Sq    RSS    AIC
## + protein:weight  1   127.879 439.78 71.811
## <none>                        567.66 74.916
## + weight:age      1    35.444 532.22 75.626
## + protein:age     1     0.892 566.77 76.884
##
## Step:  AIC=71.81
## carbohydrate ~ age + weight + protein + weight:protein
##
##               Df Sum of Sq    RSS    AIC
## <none>                     439.78 71.811
## + protein:age  1   18.1714 421.61 72.967
## + weight:age   1    3.6986 436.09 73.642
```

```
   signif( coef(auto.forward), 3 )
```

```
##    (Intercept)            age         weight        protein weight:protein
##       -78.0000        -0.0948         0.7460         9.2900        -0.0625
```

El mejor aic es el de carbohydrate ~ weight:protein , puesto que arroja el valor menor

```
modelobase1<-lm(carbohydrate ~ weight+protein, data=carbohydrate)
summary(modelobase1)
```

```
##
## Call:
## lm(formula = carbohydrate ~ weight + protein, data = carbohydrate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6812  -3.9135   0.9464   4.0880   9.7948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.13032   12.57155   2.635  0.01736 *
## weight      -0.22165    0.08326  -2.662  0.01642 *
## protein      1.82429    0.62327   2.927  0.00941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.971 on 17 degrees of freedom
## Multiple R-squared:  0.4454, Adjusted R-squared:  0.3802
## F-statistic: 6.827 on 2 and 17 DF,  p-value: 0.006661
```

```
coef(modelobase1)
```

```
## (Intercept)      weight     protein
##   33.130320   -0.221649    1.824291
```

Un aumento en una unidad de peso, provoca una disminucion de más 0.221649 en la cantidad de carbohidratos, manteniendo las demás variables constante. Un unidad del peso sube, indica que la cantidad de carbihidratos baja en 0.22, nominalmente. Asimismo, un aumento en una unidad de proteina, provoca un aumento de más 1.824291 en la cantidad de carbohidratos, manteniendo las demás variables constante.

c) Interprete un intervalo de confianza al 90 % para el coeficiente de weight. Calcule directamente (sin usar summary) el estad´ıstico de la prueba F y su valor p e interpretelo.

```
confint(modelobase1, level = 0.9)
```

```
##                    5 %        95 %
## (Intercept) 11.2607641 54.99987533
## weight      -0.3664914 -0.07680669
## protein      0.7400395  2.90854280
```

Esto es que el 90% de las veces, casi seguramente, el coeficiente de la variable weight se va aencontrar entre -0.3664914 -0.07680669, el otro porcentaje restante, puede incurrir al error.

```
 Xmat<-model.matrix( ~ weight+protein,
data=carbohydrate)
 XtX <- t(Xmat) %*% Xmat # t() is transpose; %*% is matrix multiply
 y <- log(carbohydrate$carbohydrate)
 inv.XtX <- solve( XtX ) # solve returns the matrix inverse
 XtY <- t(Xmat) %*% y

  beta <- inv.XtX %*% XtY; drop(beta)
```

```
## (Intercept)      weight     protein
##  3.45905952 -0.00585892  0.05010726
```

```
beta <- solve(XtX, XtY); beta
```

```
##                   [,1]
## (Intercept)  3.45905952
## weight      -0.00585892
## protein      0.05010726
```

```
QR <- qr(Xmat)
beta <- qr.coef(QR, y); beta
```

```
## (Intercept)      weight     protein
##  3.45905952 -0.00585892  0.05010726
```

```
mu <- Xmat %*% beta
RSS <- sum( (y - mu)^2 );
RSS
```

```
## [1] 0.4519442
```

```
s2 <- RSS / ( length(carbohydrate$carbohydrate) - length(beta) )
c(s=sqrt(s2), s2=s2)
```

```
##         s         s2
## 0.16304893 0.02658495
```

```
 a<-predict(modelobase1,

         interval = "confidence",
         level = 0.90)



MsReg<-sqrt(mean((carbohydrate$carbohydrate - a[,1])^2))

Festadista<-MsReg/s2
Festadista<-MsReg/s2
 DF <- df.residual(modelobase1);
 DF
```

```
## [1] 17
```

```
 df.B <- df.residual(modelobase1);
 df.B
```

```
## [1] 17
```

```
   pf(Festadista, df1=DF, df2=df.B, lower.tail=FALSE)
```

```
## [1] 2.388905e-16
```

```
 summary(modelobase1)
```

```
##
## Call:
## lm(formula = carbohydrate ~ weight + protein, data = carbohydrate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6812  -3.9135   0.9464   4.0880   9.7948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.13032   12.57155   2.635  0.01736 *
## weight      -0.22165    0.08326  -2.662  0.01642 *
## protein      1.82429    0.62327   2.927  0.00941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.971 on 17 degrees of freedom
## Multiple R-squared:  0.4454, Adjusted R-squared:  0.3802
## F-statistic: 6.827 on 2 and 17 DF,  p-value: 0.006661
```

Bueno, los calculé mal, pero usando la intrepertración del summary, dice que

Con respecto a la base, dice que con la prueba t que la proteina es el estimador con mayor significancia y que la edad no tiene significancia. En caso del peso, tienen una significancia del 0.01. Entonces se debería ajustar el modelo y sacar la variable age, puesto que no es tan relevante. Asimismo, se ve que la p-valor se rechararía si la hipótesis nula fuera conservadora y sería del 5%, en caso de ser de un 1%, no se rechazaría. La f-test nos dice que la capacidad explicativa que tiene un grupo de variables independientes(peso y proteina) sobre la variación de la variable dependiente (carbohidrato) es de 6.827 .

d) Calcule usando la formula vista en clase, el intervalo de confianza al 90 % de una ´ prediccion de la variable dependiente si la edad del individuo es 57, su peso es 120 y ´ el nivel de prote´ına es 18. Compare el resultado obtenido con el que se obtiene a traves´ de la funcion´ predict. Interprete el intervalo de confianza.

```
# x <-(PromedioTemperatura = seq(min(x), max(x), by = 0.05))
# cosa <-
#   predict(mod_leo,
#           newdata = data.frame(x),
#           interval = "confidence",
#           level = 0.90)
```

e)

```
rstandard(modelobase1)#residual estandarizado
```

```
##           1           2           3           4           5           6
## -0.61987738 -0.01838129  0.17784413  0.91102922  0.10281376  0.85114661
##           7           8           9          10          11          12
## -0.64685191  1.09379028 -1.53635371  0.45723010  1.69736401  0.41521270
##          13          14          15          16          17          18
## -1.68267215 -1.06164246  0.80534312  0.58471574  1.05322676 -1.94109159
##          19          20
## -0.85058702  0.16528856
```

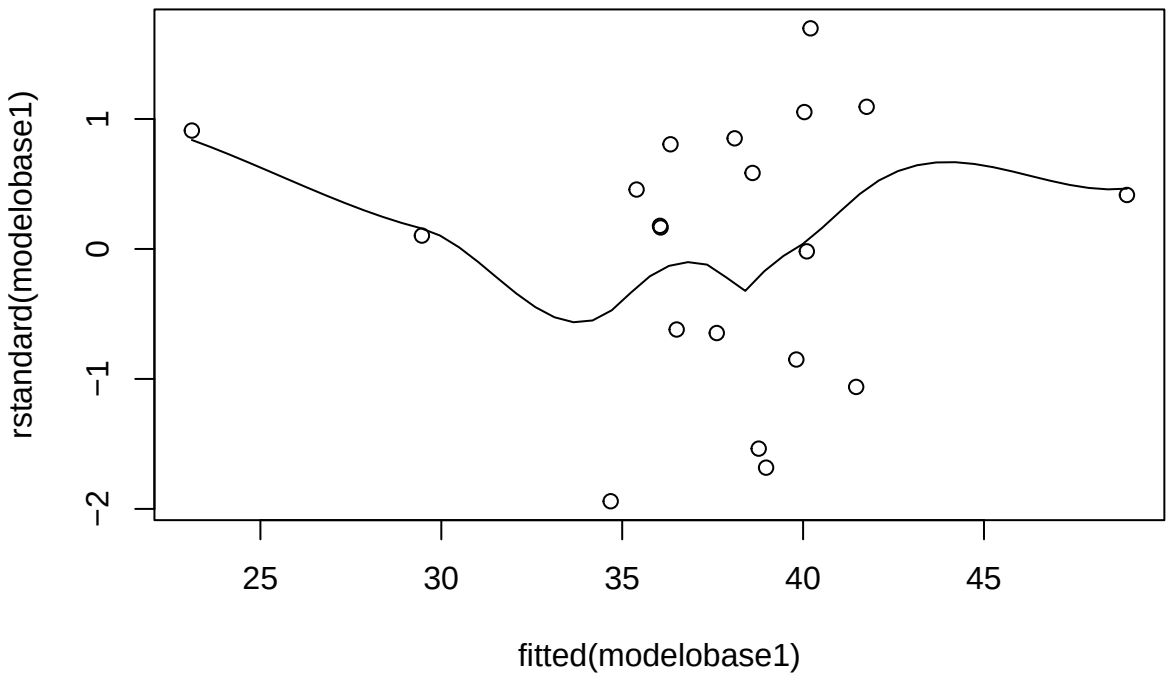
```
fitted(modelobase1) #Este ejemplo demuestra cómo encontrar los valores ajustados de un modelo de regres
```

```
##        1        2        3        4        5        6        7        8
## 36.50549 40.10298 36.04494 23.10435 29.46382 38.10813 37.61374 41.75672
##        9       10       11       12       13       14       15       16
## 38.77308 35.39725 40.20517 48.95168 38.97748 41.46671 36.33494 38.60253
##       17       18       19       20
## 40.03462 34.68120 39.81297 36.06219
```

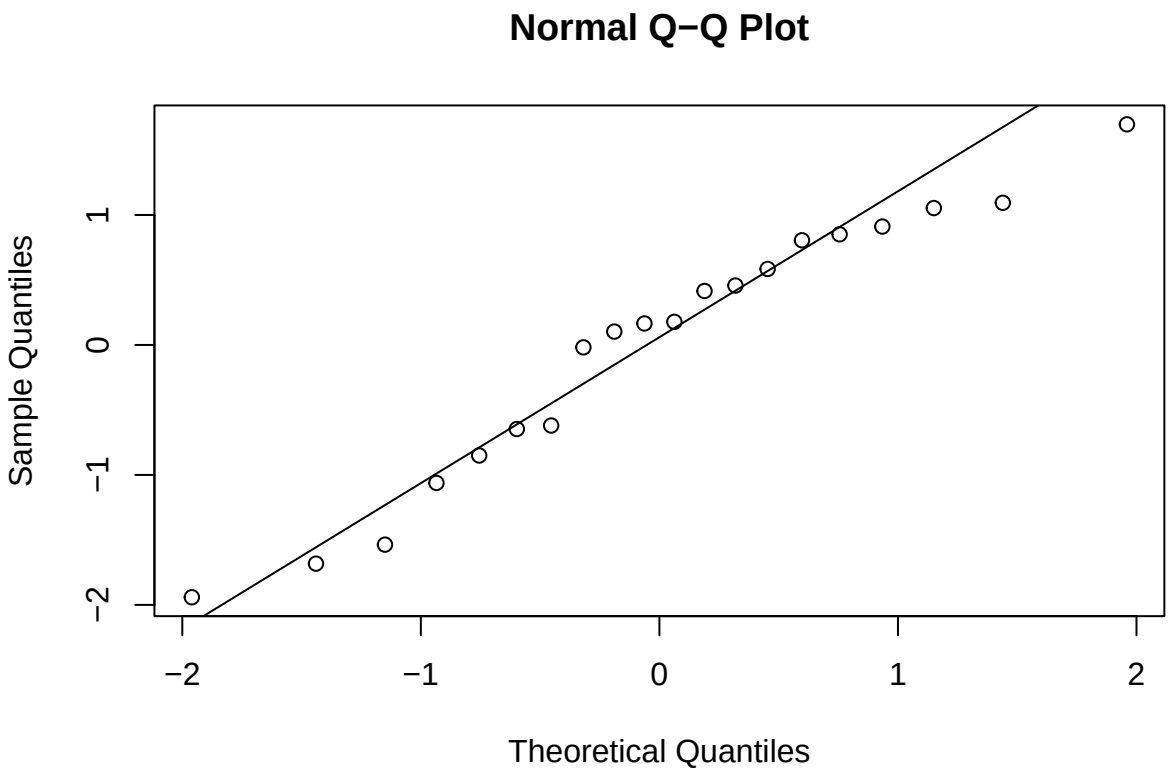
```
cooks.distance(modelobase1) #La distancia de Cook es un resumen de cuánto cambia un modelo de regresión
```

```
##            1            2            3            4            5            6
## 1.468948e-02 1.530182e-05 2.489264e-03 2.627132e-01 1.094970e-03 1.912373e-02
##            7            8            9           10           11           12
## 1.983076e-02 3.755081e-02 7.336478e-02 6.977511e-03 6.772622e-02 2.671230e-02
##           13           14           15           16           17           18
## 2.381781e-01 1.294064e-01 1.348990e-02 6.369233e-03 4.113183e-02 2.226934e-01
##           19           20
## 2.734748e-02 9.779275e-04
```

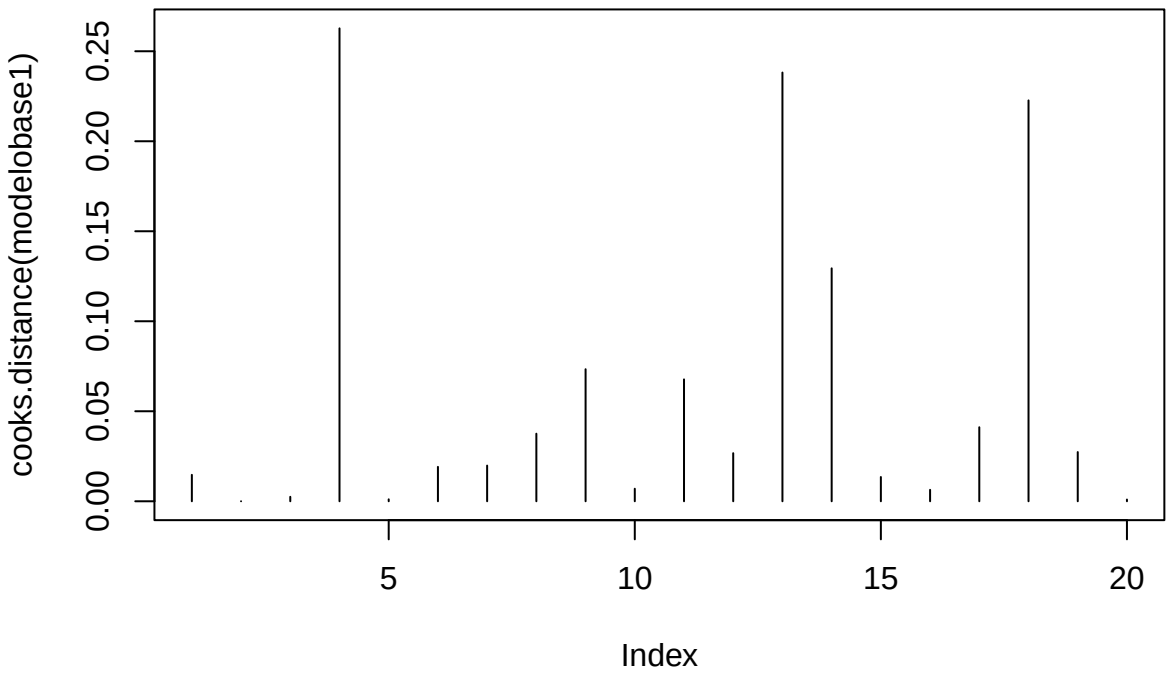
```
 scatter.smooth( rstandard(modelobase1) ~ fitted(modelobase1) )
```




```
qqnorm( rstandard(modelobase1) )
qqline( rstandard(modelobase1) )
```




```
plot( cooks.distance(modelobase1), type="h")
```



Se ve una tendencia de residuo estandarizado entre 35 y 40. No se ve tan claro la tedencia lineal. Si se pone una barrera en el -1.5 se ven que 3 observaciones quedarían afuer, y el resto se concentraría arriba de -1.5.

```
influence.measures(modelobase1)$is.inf
```

```
##    dfb.1_ dfb.wght dfb.prtn dffit cov.r cook.d   hat
## 1   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 2   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 3   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 4   FALSE    FALSE    FALSE FALSE  TRUE  FALSE  TRUE
## 5   FALSE    FALSE    FALSE FALSE  TRUE  FALSE FALSE
## 6   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 7   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 8   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 9   FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 10  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 11  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 12  FALSE    FALSE    FALSE FALSE  TRUE  FALSE FALSE
## 13  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 14  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 15  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 16  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 17  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 18  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 19  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
## 20  FALSE    FALSE    FALSE FALSE FALSE  FALSE FALSE
```

```
rowSums(influence.measures(modelobase1)$is.inf)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  0  0  0  2  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
```

Y se ven 3 outliers, esto con el modelo de cooks distance, el de la observación 4,5 y 12,se ve un apalancamiento en la observación 5, el cual sería el único. S eobserva con el normal plot no una clara linealidad, esto puede ser debido a los outliers y el apalancamiento encontrado.

2. En el archivo doctors.csv se tiene la cantidad de muertes de medicos fumadores y no- ´ fumadores ordenados por grupo de edad a partir de un seguimiento que se le hizo a grupos de medicos en el transcurso de 10 a ´ nos. La variable ~ person-years denota la cantidad de medicos ´ participantes por periodo de observacion (unidades: personas-a ´ no)

```
library(readr)
doctors <- read_csv("doctors.csv")
```

```
## Rows: 10 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): age, smoking
## dbl (2): deaths, person-years
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

a) Grafique la tasa de muerte estimada por 100000 personas-anos de muerte (~ personyears) con respecto a edad (age). Repita este grafico de manera separada para fumado- ´ res y no fumadores (smoking). Interprete los graficos y justifique el uso de un modelo ´ de conteo para estudiar la tasa de muerte.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

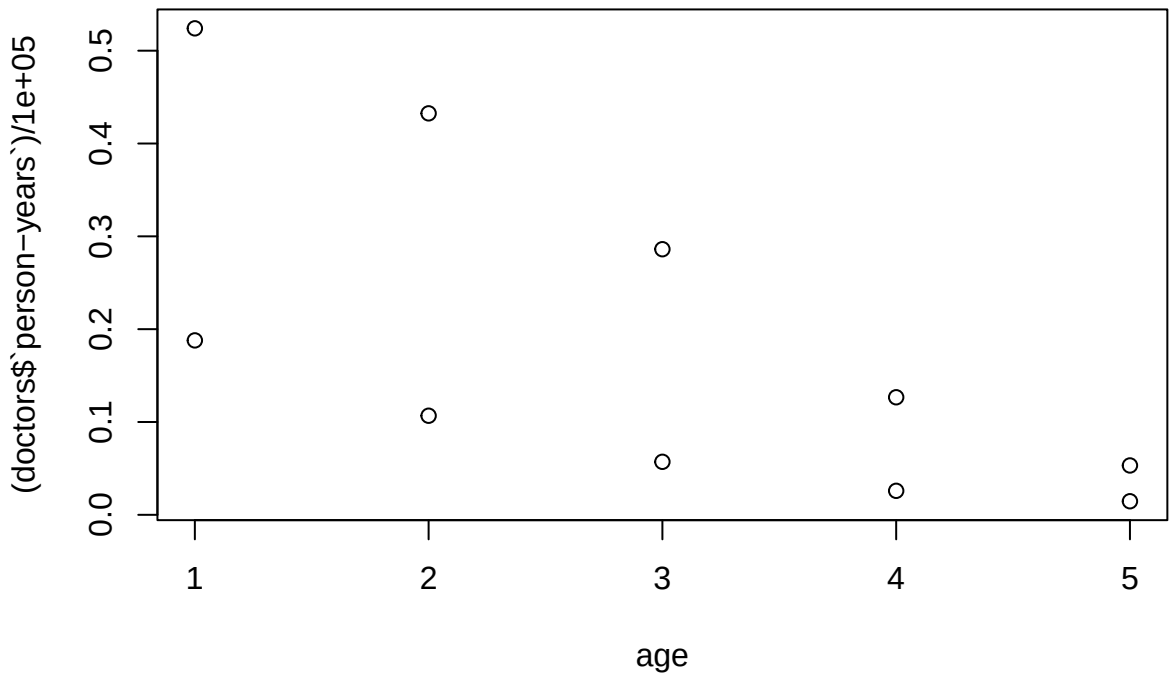
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
doctors1 <- doctors%>%mutate(age = case_when( age == '35 to 44' ~ 1,
                                              age == '45 to 54' ~ 2,
                                              age == '55 to 64' ~ 3,
                                              age == '65 to 74' ~ 4,
                                              age == '75 to 84' ~ 5)

                                    )
plot((doctors$`person-years`)/100000 ~age, data=doctors1)
```



Se observa para esta base que la mayo tasa de muerte en el rango de 35 a 44, e increblemete con forme aumenta la edad baja la tasa de muerte.

```
# plot(smoking) ~age, data=doctors1)
```

b) Ajuste un modelo de conteo apropiado para la tasa de muerte con las covariables smoking y age, recodificando la variable age de la siguiente forma: 35-44: 1, 45-54: 2, 55-64: 3, 65-74: 4 y 75-84: 5. Verifique si una interaccion entre la edad codificada y la ´ variable smoking es significativa.

```
doctors1 <- doctors%>%mutate(age = case_when( age == '35 to 44' ~ 1,
                                              age == '45 to 54' ~ 2,
                                              age == '55 to 64' ~ 3,
                                              age == '65 to 74' ~ 4,
                                              age == '75 to 84' ~ 5)

                                    )
```

```
glm(formula = (doctors1$deaths)/100000  ~ age+smoking, family = binomial,
    data = doctors1, weights = doctors1$`person-years`)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
##
## Call:  glm(formula = (doctors1$deaths)/1e+05 ~ age + smoking, family = binomial,
```

```
##     data = doctors1, weights = doctors1$`person-years`)
##
## Coefficients:
##   (Intercept)            age  smokingsmoker
##       -9.9922         0.4105         2.1484
##
## Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
## Null Deviance:       115.1
## Residual Deviance: 29.57     AIC: 71.05
```

```
modelopos<- glm((doctors1$deaths)/100000 ~ age + smoking, family=poisson,
data=doctors1 )
```

```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.000320

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.001040

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.002060

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.001860

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.001020

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.000020

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.000120

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.000280

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.000280

## Warning in dpois(y, mu, log = TRUE): non-integer x = 0.000310
```

c) Analice los residuos del modelo y verifique de manera descriptiva e inferencial que una transformacion cuadr ´ atica sobre la edad codificada es significativa.

d)Analice de manera descriptiva la posible existencia de sobredispersion. Ajuste un mo- ´ delo con sobredispersion y comp ´ arelos a trav ´ es de BIC. ¿Cu ´ al es mejor y c ´ omo se ´ relaciona ese resultado con el analisis descriptivo que hizo?

```
min(doctors1$deaths)
```

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

Para ver la sobredispersión see necesitaría ver el punto de silla, en el caso binomial sería la suma de las partes mayor o igual a 3 y para usar pearson se necesitaría el teorema del limite central, o sea la suma de las partes mayor o igual a 5. EN el caso de possson ser mayor o igual a 3 el valor mpinimo del conteo o mayor igual a 5, pata el tlc y usar pearson, en caso de que la devianza falle o no indique nada.

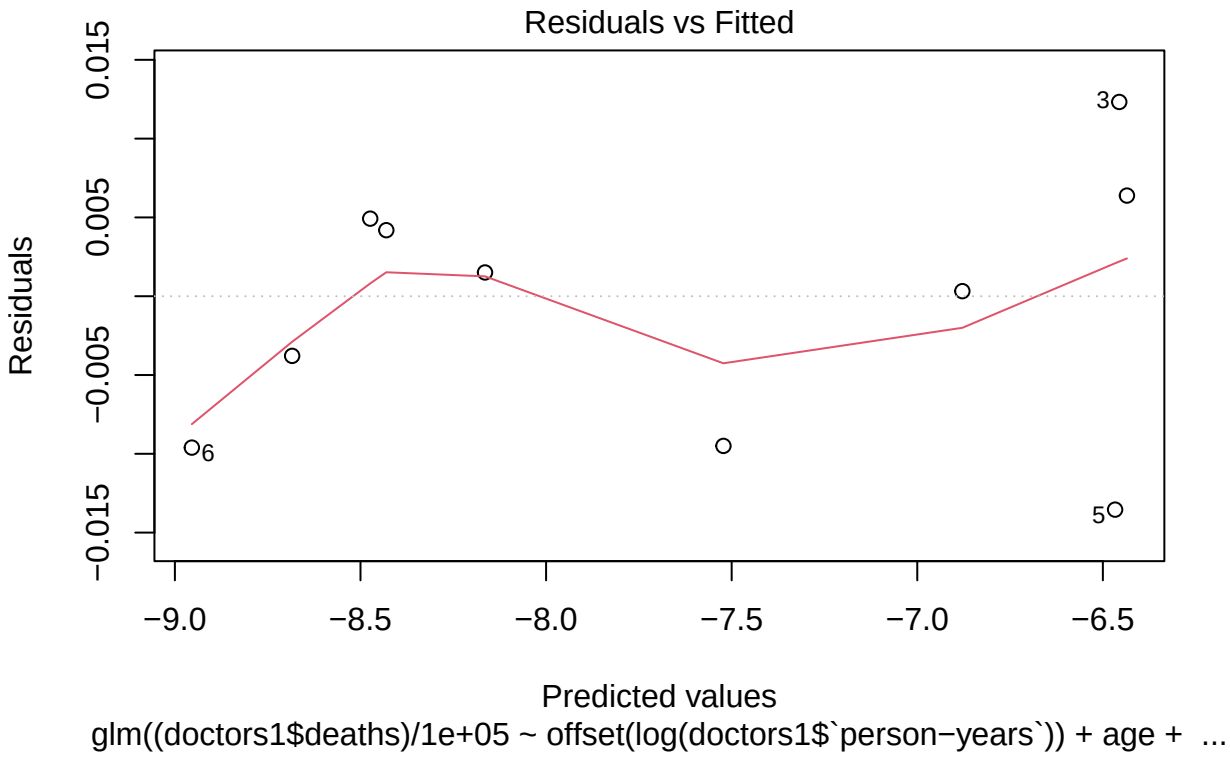
No pude calcular le bic por falta de tiempo, pero este se ve con el valor menor, en la comparación de modelos, además de que se ajusta con h, que es un apalancamiento

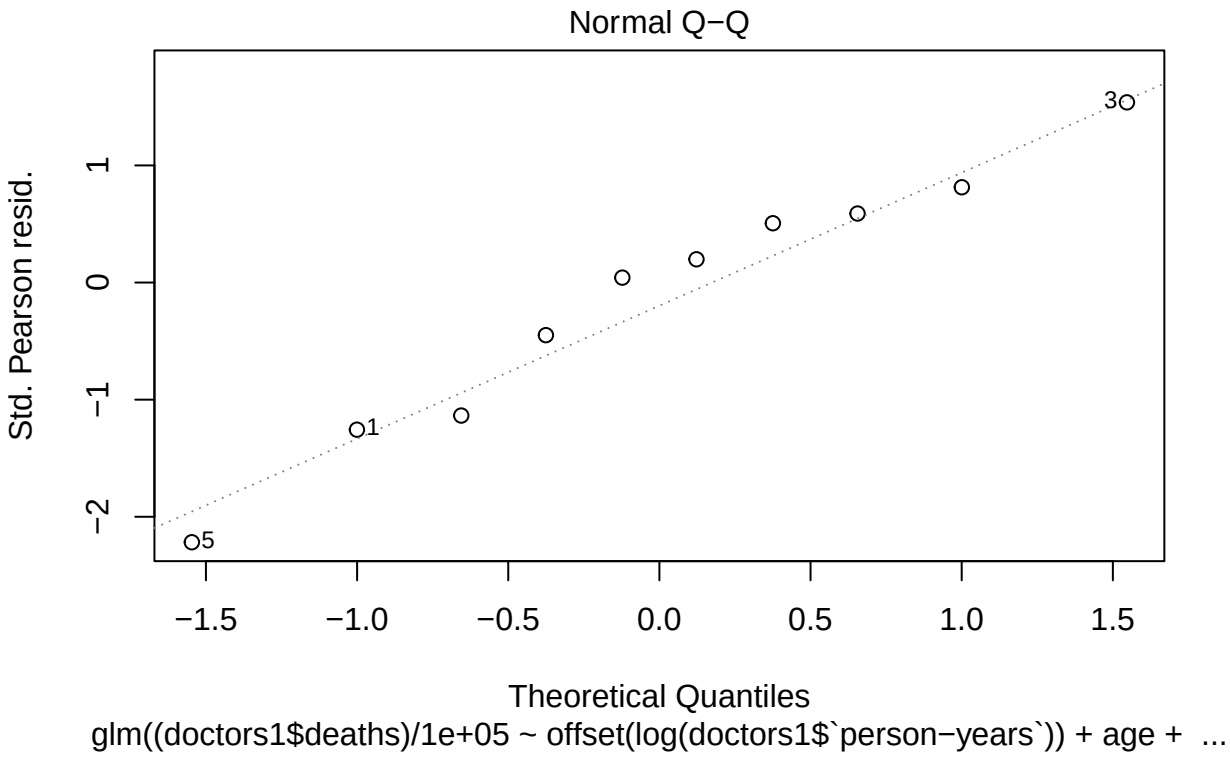
e) Con el modelo resultante en el punto anterior, interprete en cuanto aumenta o disminuye ´ el riesgo de muerte de una persona en el cuarto grupo de edad cuando es fumador con respecto a que no lo sea.

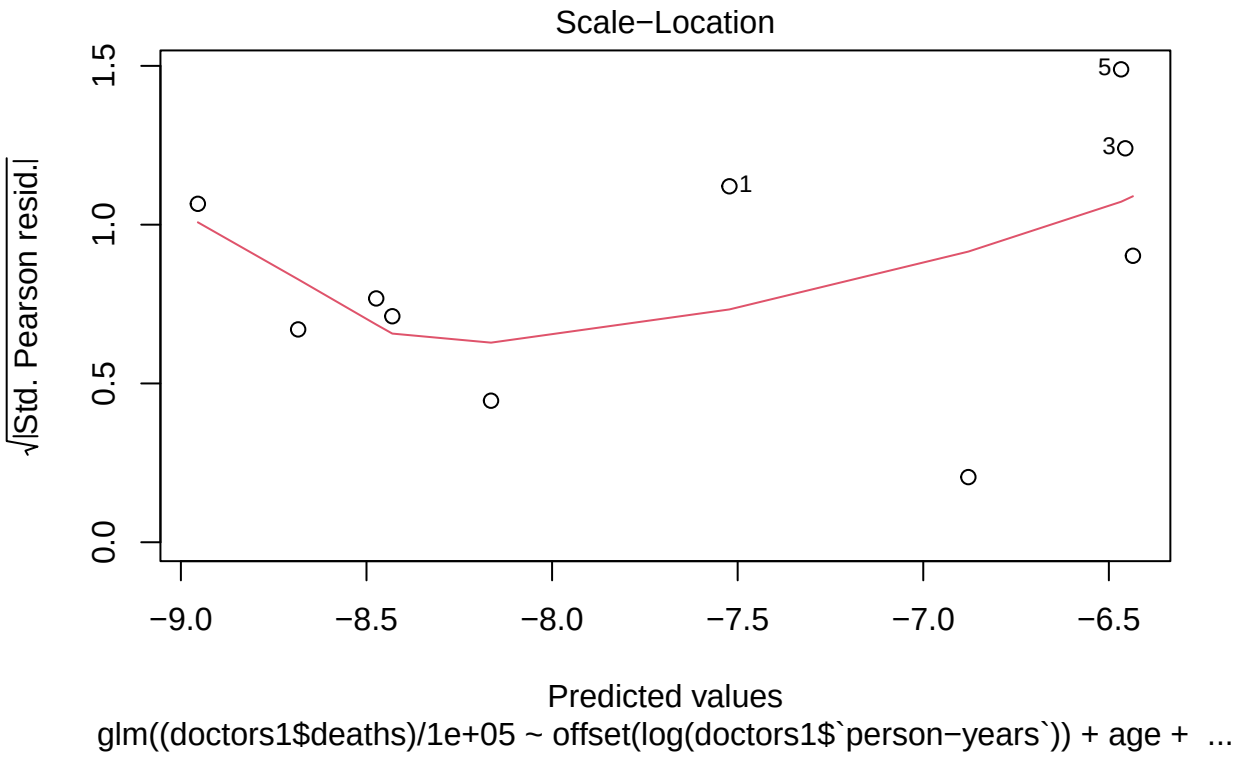
```r
fit10.11quasi <- glm( (doctors1$deaths)/100000 ~ offset(log(doctors1$`person-years`) )+age + smoking,
family="quasipoisson", data=doctors1)
summary(fit10.11quasi)
```

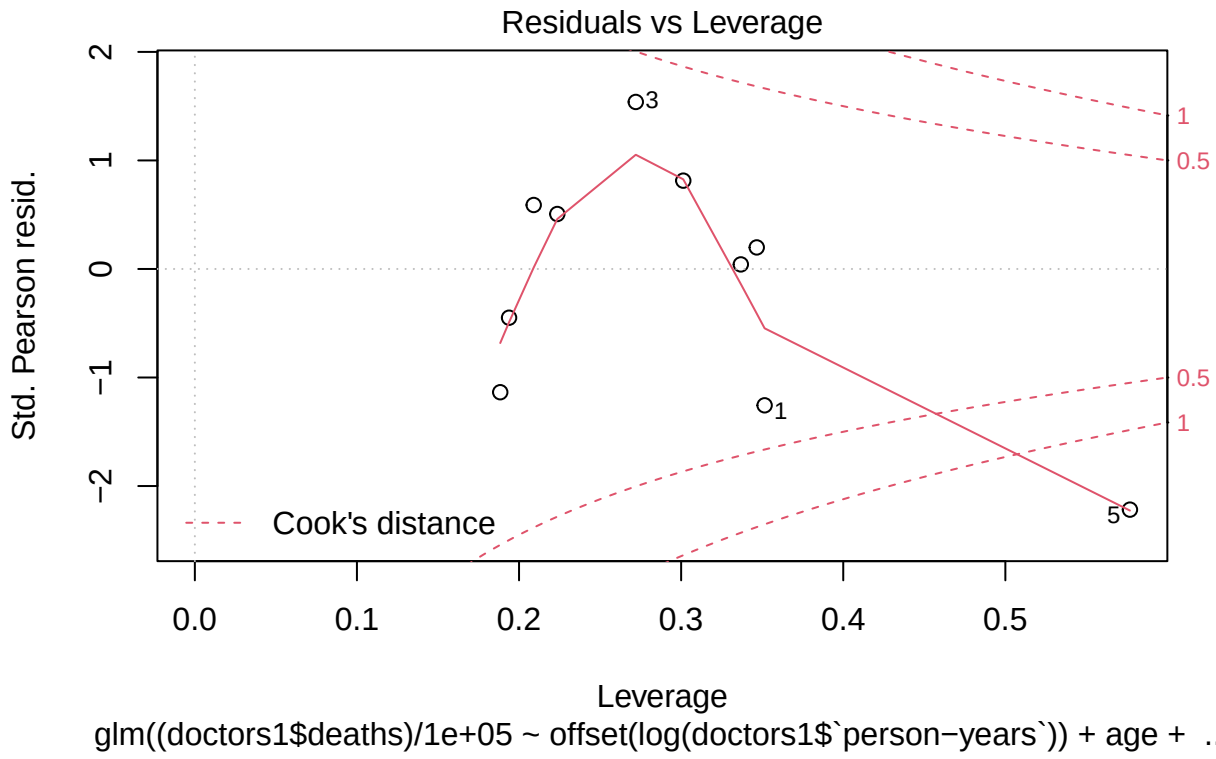
```
##
## Call:
## glm(formula = (doctors1$deaths)/1e+05 ~ offset(log(doctors1$`person-years`)) +
##     age + smoking, family = "quasipoisson", data = doctors1)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -0.0144555  -0.0087158   0.0009035   0.0045096   0.0117583
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -19.63125    0.41357 -47.468 4.82e-10 ***
## age             0.83583    0.08624   9.692 2.63e-05 ***
## smokingsmoker   0.40637    0.31826   1.277    0.242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 8.81551e-05)
##
##     Null deviance: 0.00935067  on 9  degrees of freedom
## Residual deviance: 0.00069182  on 7  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 10
```

```r
plot(fit10.11quasi)
```


Residuals vs Fitted
Residuals
0.015
0.005
−0.005
−0.015
3
6
5
−9.0
−8.5
−8.0
−7.5
−7.0
−6.5
Predicted values
glm((doctors1$deaths)/1e+05 ~ offset(log(doctors1$`person−years`)) + age + ...


Normal Q−Q
Std. Pearson resid.
−2
−1
0
1
3
1
5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
glm((doctors1$deaths)/1e+05 ~ offset(log(doctors1$`person−years`)) + age + ...


Scale−Location
√|Std. Pearson resid.|
1.5
1.0
0.5
0.0
5
3
1
−9.0
−8.5
−8.0
−7.5
−7.0
−6.5
Predicted values
glm((doctors1$deaths)/1e+05 ~ offset(log(doctors1$`person−years`)) + age + ...




```
influence.measures(fit10.11quasi)$is.inf
```

```
##    dfb.1_ dfb.age dfb.smkn dffit cov.r cook.d   hat
## 1   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 2   FALSE   FALSE    FALSE FALSE  TRUE  FALSE FALSE
## 3   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 4   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 5    TRUE    TRUE     TRUE  TRUE FALSE   TRUE FALSE
## 6   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 7   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 8   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 9   FALSE   FALSE    FALSE FALSE FALSE  FALSE FALSE
## 10  FALSE   FALSE    FALSE FALSE  TRUE  FALSE FALSE
```

No se observa aplancamiento, pero si se ven valores de influencia

```
printCoefmat(exp(coef(summary(fit10.11quasi))),digits=4)
```

```
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)   2.980e-09  1.512e+00     0.000     1.00
## age           2.307e+00  1.090e+00 16192.572     1.00
## smokingsmoker 1.501e+00  1.375e+00     3.585     1.27
```

Los odds son la razón de la prob de exito, entre la prob de fracaso. En este caso como la variable es categórita, entonces se ve como, no se ve el cons, entonces ese el base,el1.501e+00 de más de los que no fuman.