CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS (CIMAT).
UNIDAD MONTERREY

Tarea-1-MLG-Hairo Ulises-Miranda Belmonte

Hairo Ulises Miranda Belmonte

03 septiembre, 2019

0.1 Ejercicio 1

Los siguientes datos tomados de Little (1978) corresponden a 1607 mujeres casadas y fertiles entrevistadas por la Encuesta de Fertilidad Fiji de 1975. Los datos estan clasificados por edad, nivel de educación, deseo de tener hijos y el uso de anticonceptivos. En este ejemplo se considera la anticoncepción como variable dependiente y a las demás como predictoras. Todas las predictoras son variables categoricas. El objetivo es describir cómo el uso de métodos anticonceptivos varía según la edad, el nivel de educación y el deseo de tener más hijos.

a) Ajuste un modelo lineal a los datos

Base de datos:

Edad	Educacion	Hijos	Si	No
<25	Baja	Si	53	6
<25	Baja	No	10	4
<25	Alta	Si	212	52
<25	Alta	No	50	10
25-29	Baja	Si	60	14
25-29	Baja	No	19	10
25-29	Alta	Si	155	54
25-29	Alta	No	65	27
30-30	Baja	Si	112	33
30-30	Baja	No	77	80
30-30	Alta	Si	118	46
30-30	Alta	No	35	6
40-40	Baja	Si	68	78

Edad	Educacion	Hijos	Si	No
40-40	Baja	No	46	48
40-40	Alta	Si	8	8
40-40	Alta	No	12	31

Modelo de regresión lineal:

$$Anticonceptivos_{si} = \beta_0 + Edad_{25-29} + Edad_{30-39} + Edad_{40-49} + Edad_{si} + Educacion_{Baja} + \epsilon$$

β_0: intercepto, captura a las mujeres de menos de 25 años, con educación alta, y sin deseo de tener hijos.

El resto de los parámetros son variables dummies con $k-1$ factores, donde k varia respecto a la variable.

El resultado del ajuste de la regresión lineal es el siguiente:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.87	29.98	2.164	0.05573
Edad25-29	-6.5	34.62	-0.1878	0.8548
Edad30-30	4.25	34.62	0.1228	0.9047
Edad40-40	-47.75	34.62	-1.379	0.1978
EducacionBaja	-26.25	24.48	-1.072	0.3087
HijosSi	59	24.48	2.41	0.03666

Tabla 0.3: Fitting linear model: Si ~ Edad + Educacion + Hijos

Observations	Residual Std. Error	R^2	Adjusted R^2
16	48.96	0.4955	0.2433

Se pretende observar la relación entre las mujeres que hacen uso de métodos anticonceptivos, respecto a su rango de edad, su nivel de educació y el deseo de tener hijos.

Significancia:

Evaluando la significancia del los parámetros se puede inferir a las covariables que afectan el número de mujeres que sí utilizan anticonseptivos. Al .05 de significancia las variables significativas son; las

mujeres que sí desean tener hijos (de forma positiva) y ligeramente el parámetro de intercepto (de forma positiva), que captura a las mujeres que no desean tener hijos, con un nivel de educación alta y edad menor a 25 años.

Interpretación:

Esto quiere decir, que el número de mujeres que si utilizan anticonseptivo son aquellas que aún sienten el deseo de tener algún hijo, no obstante, dado que el intercepto es diferente de cero, podemos inferir que esas mujeres pueden ser aquellas que tienen un nivel de estudio elevado, con una edad menor a los 25.

Ajuste del modelo:

El resultado se considera espurio por dos razones; i) la variable respuesta parece distribuirce como una bernulli; ii) El $R^2 ajustada$ - la cual penaliza el número de covariables que se agregan al modelo-, se aleja del R^2, indicando que alguna variable está de más, y sobre todo un mal ajuste en las covariables respecto a la variable respuesta.

Por otro lado, el valor del F-statistic es de 1.964 con $p - value : 0.1701$, lo cual hace que el modelo en su conjunto no sea significativo, i.e., las covariables no explican al número de mujeres que sí utilizan anticonseptivos.

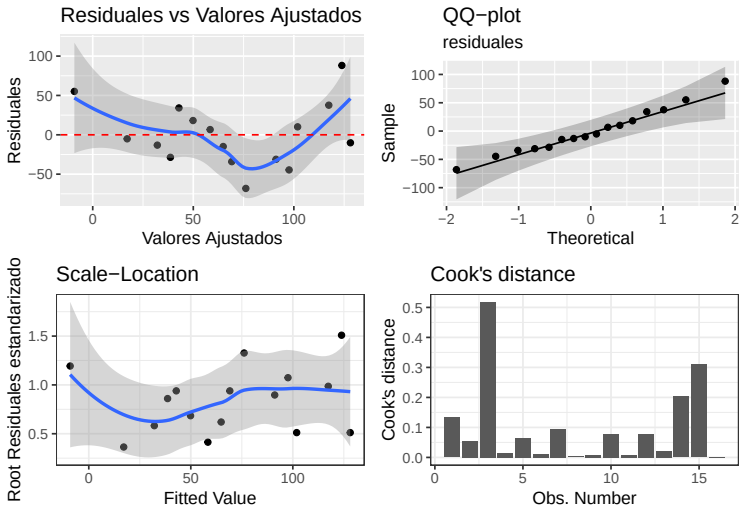
Intervalos de confianza de los parámetros estimados:

Se incluyen los intervalos de confianza de los parámetros estimados

	2.5 %	97.5 %
(Intercept)	-1.924	131.7
Edad25-29	-83.63	70.63
Edad30-30	-72.88	81.38
Edad40-40	-124.9	29.38
EducacionBaja	-80.79	28.29
HijosSi	4.459	113.5

Supuestos del modelo:

Se evaluan los supuestos del modelo, residuales con distribución normal, las covariables y la variable respuesta estimada no correlacionada con los residuales del modelo, y homocedasticidad en los términos de error.



- Gráfico superior izquierdo: se tiene a los residuales respecto a los valores ajustados, se observa que los residuales siguen el comportamiento del ajuste de la linea, por ende, se encuentran correlacionados.
- Gráfico superior derecho QQplot: se puede ver que la distribución de los residuales parecen en el centro normal, sin embargo, no en las colas.
- Gráfico Scale-Location: se observa como los residuales conforme incrementa el número de valores ajustados tienden a dispersarse, indicando heterocedasticidad, i.e., varianza no constante.
- Gráfico cook's distance: nos muestra que existe un valor que se encuentra influenciando el resultado del modelo.

En resumen no se cumplen los supuestos de regresión lineal

Análisis de varianza:

Realizamos el análisis de varianza para observar si las covariables tienen relación con la respuesta (i.e., edad, educación, e hijos se relacionan con el uso de anticonceptivos). La hipotesis nula está dada en términos de que las covariables son independientes a la variable respuesta.

Tabla 0.5: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Edad	3	6861	2287	0.9543	0.4512
Educacion	1	2756	2756	1.15	0.3087
Hijos	1	13924	13924	5.81	0.03666
Residuals	10	23967	2397	NA	NA

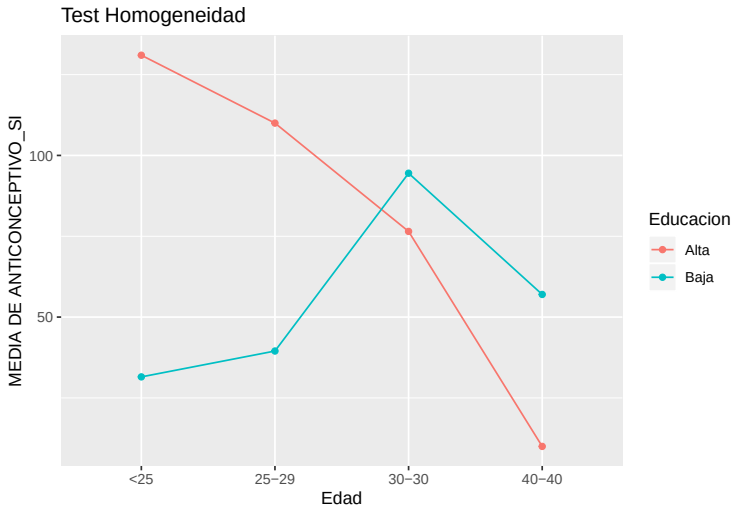
A un nivel del 0.05 de significancia, se tiene que solo el efecto en conjunto de la variable, "deseo de

tener hijos", afecta al número de mujeres que utilizan anticonceptivos.

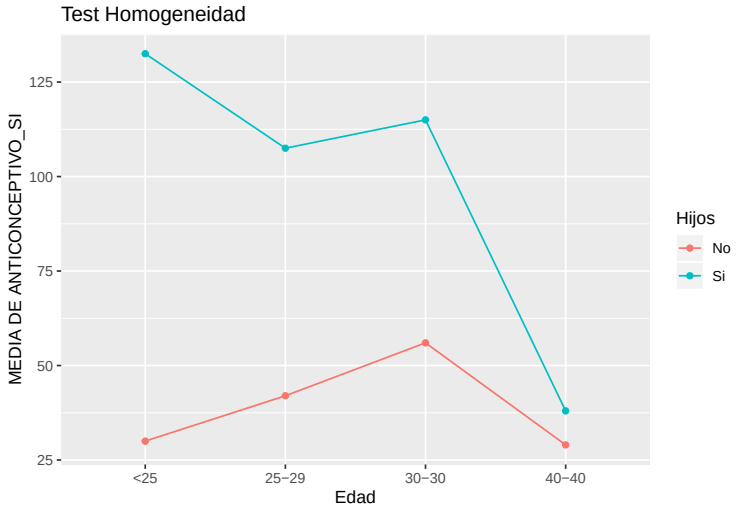
Supuesto de Homogeneidad en las pendientes:

El análisis ANCOVA exige el supuesto de homogenidad en los factores de las variables, para poder ver su efecto total respecto a la variable respuesta.

En el siguiente gráfico de observa la interacción de la Edad respecto al número de mujeres que utilizan anticonceptivos teniendo en cuenta los factores de la variable educación. Si se observa el gráfico la pendiente de los factores son:



La edad respecto al número de mujeres que utilizan anticonceptivos evaluando la pendiente de los factores de la variable "deseos de tener hijos":



Se puede observar que las pendientes son distintas y no satisfacen el supuesto de pendientes homogeneas.

Conclusión:

Se sugiere realizar regresión por cada uno de los factores, si el interés es no capturar el efecto total de las covariables.

El modelo de regresión lineal no es adecuado para el componente aleatorio Y, el número de mujeres que utilizan anticonceptivos.

b) Ajuste un modelo de regresión logística a los datos.

Se asume que la variale respuesta $Y_i \sim Bernoulli(p_i)$

Modelo:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.164	0.1616	7.204	5.844e-13
Edad25-29	-0.4111	0.174	-2.363	0.01812
Edad30-30	-0.6889	0.1715	-4.016	5.911e-05
Edad40-40	-1.606	0.1935	-8.299	1.046e-16
EducacionBaja	-0.005538	0.1281	-0.04322	0.9655
HijosSi	0.434	0.1176	3.692	0.0002227

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	165.77 on 15 degrees of freedom
Residual deviance:	38.38 on 10 degrees of freedom

Significancia:

- El intercepto es significativo y se relaciona con el uso de anticonceptivo de manera positiva
- Los distintos intervalos de los log-odss de la edad son significativos y se relacionan de forma negativa a la variable respuesta. Solo el parámetro de la mujeres con 25 a 29 años de edad es significativo al .05, el resto al .01.
- El parámetro de los log-odds del deseo de sí tener hijos es significativo y se relaciona de forma positiva
- El parámetro que representa el log-odds del nivel de educación baja, no es significativo.

La interpretación de que sean significativo los parámetros de las covariables y su relación con el uso de anticonceptivos se menciona en el inciso e), el calculo de los odd-ratio. Aquí solo se tiene que existe una relación significativa entre la edad y el deseo de tener hijos

c) Compare ambos modelos.

Comparando los modelos con el análisis de varianza:

Regresión Lineal

Tabla 0.8: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Edad	3	6861	2287	0.9543	0.4512
Educacion	1	2756	2756	1.15	0.3087
Hijos	1	13924	13924	5.81	0.03666
Residuals	10	23967	2397	NA	NA

Regresión logísitca

Tabla 0.9: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	15	165.8	NA
Edad	3	113.8	12	51.96	1.662e-24
Educacion	1	0.08251	11	51.88	0.7739
Hijos	1	13.51	10	38.38	0.0002378

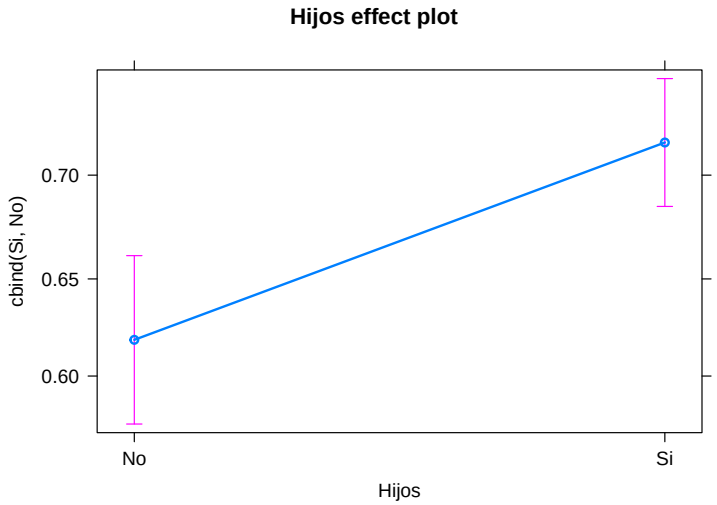
Bajo la regresión logística se tiene significancia al 0.05 de que el efecto de las variables en su conjunto no son independientes al uso de anticonceptivos.

Evaluando el críterio AIC, se observa que el menor valor lo registra la regresión logistica, siendo este el mejor modelo entre los dos.

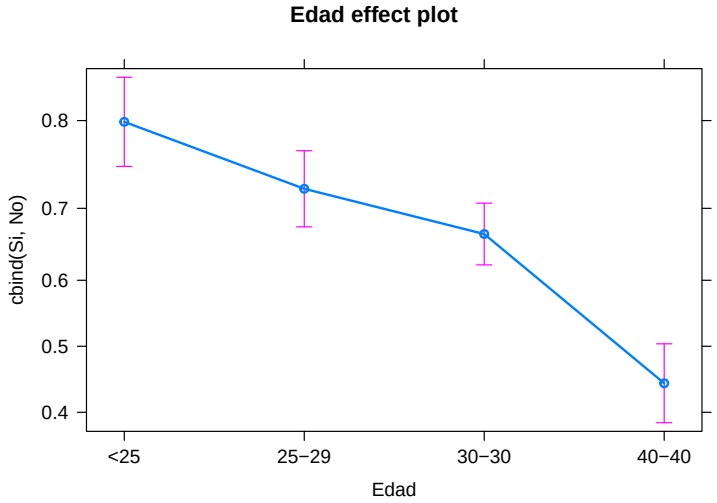
	AIC
Lineal	176.4
Logit	121.9

d) Grafique los modelos e interprete los resultados.

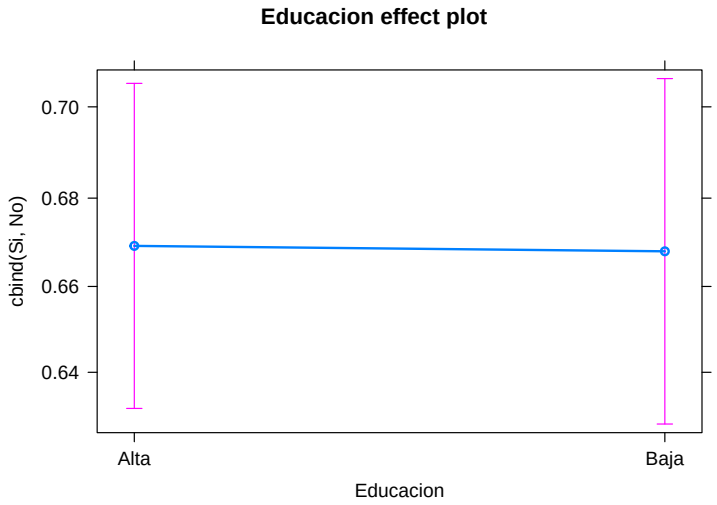
Al tratarse de un modelo con 3 predictores, no se puede obtener una representación en 2D en la que se incluyan ambos predictores a la vez. Sí es posible representar la curva del modelo logístico cuando se mantiene constante uno de los dos predictores. Cabe recalcar que aquí se siguen interpretando en témino del log-odd.



- la probabilidades de usar anticonseptivos es mayor si la mujer tiene deseos de tener un hijo.



- conforme la edad de la mujer va avanzando la probabilidad de utilizar anticonceptivos disminuye.



- El nivel de educación no afecta en el uso de anticonceptivos.

e) Calcule el odds-ratio para el modelo de regresión logística y de su interpre-tación.

$$\hat{O} = \frac{odds_{x+1}}{odds_x} = e^{\beta}$$

Se aplica exponencial a los términos en logaritmo, como resultado tenemos los odds ratios; también se agregan los intervalos de confianza de los parámetros (al 95%) estimados del modelo, los cuales se le aplica exponencial para su interpretación.

	OR	2.5 %	97.5 %
(Intercept)	3.204	2.345	4.422
Edad25-29	0.6629	0.4702	0.9307
Edad30-30	0.5021	0.3575	0.7008
Edad40-40	0.2008	0.1368	0.2922
EducacionBaja	0.9945	0.7744	1.28
HijosSi	1.543	1.225	1.943

Interpretación

- El ser mujer entre los 25 a 29 años de edad reduce el uso de anticonceptivos en 34% aproximadamente.
- El ser mujer entre los 30 a 39 años reduce el uso de anticonceptivos en 50% aproximadamente.
- El ser mujer entre los 40 a 49 años reduce el uso de anticonceptivos en 80% aproximadamente.
- El ser mujer y tener una educación a niveles escolares bajos, reduce el uso de anticonceptivos solo el 1% aproximadamente.

- El ser mujer y tener deseos de hijos incrementa en 54% el uso de anticonceptivos, aproximadamente.
- El intercepto que representa a las mujeres de menos de 25 años con un nivel de educación alto y el deseo de no tener hijos, incrementa el de manera muy significativa el uso de anticonseptivos.

Esto último hace sentido, ya que al ser una mujer preparada y de temprana edad, y sin deseos de tener hijos, hace que dicha mujer se cuide, haciendo uso de anticonceptivos.

f) Verifique la validez del modelo de regresión logística

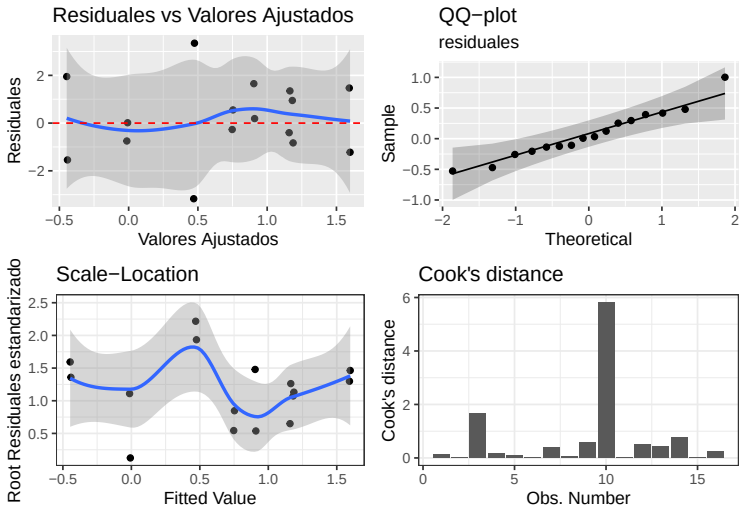
Se verifica la validez de la regresión logística utilizando la prueba de la razón de verosimilitud, con hipótesis nula: el modelo reducido es el adecuado. Por lo tanto, se busca rechazar la hipótesis nula.

Likelihood Ratio test:

Tabla 0.12: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	15	165.8	NA
Edad	3	113.8	12	51.96	1.662e-24
Educacion	1	0.08251	11	51.88	0.7739
Hijos	1	13.51	10	38.38	0.0002378

Se observa que a un nivel del 0.05, al menos una de las variables es diferente de cero, por lo tanto, se tiene evidencia para rechazar la hipótesis nula.



- Se obserba que se cumple lo dicho para los modelos lineales generalizados, la distribución no proviene de una normal y los datos presentan varianza no constante.

En conclusión, el modelo en conjunto sí es significativo y, acorde a los $p-values$ mostrados también

es significativa la contribución al modelo solo para los predictores de la edad y el deseo de tener hijos.

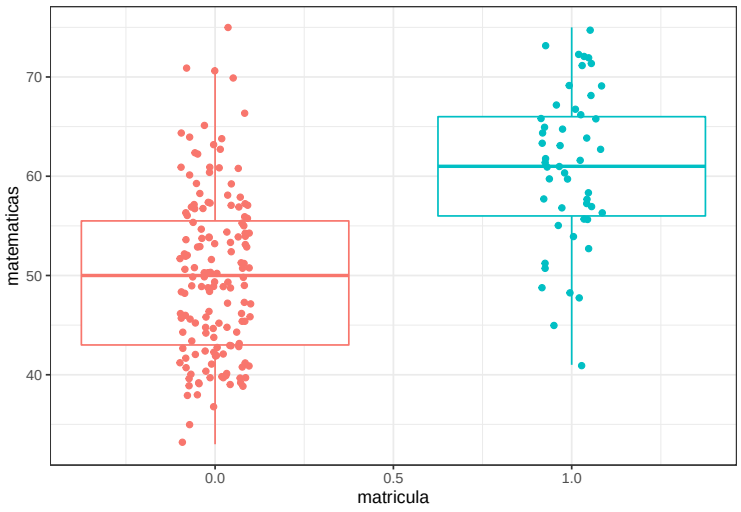
0.2 Ejercicio 2

Un estudio quiere establecer un modelo que permita calcular la probabilidad de obtener una matricula de honor al final del bachillerato en función de la nota que se ha obtenido en matemáticas. La variable matricula está codificada como 0 si no se tiene matricula y 1 si se tiene. El archivo Bachilleres.txt contiene los datos para el estudio.

Primeras 6 observaciones de la base de datos

matricula	matematicas
0	41
0	53
0	54
0	47
0	57
0	51

Descriptivo de la base:

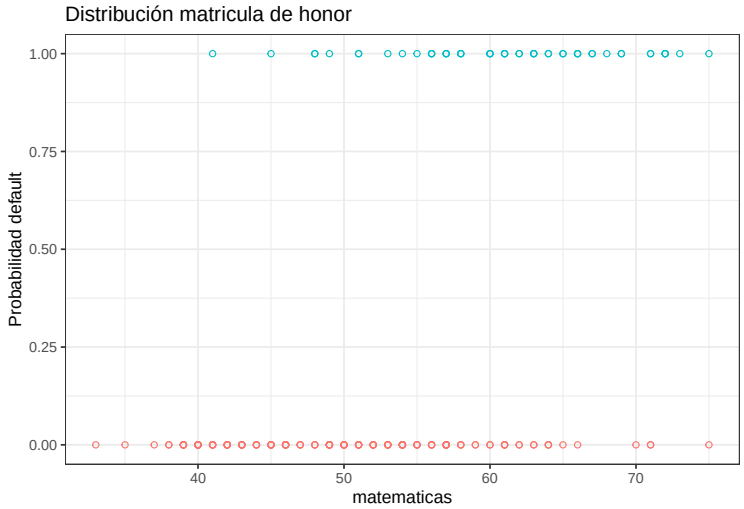


- Varianza distinta entre, tener o no tener, matricula de honor respecto a las notas de matemáticas.
- Los alumnos que no tienen matrícula de honor sus notas promedio en matemáticas están sobre el 50; caso contrario, los que si cuentan con matrícula de honor sus notas en matemáticas se encuentran por arriba del 60.

- Se observan valores atípicos en las notas de matemáticas de estudiantes que no tienen cuadro de honor.

Distribución de la variable respuesta

- Las notas bajas se relacionana más con no tener matrícula de honor



a) Ajuste un modelo de regresión logística a los datos.

Ajuste del modelo

$$matricula = \beta_0 + matematicas_{notas}$$

donde:

β_0: parámetro de los alumnos cuya nota en matemáticas es de cero

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.794	1.482	-6.61	3.85e-11
matematicas	0.1563	0.02561	6.105	1.029e-09

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	222.7 on 199 degrees of freedom
Residual deviance:	167.1 on 198 degrees of freedom

Significancia

- El log-odd del intercepto es significativo y se relaciona de manera negativa con tener matrícula

de honor.

- El log-odd de las notas en matemáticas es significativo y se relaciona de forma positiva con tener matrícula de honor

Interpretación log-odds

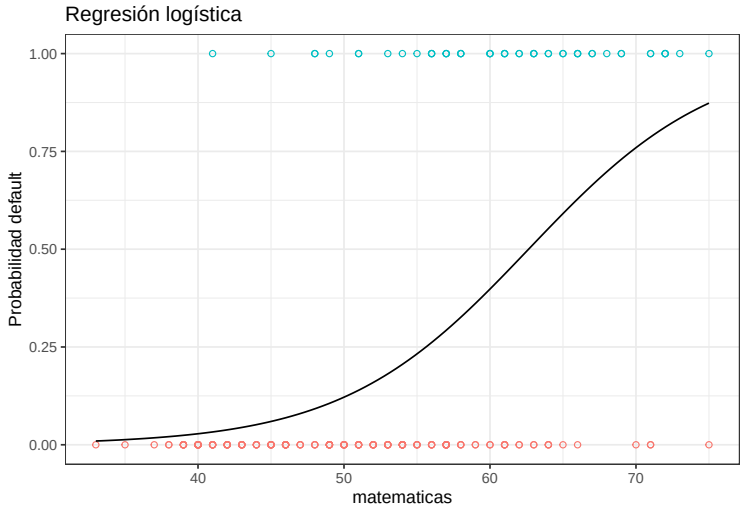
- log-odd del intercepto, indica que la intersección es el valor esperado del logaritmo de odds de que algún estudiante obtenga una matrícula teniendo cero en la nota de matemáticas.
- log-odd del parámetros de la nota en matemática, por cada unidad que se incrementa la variable matemáticas (la nota) se espera que el log-odds de la variable matrícula de honor se incremente en promedio 0.1563404 unidades.

La interpretación en términos del odds ratio se presenta en la pregunta c).

La Null deviance es la desviación para el modelo que no depende de ninguna variable. La Residual deviance es la diferencia entre la desviación del modelo que no depende de ninguna variable menos la correspondiente al modelo que incluye a la variable width

La diferencia entre ambas se distribuye como una distribución chi-cuadrado, el cual se vera en la prueba del test de verosimilitud.

b) Grafique el modelo e interprete los resultados.



Se observa la transformación de los datos y su ajuste con la función sigmoide, incluso con el modelo logístico se observa algo de confusión con las observaciones que se encuentran en la nota de matemáticas de valor 50 a 60.

c) Calcule el odds-ratio para este modelo y de su interpretación.

Se aplica exponencial a los parámetros del modelo (log-odds), y a los intervalos de confianza (al 95%)

	OR	2.5 %	97.5 %
(Intercept)	1e-04	0	8e-04
matematicas	1.169	1.116	1.234

Interpretacion

- los odds rate son muy bajos en el intercepto, lo que corresponde a una probabilidad de obtener una matrícual de 1e-04 cuando se tiene en matemáticas una nota de cero.
- el odds rate del parámetro de matemáticas indica que por cada unidad que se incrementa la variable matemáticas (notas), los odds de obtener matrícula de honor se incremente en promedio 1.169 unidades.

Una forma distinta de interpretar, y más general, es:

- Tener una nota en matématicas diferente de cero incrementa la probabilidad de tener una matrícula de honor en .169%

d) Verifique la validez del modelo.

Utilizando el test de la razón de verosimilitud, se tiene que la variable, notas en matemáticas, tiene relación respecto al tener una matrícula de honor, esto se observa dado al $p - valor$, el cual es muy pequeño, rechazando la hipótesis nula de no relación entre covariables y variable respuesta.

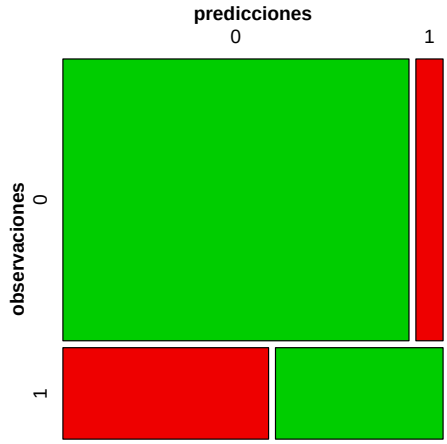
Tabla 0.17: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	199	222.7	NA
matematicas	1	55.64	198	167.1	8.718e-14

e) Compare los valores predichos con las observaciones y explique.

Se utiliza la matriz de coonfusión para obtener el número de predicciónes correctas.

	0	1
0	140	11
1	27	22



El modelo es capaz de clasificar correctamente:

$$\frac{144+22}{140+22+27+11} = .81$$

Es decir, el 81 de las observaciones utilizando los datos de entrenamiento

Conclusión

El modelo logístico creado para predecir la probabilidad de que un alumno obtenga matrícula de honor a partir de la nota de matemáticas es en conjunto significativo acorde al Likelihood ratio (p-value = 8.717591e-14). El p-value del predictor matematicas es significativo (p-value = 1.029e-09).

Modelo:

$$logit(matriculaHonor) = -9.794 + 0.1563 * NotaMate$$

$$P(matriculaHonor) = \frac{e^{-9.794+0.1563*NotaMate}}{1+e^{-9.794+0.1563*NotaMate}}$$

0.3 Ejercicio 3

El archivo Cangrejos.txt, contiene los datos de cangrejos herradura. Entre los cangrejos herradura se sabe que cada hembra tiene un macho en su nido, pero puede tener más machos concubinos. Se considera que la variable respuesta es el número de concubinos (Satellite) y las variables explicativas son: color (Color), estado de la espina central (Spine), peso (Weight) y anchura del caparazón (Width).

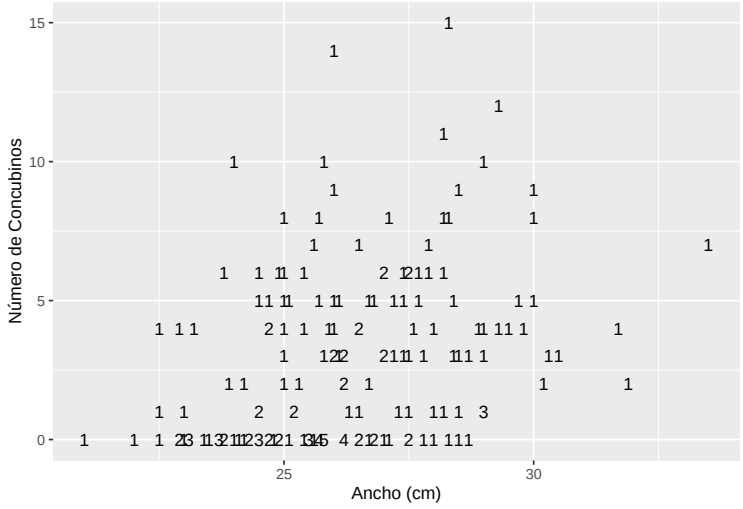
Base de datos, primeros 6 elementos:

Color	Spine	Width	Satellite	Weight
3	3	28.3	8	3050

Color	Spine	Width	Satellite	Weight
4	3	22.5	0	1550
2	1	26	9	2300
4	3	24.8	0	2100
4	3	26	4	2600
3	3	23.8	0	2100

Observamos la relación del número de concubinos respecto al ancho del caparazón.

- Se agregan aquellos cangrejos que tienen el mismo cm de anchura del caparazon, de esta forma los que dicen 1 son únicos, 2 que existen dos con esas medidas, y así.



- Se observa que el mayor número de concubinos que tiene una hembra se relaciona a que su ancho de caparazón sea alrededor de 25 a 30 cm

En un primer análisis solo considere la anchura del caparazón como variable explicativa.

Modelo

$$Satellite = \beta_0 + Width$$

a) Ajuste un modelo de regresió de Poisson a los datos.

Para el modelo de regresión de poisson se utiliza una función enlace del tipo logaritmo. Esto de acuerdo Raymundo H., Douglas, C., Geoffrey, V., y Timothy, J. (2010). Generalized Linear Models: with applications in engineering and the sciences.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.305	0.5422	-6.095	1.097e-09

	Estimate	Std. Error	z value	Pr(>|z|)
width	0.164	0.01997	8.216	2.095e-16

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	632.8 on 172 degrees of freedom
Residual deviance:	567.9 on 171 degrees of freedom

- La Null deviance es la desviaciónn para el modelo que no depende de ninguna variable.
- La Residual deviance es la diferencia entre la desviación del modelo que no depende de ninguna variable menos la correspondiente al modelo que incluye a la variable width
- La diferencia entre ambas se distribuye como una distribución chi-cuadrado con 1 grado de libertad y permite contrastar si el coeficiente de width puede considerarse nulo.

En la pregunta c) se utilizarán estos valores.

	Parametros	2.5 %	97.5 %
(Intercept)	-3.305	-4.366	-2.241
width	0.164	0.1247	0.203

Significancia

- El intercepto es significativo y se relaciona de forma negativa con el número de concubinos que tiene una hembra cangrejo.
- El parámetro de la anchura del caparazón es significativo y positivo; de este modo, un incremento en un centimentro de la anchura del caparazon aumenta en 0.164 el número de concubinos

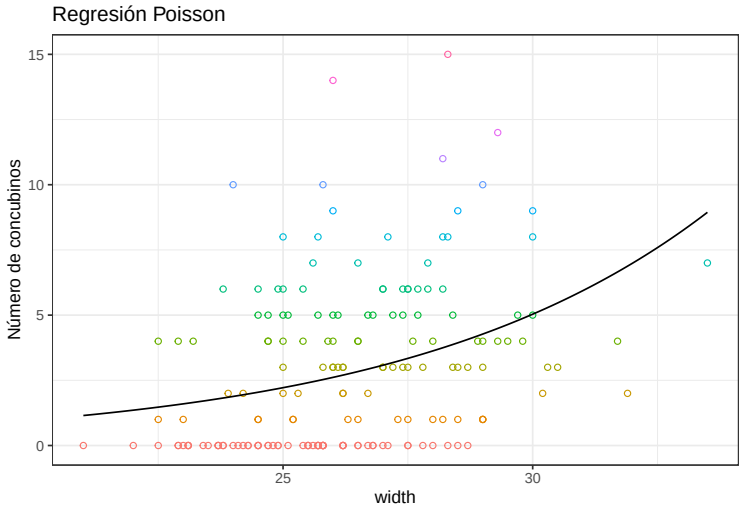
Interpretación

El efecto individual de la covariables se puede observar. Se tiene que el coeficiente de la anchura del caparazon es positivo en el predictor lineal, incrementando la media del número de concubinos. De esta manera, por cada centimetros más de anchura del caparazon, el número medio de concubinos que tiene una hembra es de:

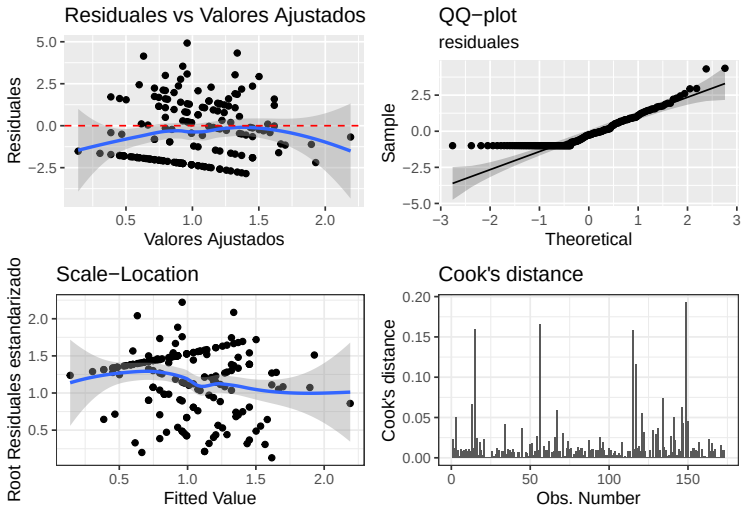
(Intercept)	width
0.0367	1.178

$$e^{Width} = 1.178$$

b) Grafique el modelo e interprete los resultados.



- Se puede ver que la recta de regresión no se ajusta del todo a los datos. Mostrando valores muy lejanos sobre la recta ajustada.



- Observando los residuales se aprecia no normalidad y una varianza nada constante, lo cual es usual en los modelos glm.

c) Verifique la validez del modelo.

Likelihood ratio test

Hacemos uso de los valores del NUll device y el residual device, como la diferencia de las desviaciones de los modelos reducidos y completos.

64.91

Se sabe que la diferencia entre ambas se distribuye como una chi-cuadrada con un grado de librtad, el valor de la distribución es:

```
## [1] 7.771561e-16
```

Esto nos permite contrastar si el modelo sin parámetros es el adecuado. Dado el p-valor de la distribución es muy chico, se tiene evidencia para rechazar la hipótesis nula, y por ende, se concluye que el modelo es adecuado al utilizar la anchura del caparazon para explicar el número de concubinos de la hembra (cangrejo).

Se puede utilizar la función anova en R para realizar la prueba likelihood ratio test:

Tabla 0.24: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	172	632.8	NA
width	1	64.91	171	567.9	7.828e-16

Se puede rechazar claramente la hipótesis nula. Hay un aportación significativa de la anchura del caparazón respecto a el número de concubinos.

d) Realice el mismo procedimiento con las variables explicativas restantes (color, estado de la espina central y peso). ¿Cuál de ellas resulta más explicativa para el modelo? ¿La original o alguna de las restantes? Explique por que.

Modelo

$$Satellite = \beta_0 + Color$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.987	0.1993	9.969	2.078e-23
color	-0.2729	0.05932	-4.601	4.204e-06

(Dispersion parameter for poisson family taken to be 1)

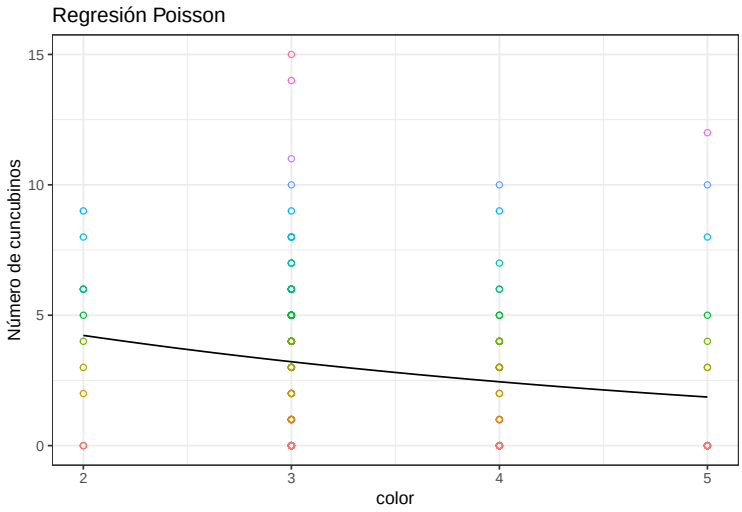
Null deviance:	632.8 on 172 degrees of freedom
Residual deviance:	610.7 on 171 degrees of freedom

Se tiene que el coeficiente del color es negativo en el predictor lineal, disminiyendo la media del número de concubinos. De esta manera, depende el tipo de color que sea, el número medio de

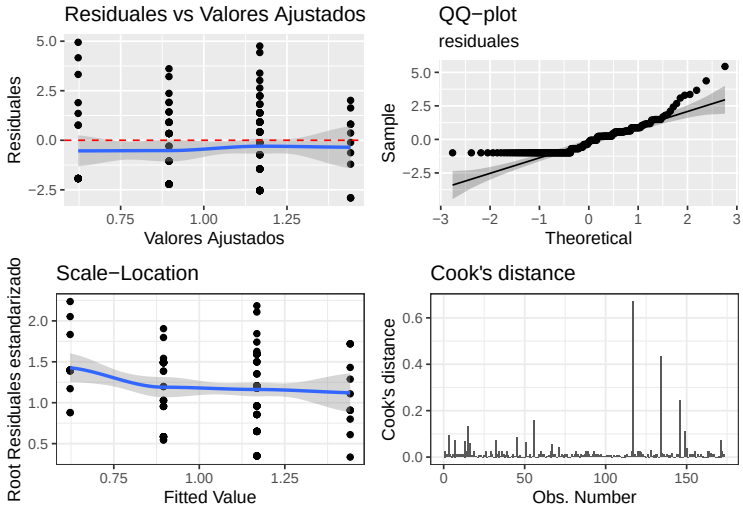
concubinos que tiene una hembra es de:

(Intercept)	color
7.295	0.7611

$$e^{color} = 0.7611$$



- El ajuste no se observa bien, ya que color es una variable con categorías.



- Observando los residuales vemos su comportamiento, los cuales no son los de observaciones normales, con una estructura de varianza heterocedastica.
- En el gráfico de coock distance observamos valores que influyen el desempeño de la regresión poisson.

Tabla 0.28: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	172	632.8	NA
color	1	22.13	171	610.7	2.547e-06

Se tiene evidencias suficiente para rechazar la hipótesis nula, por ende, el color que sea el cangrejo si afecta al número de concubinos que tiene la hembra.

Modelo

$$Satellite = \beta_0 + spine$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.345	0.1319	10.2	1.955e-24
spine	-0.1119	0.05159	-2.17	0.03003

(Dispersion parameter for poisson family taken to be 1)

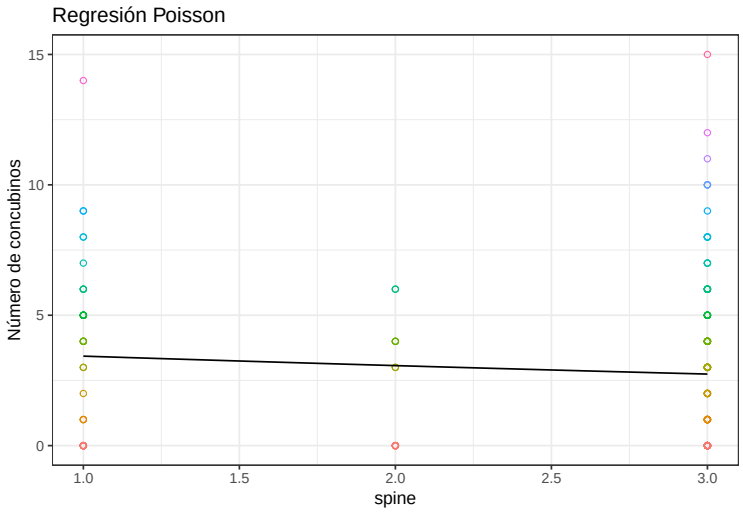
Null deviance:	632.8 on 172 degrees of freedom
Residual deviance:	628.2 on 171 degrees of freedom

Se tiene que el coeficiente de la estado de la espina central es negativo en el predictor lineal, disminuyendo la media del número de concubinos. De esta manera, dependiendo del estado de la espina central, el número medio de concubinos que tiene una hembra es de:

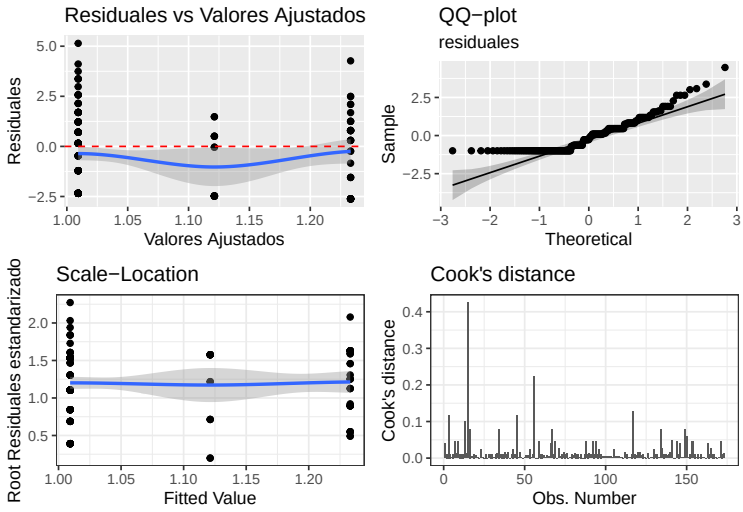
(Intercept)	spine
3.838	0.8941

$$e^{spine} = 0.8941$$

Asumiendo que las demás covariables son constantes



- AL igual que el anterior la recta no se ajusta bien a los datos categoricos.



- Al igual que el caso anterior, los residuales no vienen de una normal y presentan varianza no constante.

Tabla 0.32: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	172	632.8	NA
spine	1	4.57	171	628.2	0.03254

Se tiene ligera evidencias para rechazar la hipótesis nula, por ende, el estado de la espina central, sí afecta al número de concubinos que tiene la hembra.

Modelo

$$Satellite = \beta_0 + Weight$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4284	0.1789	-2.394	0.01665
weight	0.0005893	6.502e-05	9.064	1.258e-19

(Dispersion parameter for poisson family taken to be 1)

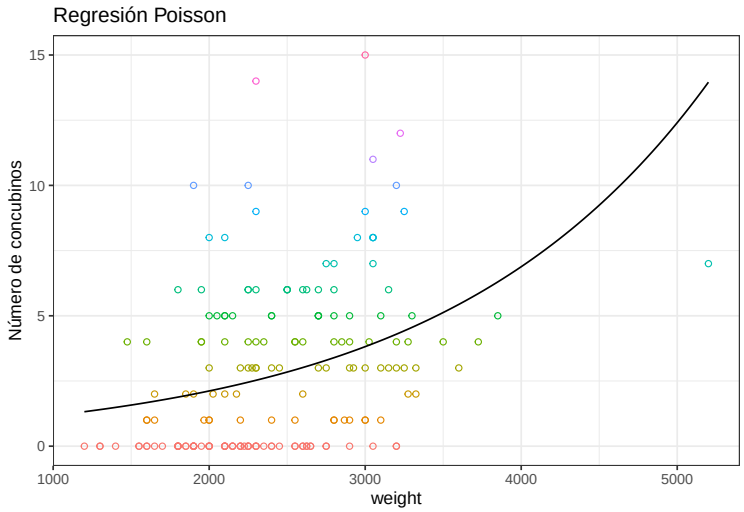
Null deviance:	632.8 on 172 degrees of freedom
Residual deviance:	560.9 on 171 degrees of freedom

Se tiene que el coeficiente del peso del cangrejo es positivo y pequeño en el predictor lineal, incrementando la media del número de concubinos. De esta manera, dependiendo del peso del cangrejo, el número medio de concubinos que tiene una hembra es de:

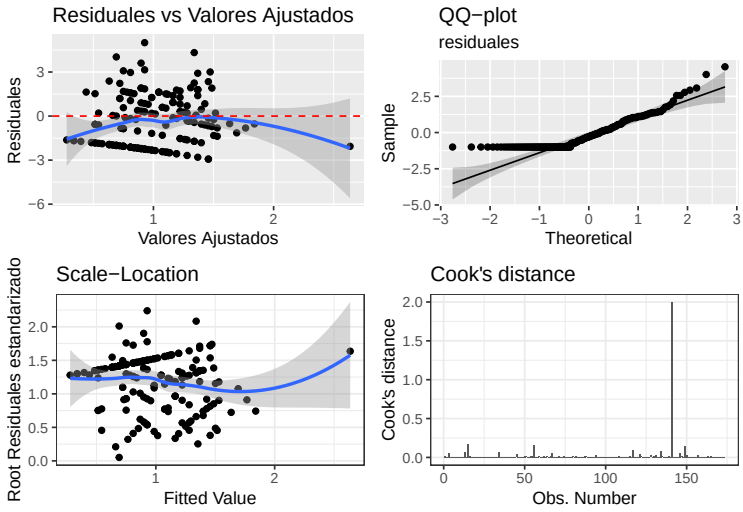
(Intercept)	weight
0.6515	1.001

$$e^{weight} = 1.001$$

Asumiendo que las demás covariables son constantes



- El ajuste mejora pero varios puntos siguen sin ajustarse a la recta de regresión.



- Se observa que los errores no tiene distribución normal, con errores que parecen que siguen a las predicciónes, indicando posible estructura de correlación.

Tabla 0.36: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	172	632.8	NA
weight	1	71.93	171	560.9	2.235e-17

Se tiene evidencias suficiente para rechazar la hipótesis nula, por ende, el peso, sí afecta al número de concubinos que tiene la hembra.

¿Cuál de ellas resulta más explicativa para el modelo?

¿La original o alguna de las restantes? Explique por que.

La original y el de los pesos del cangrejo. Al ajustar la recta de regresión se comportaban mejor con covariables continuas que discretas. Utilizando la prueba de verosimilitud, se descarta la variable del estado de la espina dorsal. Todos los parámetros para cada modelo fuero significativos.

Para seleccionar el mejor modelo se utiliza el críterio AIC:

```
##   concubino_anchura concubino_color concubino_spin concubino_peso
## 1          927.1762        969.9584       987.5194       920.1641
```

Como se observa el de menor valor es el modelo con covariable, pesos de los cangrejos, de esta forma, podría resultar más explicativo. Por otro lado, tanto el ancho del caparazon como el peso pueden están relacionados, y por ende, usar estás dos variables para estimar el número de concubinos de una hembra cangrejo, es adecuado.