


Ejemplo guía para ejercicio I

Introducción a GLMs

A black and white microscopic image showing numerous circular parasites, likely oocysts, with a distinct outer layer and a central core. Some parasites are clustered together, while others are isolated. The background is dark and grainy.

Los parásitos afectan
diversos aspectos de
sus hospederos,
como su longevidad,
tasa de crecimiento o
éxito reproductivo.

Los alumnos del LEEYUS hicieron un experimento para evaluar los efectos de distintos parásitos en la tasa de crecimiento de una especie de pulga de agua, *Daphnia magna*, que es un modelo experimental ampliamente usado en ecología.

En su experimento, infectaron pulgas de agua con tres distintos tipos de bacteria: bacteria 1, bacteria 2 y bacteria 3, además de un nivel control donde las pulgas de agua no estaban infectadas. Registraron la variable de tasa de crecimiento en mm/día, e hicieron 10 réplicas de cada condición.



Daphnia magna

Pregunta de investigación e hipótesis del experimento

Los alumnos querían responder dos preguntas de investigación:

1. ¿La tasa de crecimiento de *D. magna* es afectada por la infección?
2. ¿Cuál de las bacterias tiene un mayor impacto sobre la tasa de crecimiento?

H0: La infección bacteriana no afecta la tasa de crecimiento de *D. magna*

H1. La infección bacteriana afecta la tasa de crecimiento de *D. magna*

Para contestar su pregunta, usaron la tasa de crecimiento como variable respuesta, y el tipo de tratamiento como variable predictiva. Tipo de tratamiento es una variable categórica ordinal de 4 niveles (control, Bacteria 1, Bacteria 2 y Bacteria 3)

Estructura algebraica del modelo

Los alumnos van a ajustar un GLM con la siguiente estructura:

tasa de crecimiento ~ tratamiento experimental

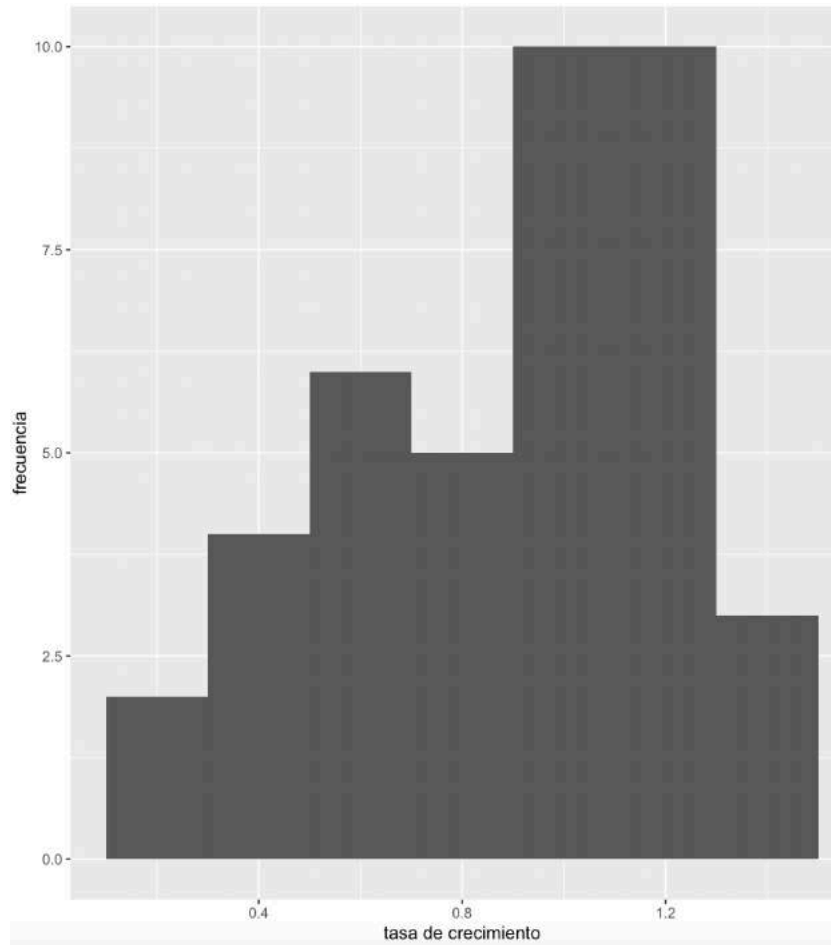
La variable predictiva, tipo de tratamiento, es una variable categórica de 4 niveles (control, Bacteria 1, Bacteria 2, Bacteria 3) por lo cual su estructura algebraica es de la forma:

$$\mathbf{f} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + c$$

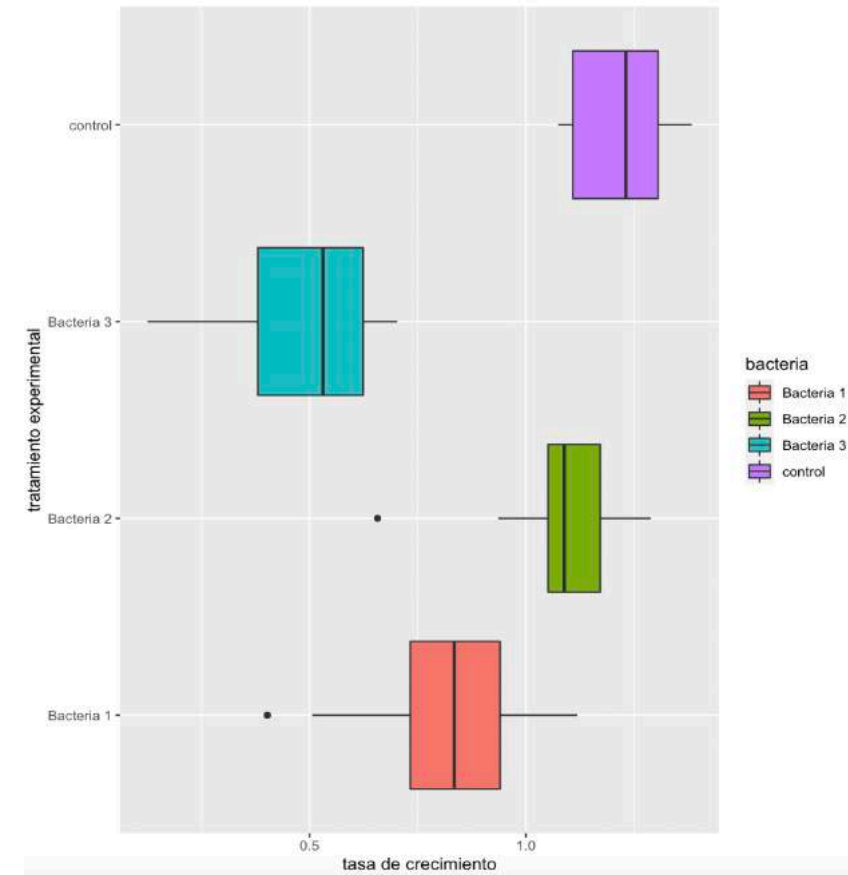
Donde \mathbf{f} se refiere a los valores ajustados que genera el modelo

Exploración gráfica

Los alumnos primero hacen un par de gráficas para explorar sus datos antes de hacer el modelo:

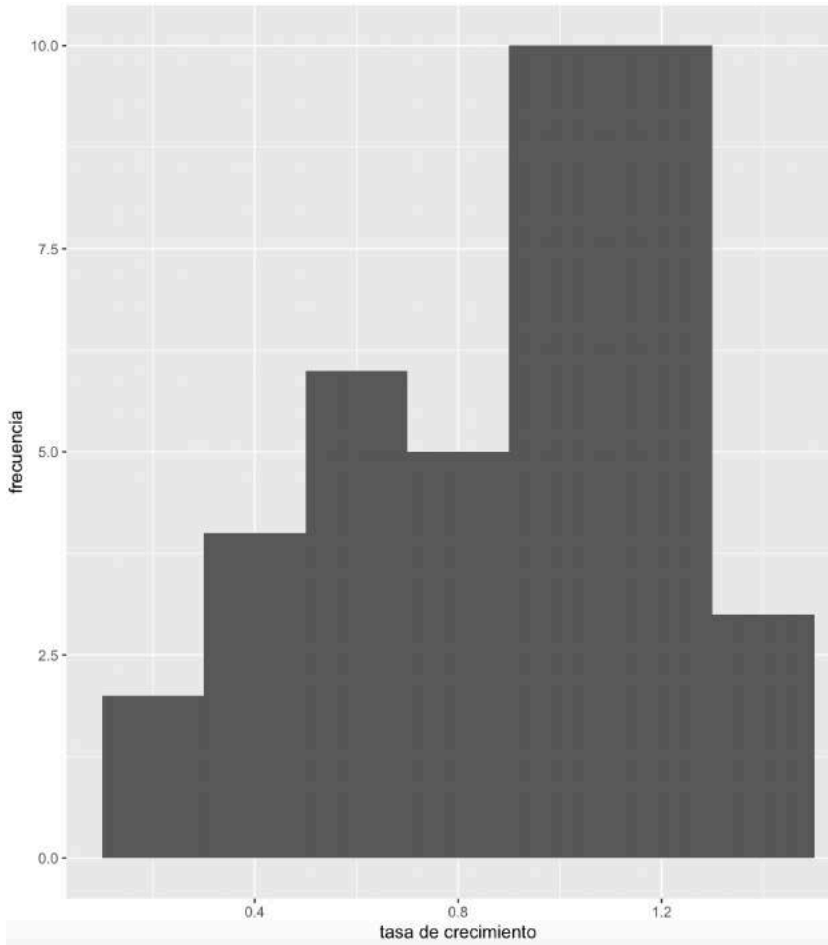


```
ggplot(daphnia, aes(x = tasa.crecimiento)) +  
  geom_histogram(binwidth = 0.2) +  
  xlab("tasa de crecimiento") + ylab("frecuencia")
```

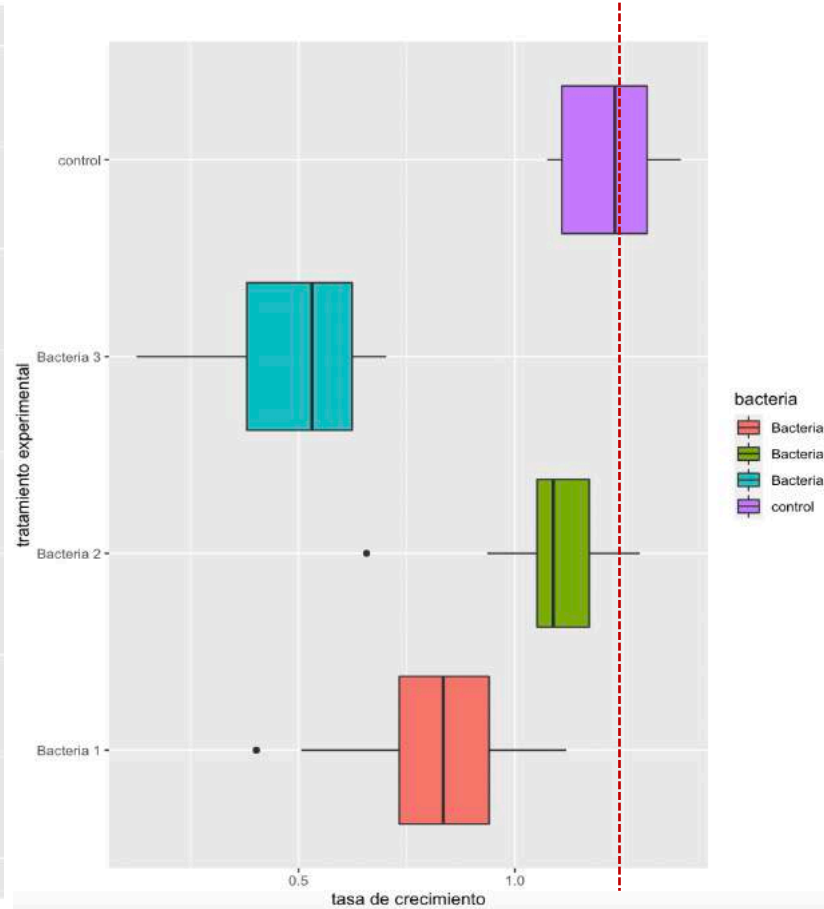


```
ggplot(daphnia, aes(x = tasa.crecimiento, y = bacteria, fill = bacteria)) +  
  geom_boxplot() +  
  xlab("tasa de crecimiento") + ylab("tratamiento experimental")
```

Exploración gráfica



Histograma



Caja y bigote

Cosas de las que nos podemos dar cuenta viendo los gráficos:

- La distribución de la variable tasa de crecimiento es relativamente simétrica (del histograma).
- Existe mucha variación en la tasa de crecimiento según el tipo de tratamiento. El tratamiento control (no infectado) genera la mayor tasa de crecimiento en *D. magna*, con una **mediana*** de ~1.2 mm/día (línea roja).
- Los gráficos parecen indicar que sí hay un efecto de la infección en la tasa de crecimiento (comparar con control), donde la Bacteria 3 es la que más parece afectar el crecimiento.

*Recuerden que un diagrama de caja y bigotes muestra la mediana, no la media. En una distribución simétrica, la mediana y la media son relativamente cercanas.

GLM - ¿La tasa de crecimiento de *D. magna* es afectada por la infección?

Modelo

tasa de crecimiento ~ tratamiento experimental

```
modelo_daphnia <- lm(tasa.crecimiento~bacteria, data = daphnia)
```

```
anova(modelo_daphnia) #anova es el comando que usa R para generar la tabla de varianza.
```

Analysis of Variance Table

Response: tasa.crecimiento

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bacteria	3	3.1379	1.04597	32.325	2.571e-10 ***
Residuals	36	1.1649	0.03236		

--- **TSS**= 4.3028

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

La variable predictiva "bacteria" (4 niveles de tratamiento) me explica 3.1379 unidades de variación de un total de 4.3028. Mi estadístico me indica que el valor de $p < 0.05$, y por tanto, rechazo H_0 y concluyo que la infección con bacterias sí tiene un efecto en la tasa de crecimiento.

Nota: recuerden que **TSS** es la suma total de cuadrados, y que **TSS** es una medida de la variación total en un variable (en referencia a una media global). Un GLM toma esta variación total y la reparte entre la que podemos explicar (con la variable predictiva), y la que no podemos explicar y queda como variación residual (1.1649 en este caso).

Paréntesis estadístico

- Nuestra hipótesis biológica es que la infección bacteriana tiene un efecto sobre la tasa de crecimiento, para lo cual hicimos un diseño experimental con 4 grupos de tratamiento.
- Esta hipótesis biológica se puede “traducir” a una hipótesis estadística: estadísticamente hablando, nos interesa saber si la media de cada grupo es diferente (indicando un efecto de tratamiento).
- La hipótesis nula (estadísticamente hablando) nos dice que todos los grupos provienen de poblaciones que tienen la misma media.
- Por lo tanto, si rechazamos la hipótesis nula, estamos diciendo que la media de cada grupo de tratamiento es estadísticamente distinta.
- El experimento controla todas las condiciones para que lo único que sea diferente en nuestros grupos sea el tratamiento experimental- y por tanto, cualquier diferencia en la media entre grupos, se pueda atribuir al tratamiento.

GLM

La tabla de varianza (`anova(modelo)`) nos permite determinar el efecto de tratamiento en la variación total de nuestros datos, y concluir que el tratamiento (infección bacteriana) sí tiene un efecto sobre la tasa de crecimiento. Pero para contestar la segunda pregunta de investigación tenemos que revisar los resultados de los coeficientes (parámetros) del modelo.

GLM- ¿Cuál de las bacterias tiene un mayor impacto sobre la tasa de crecimiento?

Resultados de coeficientes del modelo

```
summary(modelo_daphnia)
```

Call:

```
lm(formula = tasa.crecimiento ~ bacteria, data = daphnia)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.41930	-0.09696	0.01408	0.12267	0.31790

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.21391	0.05688	21.340	< 2e-16 ***
bacteriaBacteria 1	-0.41275	0.08045	-5.131	1.01e-05 ***
bacteriaBacteria 2	-0.13755	0.08045	-1.710	0.0959 .
bacteriaBacteria 3	-0.73171	0.08045	-9.096	7.34e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

En un modelo puramente categórico (ANOVA n-vías), el nivel del intercepto corresponde al nombre del nivel que venga primero en orden alfabético. SIN EMBARGO, en este caso nos interesa que el nivel de “control” sea el valor de referencia. ¿Por qué?

Residual standard error: 0.1799 on 36 degrees of freedom

Multiple R-squared: 0.7293, Adjusted R-squared: 0.7067

F-statistic: 32.33 on 3 and 36 DF, p-value: 2.571e-10

GLM- ¿Cuál de las bacterias tiene un mayor impacto sobre la tasa de crecimiento?

Call:

```
lm(formula = tasa.crecimiento ~ bacteria, data = daphnia)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.41930	-0.09696	0.01408	0.12267	0.31790

Coefficients:

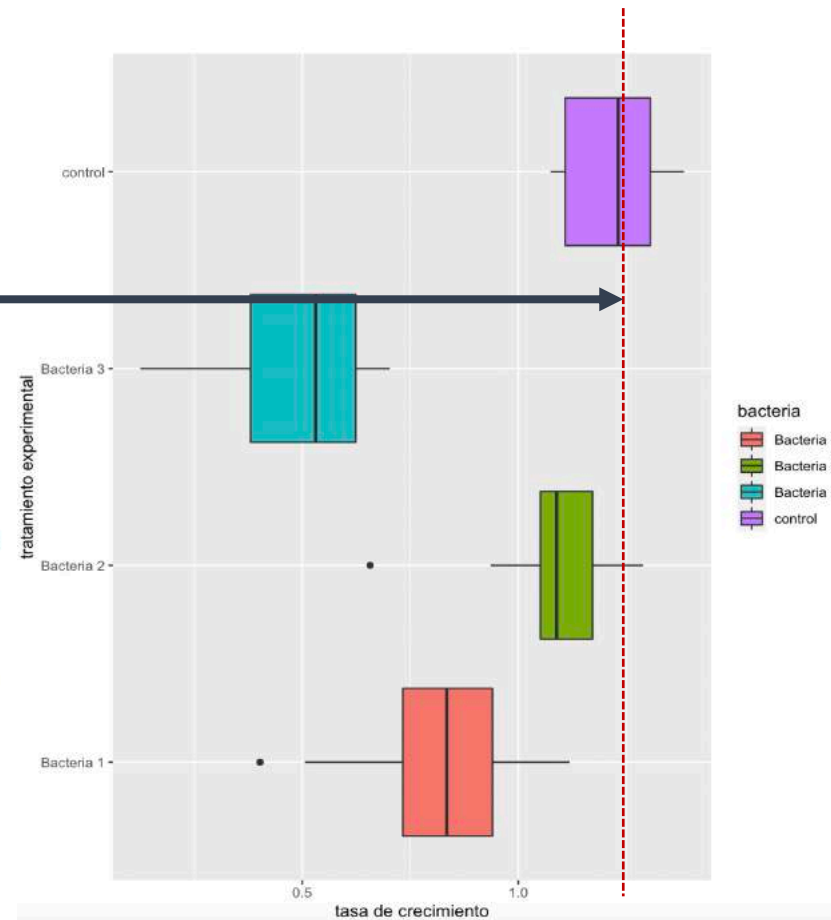
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.21391	0.05688	21.340	< 2e-16	***
bacteriaBacteria 1	-0.41275	0.08045	-5.131	1.01e-05	***
bacteriaBacteria 2	-0.13755	0.08045	-1.710	0.0959	.
bacteriaBacteria 3	-0.73171	0.08045	-9.096	7.34e-11	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1799 on 36 degrees of freedom

Multiple R-squared: 0.7293, Adjusted R-squared: 0.7067

F-statistic: 32.33 on 3 and 36 DF, p-value: 2.571e-10



Si el intercepto es la media estimada de tasa de crecimiento por el modelo para el nivel control, los coeficientes de los otros niveles son la **diferencia** entre la tasa media de crecimiento de Daphnias no infectadas y la tasa media de crecimiento de Daphnias infectadas con distintas bacterias. El efecto es mayor o menor según la especie de bacteria.

GLM- ¿Cuál de las bacterias tiene un mayor impacto sobre la tasa de crecimiento?

Coefficients:

	Estimate
(Intercept)	1.21391
bacteriaBacteria 1	-0.41275
bacteriaBacteria 2	-0.13755
bacteriaBacteria 3	-0.73171

Nivel de referencia

$$\mathbf{f} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + c \quad \mathbf{f} = \begin{bmatrix} a_1 \ 0 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + c \quad \mathbf{f} = \begin{bmatrix} 0 \\ -0.41 \\ -0.14 \\ -0.73 \end{bmatrix} + 1.21 \quad \mathbf{f} = \begin{bmatrix} 1.21 \\ 0.80 \\ 1.07 \\ 0.48 \end{bmatrix}$$

The diagram illustrates the calculation of the growth rate for different bacterial treatments. It shows four equations for the vector \mathbf{f} , each followed by a constant c . The first equation shows a general vector $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$. The second equation shows the vector $\begin{bmatrix} a_1 \ 0 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$, with a purple arrow pointing to the a_1 term, indicating the reference level. The third equation shows the vector $\begin{bmatrix} 0 \\ -0.41 \\ -0.14 \\ -0.73 \end{bmatrix}$, with a red bracket underneath, representing the difference in growth rate between the reference level and each treatment level. The fourth equation shows the final estimated growth rates: $\begin{bmatrix} 1.21 \\ 0.80 \\ 1.07 \\ 0.48 \end{bmatrix}$, with a red bracket to its right.

Tasa media de crecimiento estimada por el modelo para cada nivel de tratamiento

Diferencia en tasa de crecimiento entre nivel de referencia y cada nivel de tratamiento

Finalmente....

Después de su experimento, los alumnos concluyeron que el tratamiento de infección redujo el crecimiento de *Daphnia magna*, y que la bacteria 3 generó el mayor efecto en reducción en la tasa de crecimiento en comparación al control.

Recordatorio:

Recuerden que estamos tomando una muestra, por lo cual el modelo nos va a regresar un *estimado* de la población. Específicamente, nos va a regresar un estimado de la media para cada tratamiento – nos falta hablar de intervalos de confianza, pues todo estimado debe ir acompañado de una medida de incertidumbre.

Para ajustar este tipo de modelos, R usa el método de mínimos cuadrados, que consiste en encontrar los valores de los parámetros (los valores de los coeficientes que nos regresa el modelo) que minimice la variación residual.

Bono

¿En cuántas observaciones se basó el modelo?

Analysis of Variance Table

Response: tasa.crecimiento

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bacteria	3	3.1379	1.04597	32.325	2.571e-10 ***
Residuals	36	1.1649	0.03236		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4 niveles en total (pero como uno de esos niveles es de referencia, entonces no se toma en cuenta);

entonces la variable predictiva usa 3 grados de libertad. Entonces, en reversa: $36 + 3 = 39$

Sabemos de antemano que, de una muestra, sólo necesitamos $n-1$ de las observaciones para cuantificar la variación. Si hacemos el proceso en reversa, tenemos $39 + 1 = 40$

Por lo tanto, tenemos **40 observaciones** en la base de datos original.

Ahora al ejercicio