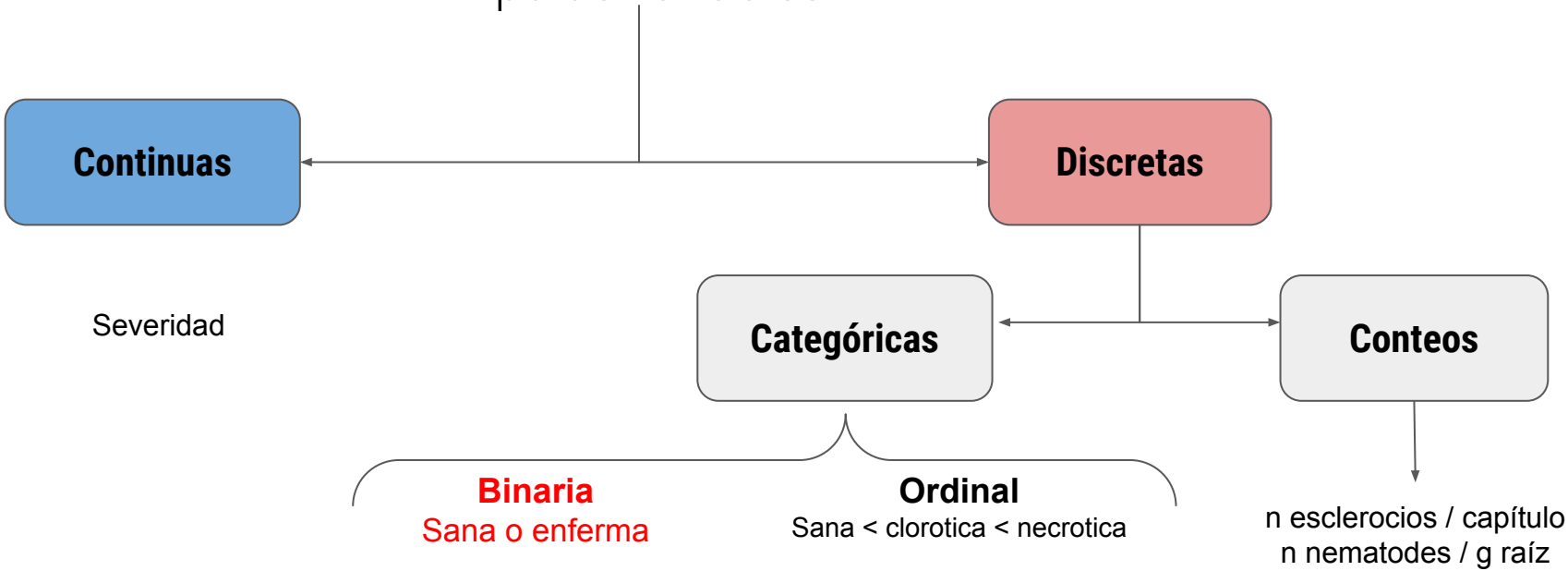
GLM para variables binomiales

Juan Pablo Edwards Molina
Juan Andrés Paredes
Bruno Pugliese

Evaluación visual de enfermedades

Tipo de variables



Normal	Binomial	Reg. ordinal o multinomial	Poisson
Modelos Lineales **LM**	Modelos lineales generalizados - **GLM**		

¿Qué hacemos?

¿Adecuamos nuestros datos a las técnicas analíticas?

o mejor

¿las técnicas analíticas a nuestros datos?

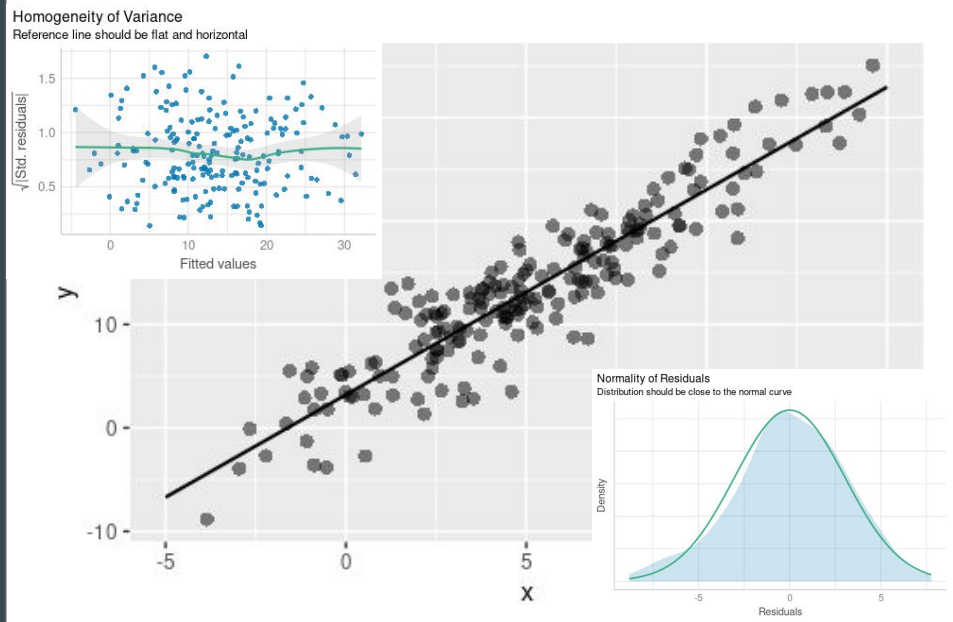
Modelos lineales (LM)

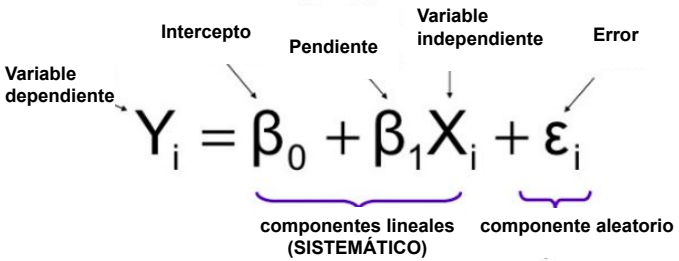
Técnica estadística para modelar **relaciones lineales** entre una variable dependiente y una o múltiple variables independientes.

Supuestos:

- independencia de las observaciones
- homocedasticidad de la varianza
- normalidad de los residuos

Regresión lineal





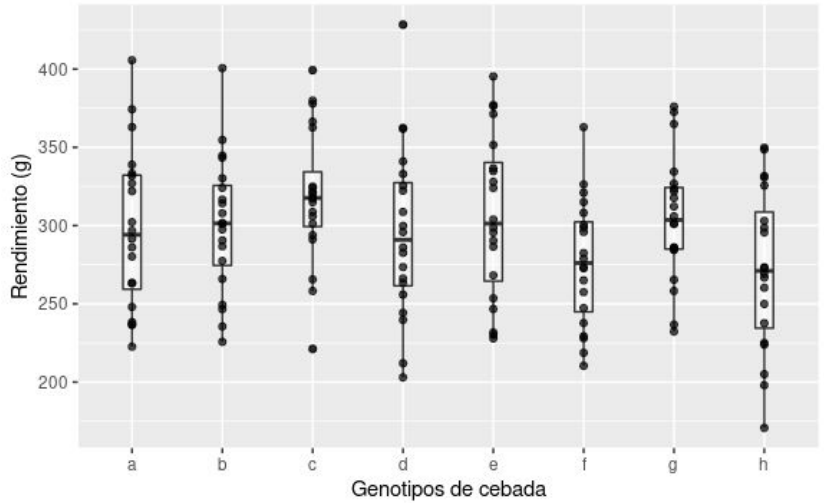
Modelos lineales (LM)

Técnica estadística para modelar **relaciones lineales** entre una variable dependiente y una o múltiple variables independientes.

Supuestos:

- independencia de las observaciones
- homocedasticidad de la varianza
- normalidad de los residuos

DBCA



Variable respuesta para el i-trt en el j-bk | **Constante Media gral.** | **Efecto del *i*-trt** | **Efecto del *j*-bk** | **Residual**

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}$$

$e_{ij} \sim N(0, \sigma_e^2)$

$b_j \sim N(0, \sigma_b^2)$

Modelos lineales generalizados (GLM)

Generalización flexible de los LM que admite variables respuesta con distribución de error distinta de una normal, al permitir que el componente lineal se relacione con la variable respuesta a través de una **función de enlace** (link).

Una propiedad de distribuciones no-normales, en general, es que la **varianza** de la distribución es **función de la media.** Esto significa que los niveles de un factor (tratamientos) tendrán diferentes varianzas (violación a los supuestos de los LM : varianzas constantes)

Supuestos

- Independencia de Y (como fueron tomados los datos? qué tipo?)
- Correcta función de enlace
- Ausencia de observaciones influyentes

LM

Componente sistemático (pred. lineal)

$$\mu = \beta_0 + \beta_1 x$$

Componente aleatorio

$$y_i = \text{Normal}(\mu_i)$$

GLM

Componente sistemático (pred. lineal)

$$\eta = \beta_0 + \beta_1 x$$

Función de enlace

$$\eta = \text{link}(\mu)$$

Componente aleatorio

$$y_i = \text{distribución}(\mu_i)$$

GLM
para variable binomial

Regresión logística

Componente sistemático

$$\eta = \beta_0 + \beta_1 x$$

Función de enlace

$$\eta = \text{logit}(\mu_i) = \log(\mu_i / 1-\mu_i)$$

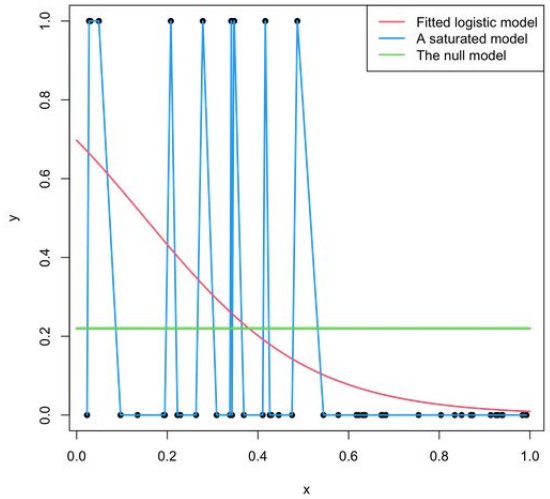
Componente aleatorio

$$y_i = \text{binomial}(\mu_i)$$

GLM
para variable binomial

Análisis de deviance

Generalización del análisis de la varianza para los GLM obtenido para una secuencia de modelos anidados (cada uno incluyendo más términos que los anteriores).
La deviance mide la desviación del GLM con respecto a un modelo perfecto para la muestra (modelo saturado), la cual se ajusta perfectamente a los datos



Modelo lineal generalizado para variable conteo

Regresión poisson

Componente sistematico

$$\eta = \beta_0 + \beta_1 x$$

Función de enlace

$$\eta = \log(\mu)$$

Componente aleatorio

$$y_i = \text{poisson}(\mu_i)$$

Variables Binomiales I

Incidencia

- Nivel intra-planta
  - frutas de naranja con antracnosis 10.1094/PDIS-01-19-0068-RE
  - virus: ToCV en hojas de tomate ($elisa^{-}$=0 ; $elisa^{+}$=1) 10.1094/phyto-06-18-0203-r
- Nivel parcela
  - Vainas de maní fuera del estándar comercial (No=1; Si=0) 10.1016/j.cropro.2020.105403
- Nivel lote
  - CABMV virus en plantas de maracuya (0-1) 10.1111/ppa.13054

Prevalencia

- Ausencia / presencia de phomopsis del girasol en un lote

Otros

- ¿Se solventó el tratamiento fungicida? No=0; Si=1
- ¿Se alcanzó el umbral de aplicación? No=0; Si=1

Probabilidad

p = n enfermos / n total

Chances

Odds = $p/(1-p)$ = *p* de enfermo / *p* de sano

Razón de chances (Odds ratio)

OR = $Odds_{trat} / Odds_{control}$

Caso 1

Aca podria ir cualquier variable binomial que se les ocurra (ver “otros” en slide 11)

trat	hj sanas 0	hj enfermas 1	p	odds	OR
check	3	7	0.7	0.7 / 0.3 = **2.33**	0.67 / 2.33 = **0.287**
fungi	6	4	0.4	0.4 / 0.6 = **0.67**	

Caso 1

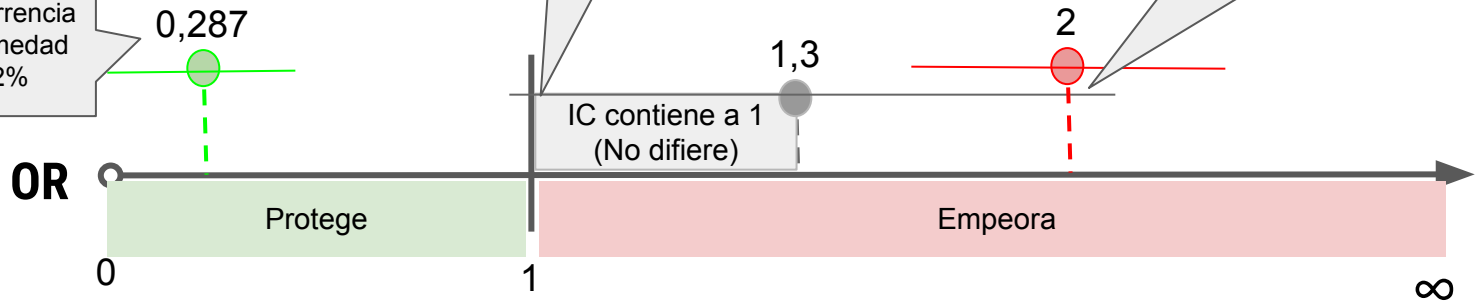
1-0,28=0,72*1-00
El trat redujo la prob de ocurrencia de la enfermedad en un 72%

Otros casos

El trat es indistinto respecto a la enfermedad

Otros casos

El trat duplicó la probabilidad de ocurrencia de la enfermedad



Ajuste de modelos glm en R: 3 posibilidades

trt	bk	y
1	1	0
1	1	0
1	1	1
1	1	0
1	1	1
1	1	0

trt	bk	n	enf
1	1	75	40
1	2	75	26
1	3	75	37
1	4	75	28
2	3	75	12
2	2	75	21

trt	bk	n	inc
1	1	75	0.53
1	2	75	0.35
1	3	75	0.49
1	4	75	0.37
2	3	75	0.16
2	2	75	0.28

Dato individualizado Cada fila representa una sola observación y la variable respuesta = 0 o 1 (o bien una variable con solo 2 valores: “sano” o “enfermo”)	glm(y~trt+bk, family = binomial(link = 'logit'), data = dat)
Datos agrupados Variable respuesta: Matriz de 2 columnas con: recuentos de 'éxitos' y recuentos de 'fallos'.	glm(cbind(enf, n-enf) ~ trt + bk, family = binomial(link = 'logit'), data = dat)
Datos agrupados Variable respuesta: Proporción entre 0 y 1 Especificar columna como 'peso' que da el número total del que proviene la proporción	glm(inc~trt+bk, weights = n, family = binomial(link = 'logit'), data = dat)

p	Es nuestra variable respuesta obtenida de: enfermos / n
log odds	$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$$
odds	*despejando el odds* $$\frac{p}{1-p} = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$
p(Y|X)	*despejando p* $$Pr(y_i = 1) = \frac{exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}$$

Así se estiman los coeficientes con GLM mediante la función de enlace logit.
Note la relación lineal entre el logit de p y las predictoras

calculamos los OR - relativos al nivel de referencia - para reportar los efectos de las predictoras

predecimos la probabilidad para valores de X

Over/underdispersion

La varianza residual es mayor/menor de lo esperado con el modelo ajustado
Más común para familias GLM con dispersión constante (fija), en particular para modelos de Poisson y binomiales, pero también puede ocurrir en familias GLM que ajustan la varianza (como la beta o binomial negativa) cuando se violan los supuestos de distribución.

Algunas reglas generales sobre el manejo de problemas de dispersión:

- La dispersión es una propiedad de los residuos, es decir, puede detectar problemas de dispersión solo DESPUÉS de ajustar el modelo. No tiene sentido mirar la dispersión de su variable de respuesta!
- La sobredispersión es más común que la subdispersión
- Si hay sobredispersión, el efecto principal es que los intervalos de confianza tienden a ser demasiado estrechos y los valores de p demasiado pequeños, lo que lleva a un error de tipo I inflado.
- Lo contrario es cierto para la sub-dispersión, es decir, es que pierde potencia.
- Una razón común para la sobredispersión es un modelo mal especificado. Cuando se detecta sobredispersión, primero se deben buscar problemas en la especificación del modelo (por ejemplo, graficando residuales contra predictores con DHARMa), y solo si esto no resuelve, las correcciones de sobredispersión tales como efectos aleatorios a nivel individual o cambios en la distribución se deben aplicar.

Conclusiones

1. Ajustamos la técnica de análisis a la naturaleza de nuestros datos
    a. Vimos que no llegamos a conclusiones similares mediante LM vs GLM
2. Los modelos mixtos nos permiten lidiar con la violacion de algunos supuestos de los GLM (independencia de las observaciones)
3. Actualmente hay paquetes de R para realizar el workflow completo de análisis (sin recurrir a cálculos manuales)

R - Outline

1. Repaso de conceptos básicos
2. DBCA (análisis de deviance) - *data phomopsis*
    a. Ajuste mediante LM y GLM - comparación
    b. Diagnósticos
    c. Interpretacion de coeficientes (log OR, OR, p)
3. Regresión logística - *data maracuyá*
    a. Single / multiple-point assessment
    b. Curva de progreso de la incidencia
    c. Predicciones

Sintaxis en R

Efectos fijos	Efectos mixtos
• {stats} lm	• {lme4} lmer • {nlme} lme +permite modelar varianza
• {stats} **glm** +family=quasibinomial	• {lme4} **glmer** • {glmmTMB} glmmTMB +tienen muchas alternativas de distribuciones • {MASS} glmmPQL (Penalized Quasi-Likelihood)

Distribución Binomial - propiedades

Y: Número de individuos con cierta carácter (EXITOS, ej., enfermedad) en una unidad experimental o muestral (ej., parcela, planta) – respuesta
n: Número de individuos observados para el carácter (ej., plantas)
p: Parámetro de localización: probabilidad de un carácter, como una enfermedad (ej., probabilidad de que una hoja, planta, etc., está enferma) (análogo a μ de normal)

- Para una simple muestra aleatoria de n plantas, la incidencia de la enfermedad (como proporción) es una estimación de **p**
- La varianza de la distribución condicional de Y es **np(1-p)** , completamente definida por **n** y **p**

Cuanto mayor n, Bin(p, n) se aproxima a la distribución normal para una muestra simple, con media **np** , y varianza **np(1-p)**

Recursos

Workshop 6: Generalized linear models

Chapter 5 Generalized linear models | Notes for Predictive Modeling

Logistic regression

Regresion Logistica: Interpretacion de Coeficientes. Pronosticos.

https://stats.oarc.ucla.edu/r/dae/logit-regression/

Using R to make sense of the generalised linear model | BARELY SIGNIFICANT

https://rpubs.com/benhorvath/logistic_regression

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/BS704-EP713_MultivariableMethods4.html

http://umh1480.edu.umh.es/wp-content/uploads/sites/44/2013/02/tema_5_1.pdf

http://glmm.wikidot.com/examples

https://stats.stackexchange.com/questions/185491/diagnostics-for-generalized-linear-mixed-models-specifically-residuals

https://www.youtube.com/watch?v=Gemf65XAH5s&list=PLUa2kfhXYC3RJ-IfkAdSf8Dg8mxM9Jg9z&index=1

https://bookdown.org/j_morales/librostat/glmbinomial.html#

glm ordinal

Journal of Plant Pathology
https://doi.org/10.1007/s42161-021-00805-5

ORIGINAL ARTICLE

Use of a growth regulator (prohexadione-Ca) and summer pruning as post symptom rescue treatments following a fire blight infection during bloom

Vincent Philion[1] · Valentin Joubert[1]

Received: 1 August 2020 / Accepted: 26 February 2021

Each of these shoots was rated for disease severity observed as a combined leaf and shoot necrosis score
0 = absence
1 = limited to central vein of inoculated leaves 2=extending to petiole
3 = reaching shoot
4 = reaching other leaves and so the shoot was apparently dead

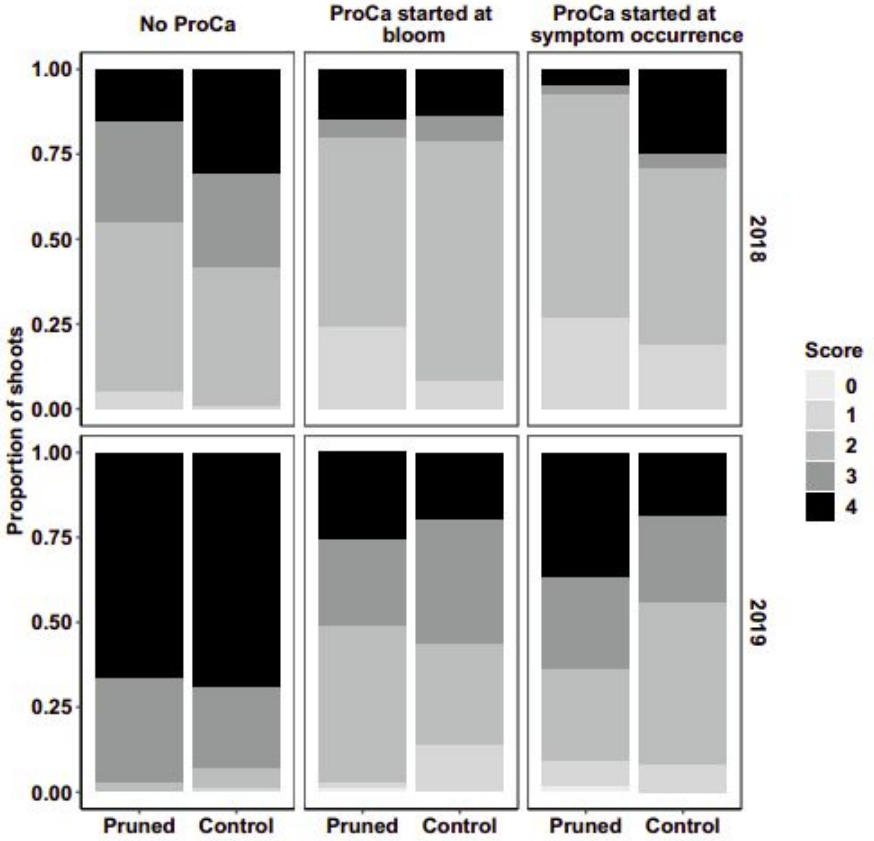


Fig. 3 Disease severity score distribution of inoculated shoots from plots either not pruned (Control) or pruned in summer (Pruned) and either unsprayed, sprayed with prohexadione-Ca (ProCa) starting at bloom, or when blossom symptoms first appeared. Apparently healthy shoots of all plots were inoculated 9 days (2018) and 14 days (2019) following the first ProCa treatment timing and observed 7 days (2018) or 9 days (2019) later. Disease score was based on necrosis extent (0 = absence, 1 = limited to central vein of inoculated leaves, 2 = extending to petiole, 3 = reaching shoot, 4 = reaching other leaves)

Table 4 Summary of the CLMM model describing the disease severity score of shoots inoculated in summer following ProCa applications and summer pruning interventions

Model terms[a]	Estimate[b]	S.E.[c]	z[d]	P-value
Year 2019	2.6	0.2	10.7	< 0.001
Year 2018: ProCa	-1.5	0.3	4.3	< 0.001
Year 2019: ProCa	-2.1	0.3	6.0	< 0.001
Year 2018: Unpruned	0.8	0.3	2.5	0.014
Year 2019: Unpruned	-0.3	0.3	0.9	0.38

[a]Significant main effects and interactions

[b]Log of the odds ratio

[c]S.E. = standard error

[d]z-value (parameter estimate/standard error) and associated probability