a) Establecer un modelo para pronosticar la ocurrencia de AML en esta población infantil, y a partir de las variables disponibles en la base de datos determinar el modelo con los factores de riesgo más relevantes.
b) Estudiar la validez del modelo estimado.
c) Discutir la capacidad del modelo obtenido para pronosticar correctamente la AML en esta población infantil y un punto de corte adecuado para la clasificación en ambos grupos de pacientes a partir del modelo ajustado.
d) Comentar las conclusiones del análisis

Teniendo en cuenta que hay numerosas variables a tener en cuenta y que casi todas las variables son de tipo cualitativo, el modelo que se ajusta mejor a la naturaleza de estos datos es el modelo logístico ya que pretendemos elaborar un sistema de discriminación de carácter dicotómico. A continuación, analizaremos la relevancia de las variables disponibles dentro del modelo. Para ello, se aplicará el modelo logit de regresión logistica múltiple.

Para comenzar, se considerará un modelo nulo en el que solo aparece la variable respuesta (*disease*) y lo compararemos con un modelo completo en el que incluimos todos las covariables y factores disponibles en el conjunto de datos con el objetivo de evaluar la relevancia de cada covariable o factor para determinar si debería incluirse o no el modelo ajustado.

```
> #Modelo logit inicial
> modelo0 <- glm(formula = disease ~ 1, family = binomial, data = datos); modelo0

Call:  glm(formula = disease ~ 1, family = binomial, data = datos)

Coefficients:
(Intercept)
    -0.1347

Degrees of Freedom: 237 Total (i.e. Null);  237 Residual
Null Deviance:      328.9
Residual Deviance: 328.9        AIC: 330.9
```

```
> modelo.global <- glm(formula = disease ~ cray + downs + age + fray + mlowray + mray + mupray + sex + cnray, family = binc
mial, data = datos); modelo.global

Call:  glm(formula = disease ~ cray + downs + age + fray + mlowray +
    mray + mupray + sex + cnray, family = binomial, data = datos)

Coefficients:
(Intercept)       crayno     downsyes          age       frayno    mlowrayno       mrayno     muprayno         sexM
   -0.51196      1.85016     17.22458     -0.01994      0.45286    -14.56382     15.05769    -16.44874      0.15997
     cnray2       cnray3       cnray4
   -1.47949      0.15151           NA

Degrees of Freedom: 237 Total (i.e. Null);  227 Residual
Null Deviance:      328.9
Residual Deviance: 294.5        AIC: 316.5
```

```
> anova(modelo0, modelo.global, test = "Chisq")
Analysis of Deviance Table

Model 1: disease ~ 1
Model 2: disease ~ cray + downs + age + fray + mlowray + mray + mupray +
    sex + cnray
  Resid. Df Resid. Dev Df Deviance
1       237     328.86
2       227     294.53 10   34.335
  Pr(>Chi)
1
2 0.000162 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
  0.05 ‘.’ 0.1 ‘ ’ 1
```

Si nos fijamos en el anova conjunto del modelo inicial ("modelo0") y el modelo completo ("modelo.global"), se puede observar que estadístico G del logaritmo de razón de verosimilitudes para el modelo completo es significativo (p-valor = 0.000162), lo que significa que el ajuste del modelo es aceptable. Sin embargo, es más que probable que en el modelo haya factores que no tengan relevancia en el modelo por lo que este modelo puede ser mejorado si le realizamos un mejor ajuste. Tal como observamos en el "summary" del modelo completo, muchos de los factores tenidos en cuenta en este modelo no son significativos, siendo los de más relevancia Crayno (p-valor=0.00667), que representa si se los niños y niñas no hayan sido expuestos a rayos X, y CnRay 1 o 2 (p-valor=0.04136), que representa el número de radiografías realizadas (en este caso es un factor y se refiere concretamente al nivel 1 o 2 radiografías). Aún así, estos datos los hemos obtenido mediante un test automatizado y puede ser que no haya tenido en cuenta la relevancia de variables que a priori no son significativas pero que aportarían información en un modelo ajustado.

```
> summary(modelo.global)

Call:
glm(formula = disease ~ cray + downs + age + fray + mlowray +
    mray + mupray + sex + cnray, family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9778  -1.0153  -0.8647   1.2197   2.0976

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.51196    0.27945  -1.832  0.06694 .
crayno          1.85016    0.68203   2.713  0.00667 **
downsyes       17.22458  869.42131   0.020  0.98419
age            -0.01994    0.02921  -0.683  0.49488
frayno          0.45286    0.30710   1.475  0.14031
mlowrayno     -14.56382 2399.54499  -0.006  0.99516
mrayno         15.05769 2399.54509   0.006  0.99499
muprayno      -16.44874 2399.54483  -0.007  0.99453
sexM            0.15997    0.28134   0.569  0.56962
cnray1 o 2     -1.47949    0.72526  -2.040  0.04136 *
cnray3 o 4      0.15151    1.02339   0.148  0.88231
cnray5 o más         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.86  on 237  degrees of freedom
Residual deviance: 294.53  on 227  degrees of freedom
AIC: 316.53

Number of Fisher Scoring iterations: 15
```

Atendiendo a los coeficientes para el intervalo de confianza al nivel 95% para sus odds-ratio, las variables señaladas como significativas muestran lo siguiente: es 6.36 veces menos probable que el individuo contraiga leucemia si no es expuesto a rayos X. Por otro lado, si el paciente ha sido sometido a 1 o 2 radiografías es 0,22 veces menos probable que el individuo contraiga leucemia.

```
> exp(modelo.global$coefficients)
 (Intercept)       crayno     downsyes          age       frayno    mlowrayno       mrayno
5.993173e-01 6.360805e+00 3.023706e+07 9.802583e-01 1.572803e+00 4.731656e-07 3.463155e+06
    muprayno         sexM   cnray1 o 2   cnray3 o 4 cnray5 o más
7.184619e-08 1.173479e+00 2.277532e-01 1.163584e+00           NA
```

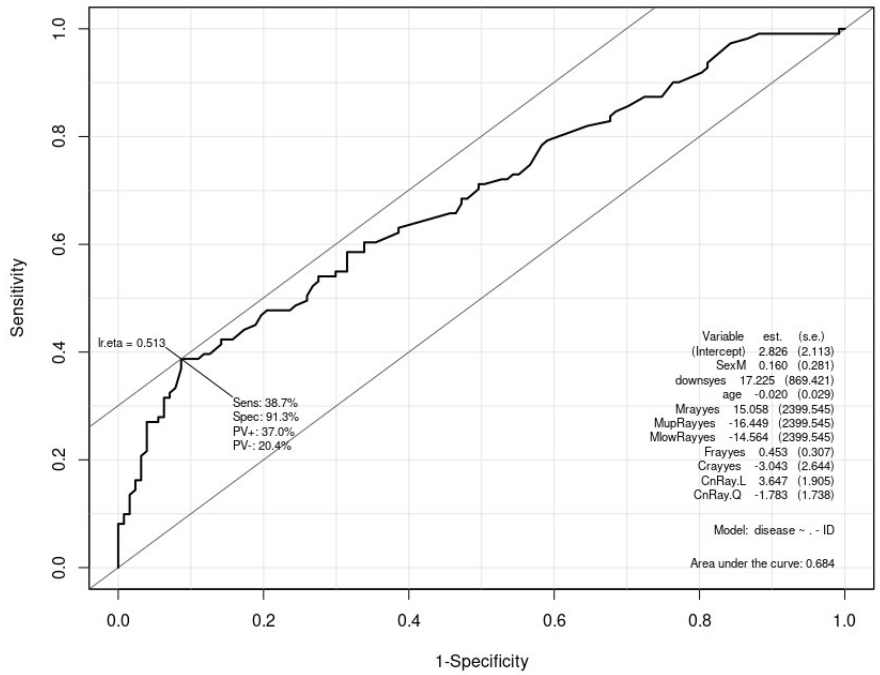
Con respecto a la bondad del ajuste del modelo completo, el test elegido para este análisis ha sido el de Hosmer-Lemeshow que nos ofrece un p-valor de 0.8712, lo que nos indica que tendríamos un 87.12% de probabilidad de equivocarnos si rechazásemos este modelo, lo que implica que tiene un buen ajuste.

```
> hoslem.test(datos$disease, fitted(modelo.global))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  datos$disease, fitted(modelo.global)
X-squared = 3.8403, df = 8, p-value = 0.8712
```

A continuación generaremos una matriz de confusión con un punte de corte de 0.5 para el modelo completo con el objetivo de establecer un caso base con el que comparar una situación en la que a partir de la cual se predice la pertenencia al grupo de interés con la situación en la que obtenemos el mejor punto de corte que se puede obtener para este modelo mediante una curva ROC para comparar las matrices de confusión y evaluar el poder clasificatorio de este modelo entre individuos que pueden contraer leucemia e individuos que no contraerían leucemia teniendo en cuenta las covariables y factores incluidos en el modelo.

```
> clasificacion(datos$disease, modelo.global, corte = 0.5)
      Clasifica correcto Clasifica incorrecto Clasifica total
1                    43                   14              57
0                   113                   68             181
Total               156                   82             238
```

En esta matriz de confusión, observamos que el poder discriminatorio no es muy bueno ya que de 284 observaciones clasifica incorrectamente 82 individuos: 14 falsos positivos y 68 falsos negativos. Esto significa que tiene una sensibilidad muy baja y es peligroso porque provocaría que un alto número de niños con leucemia fuera diagnosticado erróneamente con la implicaciones tan funestas que esto conllevaría si aplicásemos este modelo bajo estas condiciones.
El siguiente paso es establecer el punto de corte óptimo para el modelo completo mediante el método del curva ROC:



Observando la gráfica de la curva ROC, podemos observar que nos da como punto de corte óptimo 0.513, un área bajo la curva (AUC) de 0.684, una especificidad del 91.3% y una sensibilidad del 38.7%. Según el criterio general de Hosmer la AUC obtenida no tiene un poder de discriminación aceptable (<0.70 AUC). Por otro lado, destacar la alta especifidad que tiene el modelo con este punto de corte pero la baja sensibilidad que tiene por otro lado. Ello nos inclina a pensar que este modelo no tiene mucha capacidad para predecir el suceso de interés y es una clasificación demasiado aleatoria. Ahora, observamos la matriz de confusión generada con el punto de corte obtenido con la curva ROC (0.513) y vemos que el modelo mejora ligeramente su capacidad de discriminación pero aún así sigue sin ser suficiente:

```
> clasificacion(datos$disease, modelo.global, corte = 0.5)
      Clasifica correcto Clasifica incorrecto Clasifica total
1                    43                   14             57
0                   113                   68            181
Total               156                   82            238
```

No obstante, este modelo incluye factores no relevantes, por lo que no es el más óptimo. Por consiguiente, procederemos a elaborar un modelo con mejor ajuste para ver si es posible obtener un modelo que tenga un poder de discriminación aceptable y para ello debe repetirse el análisis del modelo logístico suprimiendo los términos no significativos y obtener mejor ajuste. Para ello, aplicaremos la función "step" que sigue el criterio de información de Akaike (AIC) que corresponde al valor de logaritmo de verosimilitud para obtener un modelo de mejor calidad y le aplicaremos la opción "both" que consiste en la realización de iteraciones en las que incluye un término al modelo si su presencia disminuye el AIC y se excluye del modelo un término si su ausencia disminuye el AIC. Por consiguiente, a menor AIC, mejor ajuste del modelo estimado. A continuación, se muestra la ultima iteración de la función step que muestra el modelo logístimo estimado con menor AIC posible:

```
Step:  AIC=309.96
disease ~ downs + cnray + mupray + fray

          Df Deviance    AIC
<none>         295.96 309.96
- fray     1   298.12 310.12
- mupray   1   298.96 310.96
+ mray     1   295.53 311.53
+ age      1   295.54 311.54
+ mlowray  1   295.63 311.63
+ sex      1   295.72 311.72
- cnray    3   311.54 319.54
- downs    1   310.73 322.73
```

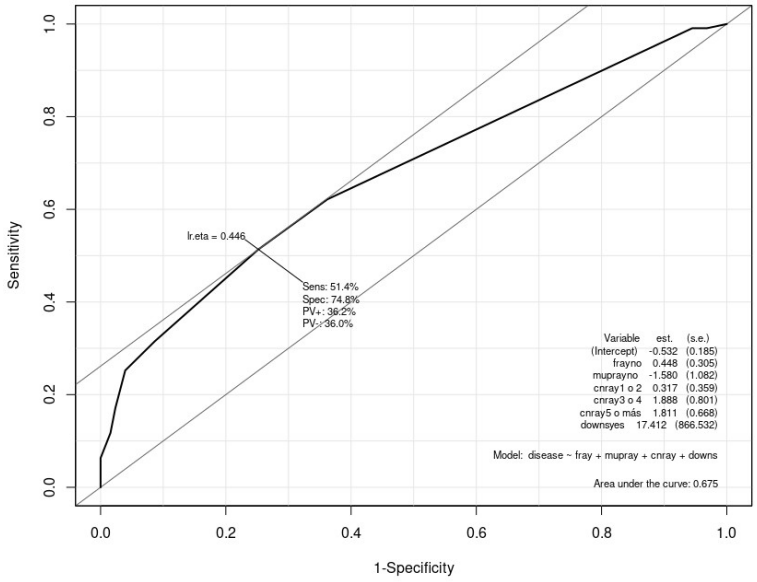
Según el criterio AIC, el modelo mejor ajustado quedaría definido de la siguiente manera:

$$disease \sim downs + CnRay + MupRay + Fray$$

Es decir, quedarían incluidos en el modelo: downs (individuo con síndrome de Downs), CnRay (número de radiografía las que ha sido sometidas el sujeto), MupRay (rayos X superior durante el embarazo) y Fray (el padre fue sometido a rayos X). Si nos fijamos el AIC de este modelo es menor (309.96) al del modelo completo (316.5).

En cuanto a la bondad de ajuste para este nuevo modelo, el test de Hosmer y Lemeshow nos da un p-valor de 1, es decir, tendríamos un 100% de probailidad de equivocarnos si rechazásemos este modelo, por lo que se deduce que el ajuste es muy bueno y, además, superior al ajuste del modelo completo analizado anteiormente.

```
> hoslem.test(datos$disease, fitted(modelo.stepAIC))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  datos$disease, fitted(modelo.stepAIC)
X-squared = 0.36457, df = 8, p-value = 1
```



Aplicando el método de la curva ROC al modelo seleccionado mediante AIC se obtiene un punto de corte de 0.446, un AUC de 0.675, una sensibilidad de 51.4% y una especificidadd del 74.8%. En cuanto al área bajo la curva podemos decir que ha mejorado con respecto al modelo completo pero a pesar de ello, sigue sin cumplir el criterio general de Hosmer y Lemeshow para determinar una capacidad de discriminación aceptable aunque se acerca al mínimo, que es 0.7. Por otro lado, continuando con la comparación con el modelo completo, se puede apreciar un aumento considerable en la sensibilidad y un descenso en la especificidad, aspecto que también podemos observar el matriz de confusión para este modelo seleccionado:

```
> clasificacion(datos$disease, modelo.stepAIC, corte = 0.416)
      Clasifica correcto Clasifica incorrecto Clasifica total
1                    69                   46             115
0                    81                   42             123
Total               150                   88             238
```

A pesar de que este modelo tiene buen ajuste y la mejora en calidad del modelo para clasificar nuestra variable de interés, disease (leucemia infantil), sigue sin tener la capacidad de predicción suficiente como para poder ser utilizado como modelo de referencia. Para lograr un modelo más preciso es conveniente recolectar más datos para poder obtener más información sobre las variables ya estudiadas y también añadir otras posibles variables de estudio con el objetivo de encontrar factores que contribuyan de manera significativa a aumentar el riesgo a contraer leucemia en la población infantil. Una propuesta podría ser estudiar la influencia de factores ambientales como la exposición a agentes cancerígenos como pueden ser la polución, la ingesta de ultraprocesados o estudiar el perfil genético de los individuos para estudiar si la expresión génica influye sensiblemente en la posibilidad de contraer leucemia y, en caso afirmativo, identificar biomarcadores para poder realizar pronósticos eficaces.