Parcial 3: Modelos Estadísticos I.

Rojas Gutiérrez Rodolfo Emmanuel

29 de mayo de 2021

Ejercicio 1:

Observación 1. *A lo largo de este ejercicio, devres denotará a la deviance residual y df a los grados de libertad de la misma, d*1 *representa a la dosis de radiación, d*2 *a la dosis de radiación al cuadrado y d*3 *a la respectiva dosis de radiación al cubo.* △

Solución. 1 Como se están considerando proporciones se hará uso de un GLM binomial. Para el modelo 1), se considero el GLM binimial con componente sistemático:[1]

$$g(\pi) = \beta_0 + \beta_1 d1, \tag{1}$$

cabe destacar que se uso la liga logit, con lo que se obtuvieron las estimaciones en R presentadas en la figura 1. Como puede verse, bajo los estadísticos de Wald se ha obtenido que todos los coeficientes

```
Call:
glm(formula = prop ~ d1, family = binomial, data = A, weights = total)

Deviance Residuals:
      1       2       3       4       5       6
 1.7700  -0.8191  -1.5681  -1.0209  -0.8518   1.4651

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.91157    0.26402 -14.815  < 2e-16 ***
d1           0.57307    0.08521   6.726 1.75e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.351  on 5  degrees of freedom
Residual deviance: 10.177  on 4  degrees of freedom
AIC: 35.842

Number of Fisher Scoring iterations: 4
```

Figura 1: Estimaciones modelo (1).

de este modelo, resultan significativamente distintos de cero, bajo un nivel de significancia del 5 %. Además, se tiene que este modelo cuenta con

$$AIC = 35.842,\ devres = 10.177 \text{ con } df = 4. \tag{2}$$

Dado que la deviance residual no es mucho mayor a sus grados de libertad, no se considerará necesario ajustar un modelo que consideré sobredispersión en los datos. Por último, se deja en la parte

[1] Donde π denota a la probabilidad de morir por leucemia, en lugar de por algún otro cáncer, dado que se estuvo expuesto a cierta dosis de radiación.

izquierda de la figura 5, una gráfica de los valores ajustados por este modelo, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos, podemos ver que el ajuste no es óptimo, parece que nos esta haciendo falta considerar cierta curvatura en los datos, lo cual se verá a continuación. Pese a ello, como ya se comento, los coeficientes son todos significativos, por lo que, este modelo puede ser un buen punto de partida para buscar una mejor alternativa, el AIC se usará más adelante para la selección del modelo.

2 Para el modelo 2), se considero el GLM binimial con componente sistemático:

$$g(\pi) = \beta_0 + \beta_1 d1 + \beta_2 d2, \tag{3}$$

cabe destacar que se uso la liga logit, con lo que se obtuvieron las estimaciones en R presentadas en la figura 2. Observe que, en este modelo el coeficiente que pre-multiplica a la variable explicativa $d1$,

```
Call:
glm(formula = prop ~ d1 + d2, family = binomial, data = A, weights = total)

Deviance Residuals:
       1         2         3         4         5         6
 0.04643  -0.18324   0.07499   0.22535  -0.19401   0.02917

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.38307    0.27259 -12.411  < 2e-16 ***
d1          -0.41832    0.32872  -1.273  0.20318
d2           0.19440    0.06359   3.057  0.00223 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.35089  on 5  degrees of freedom
Residual deviance:  0.13063  on 3  degrees of freedom
AIC: 27.796

Number of Fisher Scoring iterations: 4
```

Figura 2: Estimaciones modelo (3).

resulta ser no significativo bajo un nivel de significancia del 5 %, sin embargo, los demás coeficientes si resultan ser significativos bajo el nivel mencionado. Adicionalmente, se tiene que este modelo cuenta con:

$$AIC = 27.796,\ devres = 0.13063 \text{ con } df = 3. \tag{4}$$

Aquí, la deviance residual es muy pequeña mucho menor que sus grados de libertad, lo que nos habla de un buen ajuste del modelo.[2]. Por otro lado, en la parte de en medio de la figura 5, se puede observar una gráfica de los valores ajustados por este modelo, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos, vemos que la curvatura que hacía falta considerar en el modelo (1) parece haber desaparecido, sin embargo, pareciera haber cierto grado de sobre-ajuste del modelo.

3 Para el modelo 3), se considero el GLM binimial con componente sistemático:

$$g(\pi) = \beta_0 + \beta_1 d1 + \beta_2 d2 + \beta_3 d3, \tag{5}$$

cabe destacar que se uso la liga logit, con lo que se obtuvieron las estimaciones en R presentadas en la figura 3. Aquí, parece que ya estamos utilizando demasiadas covariables en el modelo, puesto

[2]Aunque no puede descartarse un sobreajuste, debido a que al menos un coeficiente no es significativo

```
glm(formula = prop ~ d1 + d2 + d3, family = binomial, data = A,
    weights = total)

Deviance Residuals:
      1          2          3          4          5          6
 0.03488  -0.15687   0.08481   0.21012  -0.22282   0.04073

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.379800   0.281190 -12.020   <2e-16 ***
d1          -0.449365   0.746230  -0.602    0.547
d2           0.213066   0.407586   0.523    0.601
d3          -0.002545   0.054865  -0.046    0.963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.35089  on 5  degrees of freedom
Residual deviance:  0.12848  on 2  degrees of freedom
AIC: 29.793

Number of Fisher Scoring iterations: 4
```

Figura 3: Estimaciones modelo (5).

que ninguna de ellas resulta ser significativamente distinta de 0, bajo un nivel de significancia del 5 %, pese a que en los modelos anteriores, que consideraban una cantidad menor de covariables si que lo eran. Por otro lado, este modelo cuenta con:

$$AIC = 29.793,\ devres = 0.12848 \text{ con } df = 2. \tag{6}$$

Vemos que nuevamente la deviance residual es mucho menor a sus grados de libertad, inclusive aún, más que en el modelo considerado en el inciso 2. Por último, se deja en la parte derecha de la figura 5 una gráfica de los valores ajustados por este modelo, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos. En está gráfica no queda duda del nivel de sobreajuste por parte de este modelo.

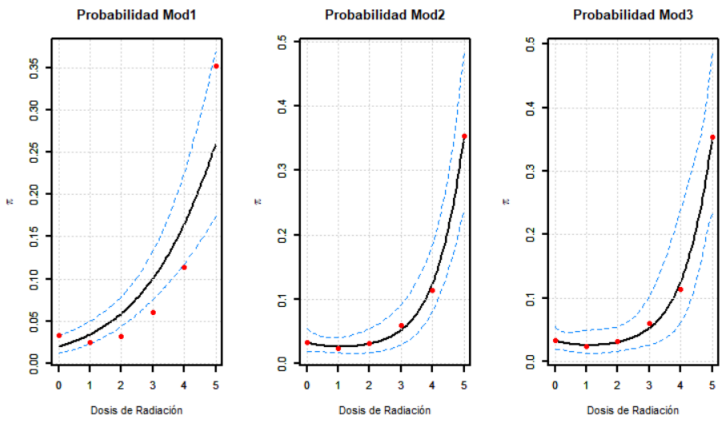

Figura 4: Valores ajustados por los tres modelos, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos.

4 De los tres modelos comparados en los incisos anteriores, claramente el mejor resulta ser

el modelo (3) considerado en el inciso 2), ya que el nivel de ajuste en las gráficas parece ser adecuado y cuenta con el menor AIC de los tres modelos considerados, sin embargo, cuenta con un ligero problema, uno de sus coeficientes resulta no ser significativamente distinto de cero, bajo un nivel de significancia del 5 %, en este inciso buscaremos optimizar dicho modelo, de ser posible. Primeramente, para corroborar que las variables $d1$ y $d2$ son las que más información aportan a este modelo, se deja el siguiente análisis anova:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: prop

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     5     54.351
d1    1   44.174         4     10.177 3.005e-11 ***
d2    1   10.046         3      0.131  0.001527 **
d3    1    0.002         2      0.128  0.963014
---
```

Figura 5: Análisis anova.

En este análisis observamos que el agregar el predictor $d3$ al modelo realmente no resulta relevante, dado que el p-valor de la prueba en el que no se considera este coeficiente es bastante alto, prácticamente cercano a 1. Por otro lado, los otros dos coeficientes resultan bastante significativos, bajo un nivel de significancia del 5 %. Puesto que ya ajustamos, el modelo que considera solo a $d1$ y el que considera a $d2$ y a $d1$, donde como ya se mencionó fue claramente superior el modelo que considerá a ambas covariables, veremos que tal resulta el ajuste del modelo que únicamente considera a $d2$, es decir el glm binomial con componente sistemático:

$$g(\pi) = \beta_0 + \beta_2 d2, \tag{7}$$

cabe destacar que se uso la liga logit, con lo que se obtuvieron las estimaciones en R presentadas en la figura 6. Observe que, todos los coeficientes de este modelo resultan ser significativamente

```
Call:
glm(formula = prop ~ d2, family = binomial, data = A, weights = total)

Deviance Residuals:
      1       2       3       4       5       6
 0.8645  -0.4138  -0.5737  -0.3053  -0.5649   0.3507

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.62345    0.21881 -16.560  < 2e-16 ***
d2           0.11656    0.01512   7.711 1.25e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.3509  on 5  degrees of freedom
Residual deviance:  1.7831  on 4  degrees of freedom
AIC: 27.448

Number of Fisher Scoring iterations: 4
```

Figura 6: Estimaciones modelo (7).

distintos de cero, bajo un nivel de significancia del 5 %. Además, este modelo cuenta con:

$$AIC = 27.448,\ devres = 1.7831 \text{ con } df = 4. \tag{8}$$

Puede observar que la deviance residual es menor que sus grados de libertad, sin llegar a ser excesivamente pequeña como en los casos de los modelos en (5) y (3), lo que es un buen primer indicio de un ajuste más que razonable del modelo.[3] Por otro lado, en la figura 7, puede observar una gráfica de los valores ajustados por este modelo, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos. En la misma, se ve un ajuste más que razonable, sin parecer que el modelo trata de pasar por todas las observaciones en el mismo, es decir no parece existir sobre ajuste, como en los modelos en (3) y (5). Más aún, cuenta con un AIC de 27.488, lo que es menor que el hasta ahora mejor modelo, modelo (3) con un AIC de acuerdo con (4) de 27.796. Así, se elige este como el mejor modelo posible, con las covariables dadas y sus derivados. ■

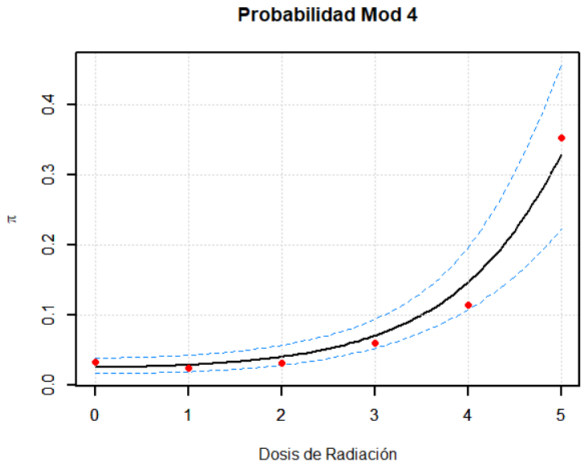

Figura 7: Valores ajustados por el modelo (7), más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos.

Ejercicio 2:

Solución. **a)** Para este inciso se considerará el ajuste del GLM poisson, con componente sistemático:

$$g(\mu) = \beta_0 + \beta_1 X. \tag{9}$$

donde $\mu = E[Y|X]$ y g es la liga logarítmica. Haciendo uso de R se obtuvieron las estimaciones de los coeficientes presentadas en la figura 8: Como puede notarse, todos los coeficientes de este modelo resultan significativamente distintos de cero, bajo un nivel de significancia del 5 %. Además, este modelo cuenta con:

$$AIC = 41.052,\ devres = 2.9387 \text{ con } df = 7. \tag{10}$$

Se nota que la deviance residual resulta bastante más chica que sus grados de libertad, lo que nos habla de la no existencia de sobredispersión,[4] además, dado que la deviance residual del modelo es

[3] Esto además nos vuelve a indicar la no existencia de sobredispersion

[4] Por lo que no habrá que considerar un modelo que considere este aspecto.

```
Call:
glm(formula = Y ~ X, family = "poisson", data = B)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8472  -0.2601  -0.2137   0.5214   0.8788

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8893     0.1421  13.294  < 2e-16 ***
X             0.6698     0.1787   3.748 0.000178 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18.4206  on 8  degrees of freedom
Residual deviance:  2.9387  on 7  degrees of freedom
AIC: 41.052

Number of Fisher Scoring iterations: 4
```

Figura 8: Estimaciones modelo (9).

muy pequeña, se puede hablar de un buen ajuste del modelo. Esto último, puede corroborarse en la gráfica a la izquierda de la figura 10, donde se aprecia una gráfica de los valores ajustados por este modelo, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos, además, no pareciera existir sobreajuste al menos visualmente. El criterio de Akaike será utilizado mas adelante, para comparar este modelo con el que será ajustado en el siguiente inciso.

b) Para este inciso se considerará el ajuste del GLM poisson, con componente sistemático:

$$g(\mu) = \beta_0 + \beta_1 X. \tag{11}$$

donde $\mu = E[Y|X]$ y g es la liga identidad. Para el que, nuevamente haciendo uso de R, se obtuvieron las estimaciones presentadas en la figura 9. Note que, al igual que en el modelo anterior todos los

```
Call:
glm(formula = Y ~ X, family = poisson("identity"), data = B)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7019  -0.3377  -0.1105   0.2958   0.7184

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.4516     0.8841   8.428  < 2e-16 ***
X             4.9353     1.0892   4.531 5.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18.4206  on 8  degrees of freedom
Residual deviance:  1.8947  on 7  degrees of freedom
AIC: 40.008

Number of Fisher Scoring iterations: 3
```

Figura 9: Estimaciones modelo (11).

coeficientes resultan significativamente distintos de cero, bajo un nivel de significancia del 5 %, lo que es una señal inequivoca de un buen ajuste del modelo. Adicionalmente, se puede ver que este modelo cuenta con:

$$AIC = 40.008,\ devres = 1.8947 \text{ con } df = 7. \tag{12}$$

Nuevamente, se desataca que al igual que en el modelo anterior, la deviance residual resulta mucho menor a sus grados de libertad, además de que, no resulta excesivamente lejana de cero, por lo que,

se tiene otro buen indicativo del ajuste del modelo. Finalmente, y como ya es costumbre se deja en la parte derecha de la figura 10, una gráfica de los valores ajustados por este modelo, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos. Donde, al igual que en el caso anterior se ve en general un ajuste bastante razonable a los datos observados. Nuevamente, el AIC nos será de utilidad en los siguientes incisos.

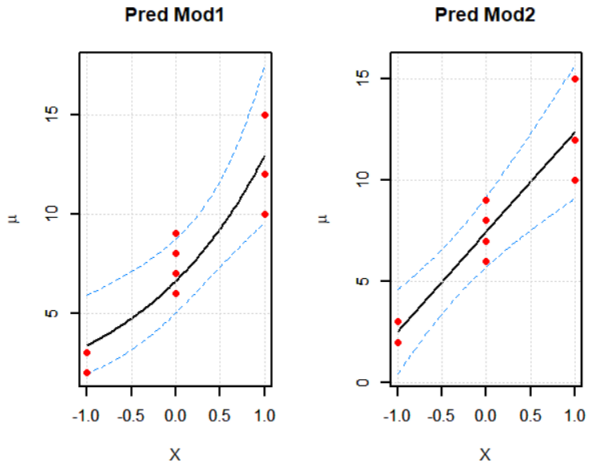

Figura 10: Valores ajustados por los dos modelos, más intervalos de confianza al 95 % en lineas azules y observaciones en puntos rojos.

c) y d) Bajo los dos modelos analizados en los incisos anteriores, se observo un ajuste bastante similar a través de los cálculos realizados, ambos cuentan con coeficientes significativamente distindos de cero, una deviance residual pequeña y mucho menor a sus grados de liberta e incluso poseen AIC muy parecidos:

$$\begin{aligned} AIC \text{ del modelo en (9)} &: 41.052, \\ AIC \text{ del modelo en (11)} &: 40.008. \end{aligned} \tag{13}$$

Además, observando las gráficas en (10), nos percatamos de que visualmente el ajuste es también similar, salvo la evidente curvatura añadida por la liga logarítmica, tenemos un comportamiento similar en los intervalos de confianza, (son mas anchos hacía los extremos y más cerrads hacia el centro), y el mismo patrón creciente en las predicciones, por lo que, se decidió hacer uso de los residuales de cuantiles,[5] los cuales deberían de distribuirse aproximadamente como normales estándar de estar siendo utilizado el glm correcto, con este idea en mente se construyeron las siguientes gráficas QQ, presentadas en la figura con los residuales de cuantiles de ambos modelos: Como puede observarse, ambas gráficas presentan un ajuste razonable de los residuales hacia el centro de la distribución normal estándar, y ambas gráficas presentan problemas hacia las colas, por lo que, incluso bajo este criterio ambos modelos resultan muy similares. Por lo que, nos basaremos en

[5]Ver Dunn y Smyth (2018) capítulo 8 para una definición de los mismos.

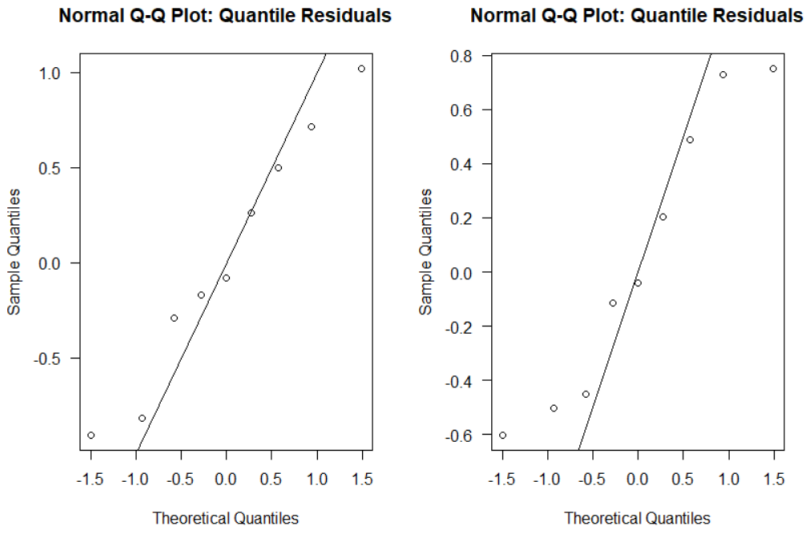

Figura 11: Gráficas QQ residuales de cuantiles, a la izquierda la del modelo en (9) y a la derecha la del modelo en (11).

el criterio menos subjetivo de todos estos para seleccionar al mejor de los dos, este es el AIC, criterio bajo el cual puede comprobarse en (13), es una mejor elección el modelo con liga identidad, ya que este posee el menor AIC. Empíricamente esto puede verse en los gráficos de dispersión, presentados en la figura 12. En los cuales, se presenta un gráfico de dispersión Y vs X a la izquierda, $log(Y)$ vs $log(X)$ a la derecha. Aqui, se puede apreciar cierta relación lineal entre X e Y, mientras que entre X y $ln(Y)$ pareciera existir cierta curvatura en la relación que entre ambos variables, por lo que, si se quiere continuar usando dicha liga, podría considerarse meter algún termino polinomial de X. ■

1. Anexo

Como comentario final para el ejercicio 1, el modelo que considera únicamente a $d3$ tiene por muy poco un menor[6] AIC que el modelo elegido, sin embargo, parece tener cierto sobre ajuste, ver figura 13, que era un propiedad no deseada desde el principio.

[6] $AIC = 26.19$

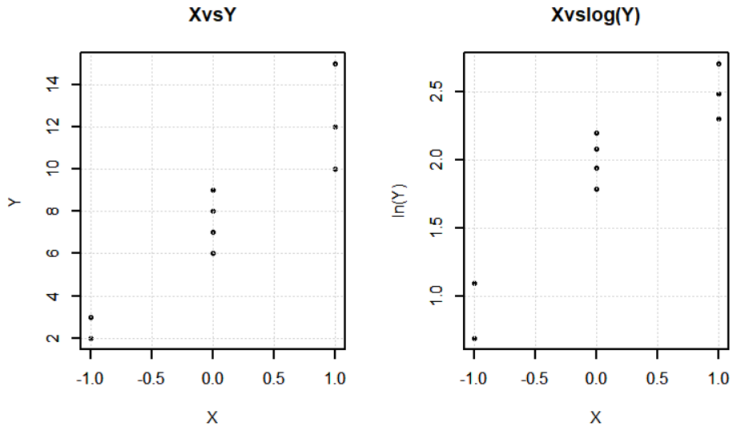

Figura 12: Gráfico de dispersión Y vs X a la izquierda, $ln(Y)$ vs X a la derecha.

Referencias

Dunn, P. K. & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R* (1st ed. 2018). Springer New York : Imprint: Springer. https://doi.org/10.1007/978-1-4419-0118-7

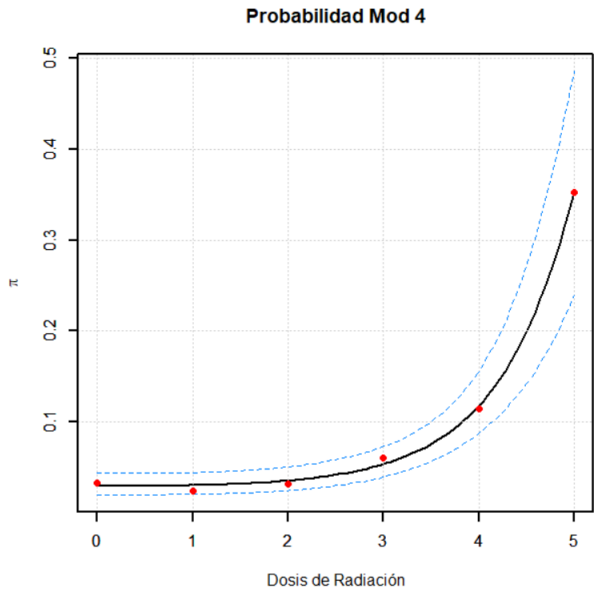

Figura 13: Modelo que solo considera a $d3$