Modelación lineal- taller 4

Cristian Gañan

## Punto 1

**Los siguientes datos encuestan mujeres casadas y fértiles por edad, tipo de educación, deseo de más hijos, si usan o no anticonceptivos. Modele la Anticoncepción como variable dependiente y a las demás como predictoras.**

library(tidyverse)  
punto1<- read.csv2("taller4.csv")  
punto1<- punto1 %>% rename(edad= ï..edad, hijos= mashi) %>%   
 mutate(edad= as.numeric(as.character(edad)))  
punto1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| edad | educ | hijos | noant | ant | tot |
| 22.5 | nop | Si | 53 | 6 | 59 |
| 22.5 | nop | No | 10 | 4 | 14 |
| 22.5 | pro | Si | 212 | 52 | 264 |
| 22.5 | pro | No | 50 | 10 | 60 |
| 27.5 | nop | Si | 60 | 14 | 74 |
| 27.5 | No | No | 19 | 10 | 29 |
| 27.5 | pro | Si | 155 | 54 | 209 |
| 27.5 | pro | No | 65 | 27 | 92 |
| 35.0 | nop | Si | 112 | 33 | 145 |
| 35.0 | nop | No | 77 | 80 | 157 |
| 35.0 | pro | Si | 118 | 46 | 164 |
| 35.0 | pro | No | 68 | 78 | 146 |
| 45.0 | nop | Si | 35 | 6 | 41 |
| 45.0 | nop | No | 46 | 48 | 94 |
| 45.0 | pro | Si | 8 | 8 | 16 |
| 45.0 | pro | No | 12 | 31 | 43 |

Se hace primero una coerción en la edad para ficilitar la modelación. Es facil notar que las variables hijos y educ son datos de tipo factor y se puden modelar mediante conteos, por esta razon es mas factible para este caso un glm() con conteos poisson:

dato1<- punto1 %>% nest()  
modelo<- function(data){  
 glm(ant ~ noant+hijos+educ+edad, data = data, family = "poisson")  
}  
  
dato1<- dato1 %>%   
 mutate(model= map(data, modelo))  
  
  
model2<- dato1 %>%   
 mutate(model2= map(model, broom::tidy)) %>%   
 unnest(model2, .drop = TRUE)  
  
model2

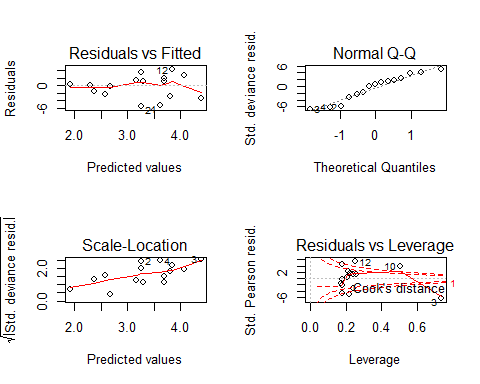
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.2677714 | 0.3957833 | 0.6765605 | 0.4986849 |
| noant | 0.0170053 | 0.0013052 | 13.0293007 | 0.0000000 |
| hijosSi | -1.3070681 | 0.1289814 | -10.1337718 | 0.0000000 |
| educnop | 0.3351922 | 0.3457232 | 0.9695392 | 0.3322762 |
| educpro | 0.3889661 | 0.3407543 | 1.1414856 | 0.2536679 |
| edad | 0.0622441 | 0.0080814 | 7.7021866 | 0.0000000 |

model3<- dato1 %>%   
 mutate(model3= map(model, broom::glance)) %>%   
 unnest(model3, .drop = TRUE)  
  
model3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual |
| 317.0881 | 15 | -69.62848 | 151.257 | 155.8925 | 60.75497 | 10 |

**Tabla 1 resultados modelo**

par(mfrow= c(2,2))  
plot(glm(ant ~ noant+hijos+educ, data = punto1, family = "poisson"))



**Figura 1** Modelo ajustado

a<- anova(glm(ant ~ noant+hijos+educ, data = punto1, family = "poisson"), test= "Chisq")

Según la **Figura 1** y la **tabla 1** este modelo se ajusta bien a los datos, con una deviance pequeña, ademas, haciendo un analisis de deviance al modelo puede noratarse que es significativo lo que permite establecer que el modelo ajusta bien a los datos, los residuales se comportan de forma homocedastica y todas las variables son significativas, sin embargo, el modelo parece complicado por tantas variables, se probó entonces un modelo alternativo con ant, noant y edad ya estas son variables cualitativas que podrian describir el fenomeno de una forma mas sencilla en cuanto a variables se refiere:

z<- lm(ant ~ noant+edad, family = "poisson", data = punto1)  
deviance(z)

## [1] 5676.487

Según el resultado anterior del modelo hallandole su devience para comparar con el modelo de todas las variables, se puede concluir que el modelo con noan, edad y ant no modelan bien los datos, su deviance es muy elevada en comparacion al modelo con todas las variables, este ultimo sería en consecuencia un buen modelo para estos datos.

## Punto 2

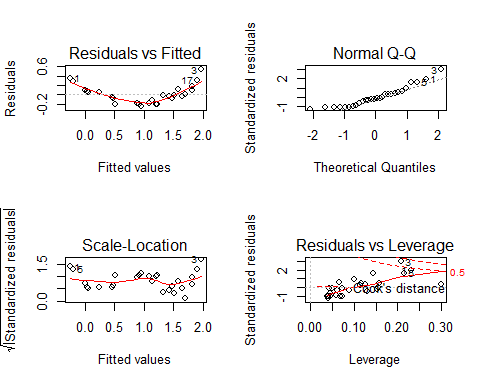
**Los siguientes datos corresponden al volumen de aserrío, obtenidos en forma general con el 20% de la altura total de los árboles. Modélelo haciendo uso del concepto de casi verisimilitud.**

La ídea principal de la casi verosimilitud, es la realación media varianza, para realizar esto lo que se hizo fue construir varios modelos, y despues comparar sus variaciónes con el fin de determinar el mejor modelo que ajustara a los datos.

punto2<- read.csv2("punto24.csv")  
  
punto2<- punto2 %>% rename(v= ï..v) %>%   
 mutate(v= as.numeric(as.character(v)), ht= as.numeric(as.character(ht)))

Como primera medida se grafica en un histograma la variable respuesta que para esta caso eligió a v, se hace esto para encontrar un patron en los datos y determinar “a ojo” un posible distribuciòn de la variable; lo anterior deja por sentado que se realizara un modelo general y no uno lineal, pues al hacerlo se encontro que los residuales no eran homocesaticos ,ademas, no eran normales. **figura 2**

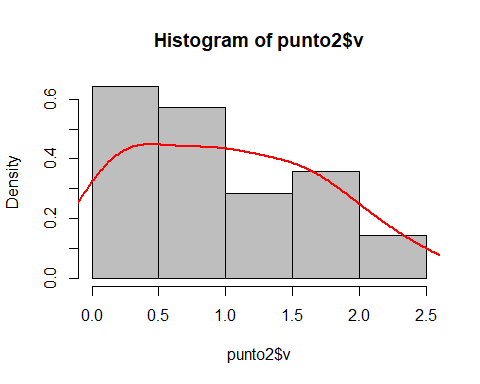
mod<- lm(v ~ d+ht, data = punto2)  
  
par(mfrow= c(2,2))  
plot(mod)



**figura 2**: Modelo lineal

Por lo anterior se procede a hacer modelos generales, primero se mira la distribución de la variable respuesta en un histograma:

hist(punto2$v, probability = TRUE, col= "grey")  
lines(density(punto2$v), lwd= 2, col= "red")



**Figura 3**: Distribución de v

Como se puede observar en la **Figura 3** v tiene cierto comportamiento de distribución poisson sesgada un poco a la izquierda ,es decir, que tiene un pequeño; teniendo este punto de partida se hara un model general con distrubución poisson y otros, teniendo en cuenta la hipotesis que los datos son de distribución poisson que se espera probar al desarrollar el ejercicio.

mod2<- glm(v ~ d+ht, data = punto2, family = "poisson")  
a<- broom::glance(mod2)  
  
mod3<- glm(v ~ d+ht, data = punto2, family = "gaussian")  
b<- broom::glance(mod3)  
  
mod4<- glm(v ~ d+ht, data = punto2, family = "Gamma")  
c<- broom::glance(mod4)  
  
  
mod5<- glm(v ~ d+ht, data = punto2, family = "quasi")  
d<- broom::glance(mod5)  
  
mod6<- glm(v ~ d+ht, data = punto2, family = "quasipoisson")  
e<- broom::glance(mod6)  
  
mod7<- glm(v ~ d+ht, data = punto2, family = "inverse.gaussian")  
f<- broom::glance(mod7)  
  
mol<- as.data.frame(rbind(a,b,c,d,e,f))  
mol$model<- c("poisson", "gaussian", "Gamma", "quasi","quasipoisson",  
 "inverse.gaussian")  
mol<- mol %>% arrange(deviance)  
mol

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | model |
| 15.05591 | 27 | -Inf | Inf | Inf | 0.400907 | 25 | poisson |
| 15.05591 | 27 | NA | NA | NA | 0.400907 | 25 | quasipoisson |
| 13.52964 | 27 | 6.193638 | -4.387276 | 0.9415418 | 1.053296 | 25 | gaussian |
| 13.52964 | 27 | NA | NA | NA | 1.053296 | 25 | quasi |
| 22.23375 | 27 | -8.112154 | 24.224308 | 29.5531261 | 6.195845 | 25 | Gamma |
| 49.70994 | 27 | -25.160043 | 58.320087 | 63.6489047 | 32.196963 | 25 | inverse.gaussian |

**Tabla 2: Resultados de modelos posibles**

En la **Tabla 2** se muestra en forma desendente de deviance los modelos, como se puede evidenciar, el modelo con poisson y quasipoisson son los que tienen menor devience, sin embargo, sus AIC y BIC no estan reportados, lo que hace dificil compararlos con otros modelos, en teoria la hipotesis que se dio al principo si “fue acertada” pero dado que los valores de AIC y BIC no estan hace que el modelo tenga una prueba no confiable pues solo se estaría basado su viabilidad en un solo test. Gaussian al contrario de los dos modelos anteriores, tiene una deviance “aceptable” y los AIC y BIC, son tambien bajos. Este proceso arroja como resultado teniendo en cuenta deviance, AIC y BIC que el mejor modelo seria en gussiano, pero no esta mal elegir un poisson, para aclarar este conficto, se procede a hacer una prueba mas, para observar cual seria el mejor ajuste.

Para tener una confianza y mas seguridad en elegir un buen modelo se harán pruebas de bondad de ajuste, la idea con esto es tener una prueba mas robusta y poder conseguir un buen modelo, se tuvo en cuenta los resultados antriores, por lo cual esta prueba se realiza para las distribuciones poisson, gamma, gaussian se eligen estas tres pues hay algunas que tienen la misma deviance u otros como inverse.gaussian la deviance es elevada en comparación a los demas modelos. Es

tab<- function(data, method, model){  
 ifelse(method == "poisson",   
 sum(((data - exp(predict(model)))^2)/exp(predict(model))),   
   
 sum(((data - (predict(model)))^2)/predict(model)))  
}

punto2 %>% summarise(poisson= tab(v, method = "poisson", model = mod2), gaussian= tab(v, method = "gaussian", model = mod3),gamma= tab(v, method = "gamma", model = mod4), chitab= qchisq(0.95, df= 27, lower.tail= T))

|  |  |  |  |
| --- | --- | --- | --- |
| poisson | gaussian | gamma | chitab |
| 0.3690994 | 12.50295 | 69.10726 | 40.11327 |

Según los resultados de la **Tabla 4** donde se muestran los calculado para las distristribuciónes, se puede concluir que el modelo de poison ó gaussian son los que mejor ajustan los datos pues sus calculados son menores a tabulado ahora, teniendo en cuenta que poisson tambien presento la menor devience se puede concluir que la distribución de poisson ajusta mejor los datos pero que la de gaussian tambien haría un buen modelo; esto coincide con la hipotesis inicial derivada del histograma de v.

finalmente se muestran los estimados del modelo:

broom::tidy(mod2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -2.9920081 | 1.1521703 | -2.5968454 | 0.0094084 |
| d | 0.0565229 | 0.0426223 | 1.3261348 | 0.1847950 |
| ht | 0.0329725 | 0.0928729 | 0.3550281 | 0.7225685 |

Como se observa, no son significativos los parametros, sin embargo, ya se demostro que el modelo ajusta los datos.