Feigl y Zelen (1965) analizaron datos de 33 pacientes con leucemia para los que se anotó si su tiempo de supervivencia era superior a 52 semanas (de hecho, ellos anotaron el tiempo de supervivencia y no sólo si era o no mayor a 52 semanas), que correspondería a un valor igual a 1, *éxito*, **de la variable dependiente Y, con probabilidad p, siendo $Y = 0$ si ese tiempo de supervivencia era inferior o igual a 52 semanas,** *fracaso*, **de probabilidad $1 - p$.**

Como covariables independientes que se piensa pueden explicar a ésta, se consideraron la covariable WBC, *número de glóbulos blancos por milímetro cúbico de sangre*, **(o leucocitos, o en inglés White Blood Cell Count) indicando un valor alto de esta covariable la existencia de infección, y la covariable AG,** *presencia* $(AG = 1)$ **o** *ausencia* $(AG = 0)$ *de cierta característica morfológica de los glóbulos blancos*. **A estos datos se ajustará en Modelo de Regresión Logística clásico y otro robusto.**

Para los datos de Feigl y Zelen (1965) se pretende ajustar un Modelo de Regresión Logística (clásico primero) de la forma

$$\log \frac{p}{1-p} = \beta_0 + \beta_1\, WBC + \beta_2\, AG$$

Los datos observados aparecen en el fichero de datos `leucemia`, proporcionado entre el Material Didáctico del curso. (Los valores de WBC del fichero fueron divididos por 10^4 con lo que habrá que multiplicarlos por esta cantidad en la fórmula del modelo ajustado.)

Como los datos a utilizar deben de estar en forma de *estructura de datos*, ejecutamos `(1)` para incluirlos en R^{mo} con ese formato al utilizar la función `read.table`. A continuación lo comprobamos.

```
> leucemia<-read.table("a:\\leucemia",header=T)                         (1)

> leucemia
   Super    WBC AG
1      1  0.230  1
2      1  0.075  1
3      1  0.430  1
................................................
32     0 10.000  0
33     0 10.000  0
```

Ahora, en `(2)`, utilizamos la función `glm` apareciendo los resultados en `(3)`, los cuales valoramos ejecutando `(4)`.

```
> solu<-glm(Super~WBC+AG,family=binomial,data=leucemia)                         (2)

> solu                                                                          (3)

Call:  glm(formula = Super ~  WBC + AG, family = binomial, data=leucemia)

Coefficients:
(Intercept)          WBC           AG
    -1.3074      -0.3177       2.2611

Degrees of Freedom: 32 Total (i.e. Null);  30 Residual
Null Deviance:     42.01
Residual Deviance: 31.06        AIC: 37.06

> summary(solu)                                                                 (4)

Call:
glm(formula = Super ~ WBC + AG, family = binomial, data = leucemia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5224  -0.6417  -0.4534   0.8362   2.1569

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3074     0.8140  -1.606   0.1083
WBC          -0.3177     0.1856  -1.712   0.0870 .                              (7)
AG            2.2611     0.9517   2.376   0.0175 *

                (5)        (6)
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ''  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 42.010  on 32  degrees of freedom
Residual deviance: 31.062  on 30  degrees of freedom
                     (8)
AIC: 37.062

Number of Fisher Scoring iterations: 4
```

Los estimadores de los coeficientes aparecen en (5), sus errores estándar en (6) (iguales a los que aparecen en la Tabla 7.1 del texto de Maronna, Martin y Yohai,

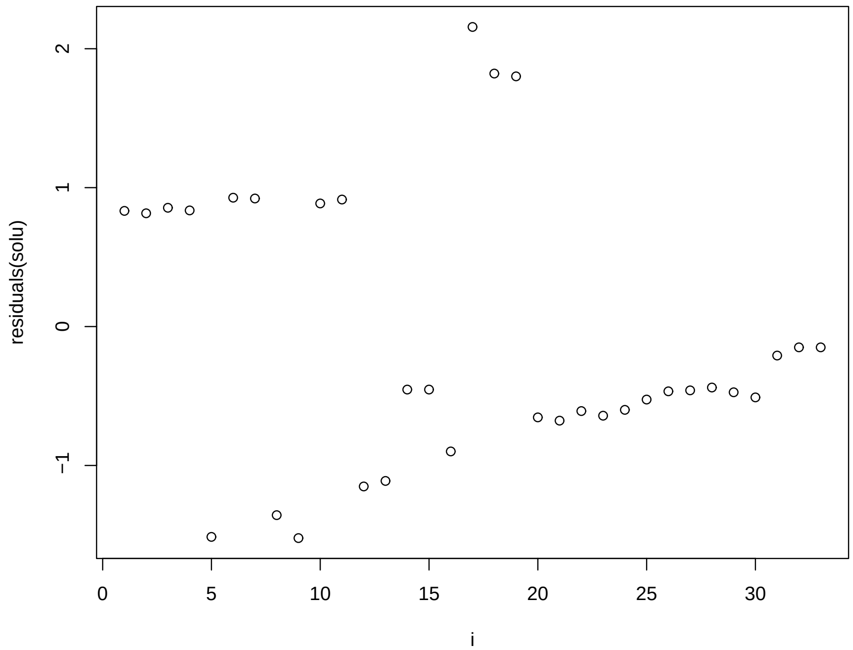

Figura 1: : Gráfico de los Residuos

2006, página 237) y los p-valores de los contrastes de la hipótesis nula de ser éstos cero, parecen indicar en (7) que no son significativas (es decir, que no se deberían de aceptar) las dos covariables independientes consideradas (con dudas podría serlo AG). Si se aceptaran ambas, quedaría como modelo ajustado el siguiente,

$$\log \frac{p}{1-p} = -1'3074 - 0'3177\, WBC(\times 10000) + 2'2611\, AG.$$

El valor del estadístico *deviance* que aparece en (8), igual a $G^2 = 31'062$, se utiliza en el contraste de la hipótesis nula de adecuarse correctamente el modelo anterior a los datos observados y que corresponde a una $\chi^2_{n-(k+1)} = \chi^2_{33-3} = \chi^2_{30}$. El p-valor de este test será, por tanto,

```
> 1-pchisq(31.062,30)
[1] 0.4123636
```

indicando que debe aceptarse, por contra, la bondad del ajuste del modelo obtenido.

Si representamos los residuos del modelo ajustado en la Figura 1 mediante la siguiente secuencia,

```
> i<-seq(1,33)
> plot(i,residuals(solu))
```

observamos que el dato número 17 es una observación influyente (un outlier). De hecho corresponde a un individuo con cien mil glóbulos blancos (lo que parece indicar que existe infección), pero que sorprendentemente sobrevivió más de 52 semanas. Las observaciones 18 y 19 son también un tanto atípicas puesto que son individuos que han sobrevivido mucho tiempo y tienen un valor $AG = 0$.

Para ajustar un Modelo de Regresión Logística Robusto, primero fijamos el valor de la constante de Huber en $1'2$ utilizamos la función `glm.rob` en la estimación robusta de los parámetros de la Regresión Logística, los cuales obtenemos en `(1)`.

```
> chuber<-1.2

> B<-glm.rob(as.matrix(leucemia[,c(2,3)]),as.matrix(leucemia[,c(1)]),
+ choice="logit")

> B$coeff
[1]  0.1646176 -2.0318031  2.4926958                                    (1)
```

Si ahora queremos analizar con cuál modelo nos quedamos, podemos hacer contrastes anidados, como los que hicimos en el ejemplo anterior, en los que estableceremos como hipótesis alternativa un modelo con un número determinado de covariables y como hipótesis nula un submodelo de éste. Si rechazamos la hipótesis nula, con un p-valor bajo, podemos concluir que la covariable no incluida en el modelo de la hipótesis nula (en el submodelo) es relevante a la hora de explicar a la variable dependiente. Todo esto lo haremos con la función anterior `quasi.rob`

Primero platearemos la hipótesis alternativa de un modelo con las dos covariables consideradas, WBC y AG frente a la hipótesis nula del submodelo sin la covariable AG. Es decir, contrastaremos las hipótesis

$H_0 : WBC$
$H_1 : WBC, AG$

Para ello ejecutamos la secuencia siguiente en donde destacamos como en la línea marcada con `(2)` incluimos, como primer argumento de la función, un modelo con las dos covariables que aparecen en las columnas 2 y 3 de la matriz de datos, y como en la línea `(3)` le decimos, con el argumento `out.col=2`, que como hipótesis nula considere el submodelo sin la covariable que aparece en la columna 2 de las anteriores, es decir, en la columna 3 de la matriz de datos, es decir, sin AG.

El p-valor de este test lo obtenemos ejecutando `(4)` que no es concluyente en cuanto al rechazo de la hipótesis nula del submodelo (desde luego la rechaza para un nivel de significación $0'05$), indicando cierta significación (i.e., algo explica) la covariable AG.

```
> a1<-quasi.rob(as.matrix(leucemia[,c(2,3)]),as.matrix(leucemia[,c(1)]),    (2)
out.col=2,choice="logit")                                                   (3)

> a1$pvalue                                                                 (4)

             [,1]
  [1,] 0.04645812
```

Ahora contrastaremos la otra posibilidad cual es la de eliminar la covariable WBC, es decir, contrastar las hipótesis

$H_0 : AG$
$H_1 : WBC, AG$

Para ello ejecutamos la siguiente sentencia indicándole en `(5)`, que ahora no considere la covariable que aparece en el lugar 1 del la matriz previa de datos de las covariables; es decir, la de la columna 2 de la matriz de datos, es decir, que prescinda en la hipótesis nula de WBC.

El p-valor lo obtenemos ejecutando `(6)`, el cual indica que se puede aceptar la hipótesis nula y prescindir de la covariable WBC.

```
> a2<-quasi.rob(as.matrix(leucemia[,c(2,3)]),as.matrix(leucemia[,c(1)]),
out.col=1,choice="logit")                                                   (5)


> a2$pvalue                                                                 (6)

            [,1]
  [1,] 0.1371982
```

Por tanto, como ya hemos decidido quedarnos sólo con la covariable AG, volvemos a ajustar el modelo de Regresión Logístico robusto ejecutando

```
> glm.rob(as.matrix(leucemia[,c(3)]),as.matrix(leucemia[,c(1)]),
+ choice="logit")$coeff

[1] -1.945900  2.063683
```

quedándonos, por tanto, con el modelo de Regresión Logística robusto

$$\log \frac{p}{1-p} = -1'9459 + 2'063683\, AG.$$