GLM-GLMM

Santos G

Tabla de contenidos

1 Contexto de proyecto

En esta sección se introducen los Modelos Lineales Generalizados (GLM) y Modelos Lineales Mixtos Generalizados (GLMM) aplicados a datos ecológicos. A diferencia de los modelos lineales simples o múltiples, los GLM permiten modelar variables de respuesta que no siguen una distribución normal, como conteos (Poisson o binomial negativa), proporciones o datos binarios (presencia/ausencia). Los GLMM extienden aún más este marco al incorporar efectos aleatorios, útiles para controlar la variabilidad entre grupos o unidades experimentales (por ejemplo, sitios de muestreo, individuos o años).

2 Carga de librerías y dataset

```
# Librerías necesarias (carga al inicio del documento)
library(tidyverse)     # manipulación y ggplot2
library(broom)         # tidy(), glance()
library(knitr)         # kable()
library(lme4)          # glmer()
```

```r
library(broom.mixed)   # tidy() para modelos mixtos
library(pROC)          # AUC / ROC (diagnóstico)
library(janitor)       # Clean names
library(palmerpenguins) # dataset de pingüinos

# Cargar dataset
df_raw <- penguins %>% as_tibble()
```

3 Preparación de datos (crear variable binaria y limpiar)

```r
# Crear variable binaria: "pico largo" (ejemplo umbral = 45 mm)
df_glm <- df %>%
  janitor::clean_names() %>%
  mutate(
    long_bill = case_when(
      !is.na(bill_length_mm) & bill_length_mm > 45 ~ 1,
      !is.na(bill_length_mm) & bill_length_mm <= 45 ~ 0,
      TRUE ~ NA_real_
    )
  ) %>%
  select(species, island, flipper_length_mm, bill_length_mm, long_bill) %>%
  drop_na(flipper_length_mm, long_bill) # quitar filas sin predictor o respuesta

# Comprobar distribución de la variable binaria
df_glm %>%
  count(long_bill) %>%
  knitr::kable()
```

Tabla 1: Distribución de la variable respuesta (long_bill)

long_bill	n
0	177
1	165

Se observa que 177 pingüinos (≈52%) tienen pico corto o normal (≤45 mm), mientras que 165 pingüinos (≈48%) tienen pico largo (>45 mm) (ver Tabla 1). Es decir, la variable respuesta está balanceada — hay una proporción similar de casos “0” y “1”. Esto es excelente, porque un modelo logístico funciona mejor cuando las clases no están demasiado desbalanceadas (por ejemplo, 90% / 10%).

En términos biológicos, esto indica que dentro del conjunto de datos de pingüinos del archipiélago Palmer, las longitudes del pico se distribuyen de forma relativamente equilibrada alrededor del umbral de 45 mm. Por tanto, existen proporciones comparables de individuos con picos largos y cortos, lo cual permitirá explorar si esta característica está asociada a diferencias morfológicas (como el largo del aleta) o taxonómicas (como la especie).

4 Ajuste del GLM binomial (logit)

```r
# Ajuste GLM binomial: probabilidad de pico largo según flipper_length y especie
modelo_glm <- glm(long_bill ~ flipper_length_mm + species,
                  data = df_glm,
                  family = binomial(link = "logit"))

# Tabla de coeficientes y OR (odds ratios)
glm_coef <- broom::tidy(modelo_glm, conf.int = TRUE) %>%
  mutate(
    estimate = round(estimate, 4),
    std.error = round(std.error, 4),
    p.value = ifelse(p.value < 0.001, "< 0.001", round(p.value, 3)),
    OR = round(exp(estimate), 3),
    OR_lwr = round(exp(conf.low), 3),
    OR_upr = round(exp(conf.high), 3)
  )

knitr::kable(glm_coef)
```

Tabla 2: Coeficientes (logit) y Odds Ratios (OR)

term	estimate	std.error	statistic	p.value	conf.low	conf.high	OR	OR_lwr	OR_upr
(Intercept)	-33.5120	7.1452	-4.690167	< 0.001	-48.380869	-20.223668	0.000	0.000	0.000
flipper_length_mm	0.1533	0.0363	4.226267	< 0.001	0.0854135	0.228350	1.166	1.089	1.257
speciesChinstrap	6.2646	0.7741	8.092529	< 0.001	4.9003855	7.986774	525.631	134.342	2941.792
speciesGentoo	1.9542	0.9068	2.154914	0.031	0.2595677	3.869489	7.058	1.296	47.918

- **Intercepto ($\beta_0 = -33.51$, $p < 0.001$):**
  Representa el log-odds de tener un pico largo para un pingüino *Adelie* (grupo de referencia) con una longitud de ala igual a 0 mm. Aunque no tiene un significado biológico directo, es necesario para definir la ecuación logística del modelo.
- **Longitud del ala ($\beta_1 = 0.153$, $p < 0.001$):**
  El efecto es positivo y altamente significativo, lo que indica que a medida que aumenta la longitud del ala, también aumenta la probabilidad de tener un pico largo. En términos de odds ratio (OR = 1.166), por cada aumento de 1 mm en la longitud del ala, las probabilidades de tener un pico largo aumentan en aproximadamente 16.6 %, manteniendo constante la especie.
- **Especie Chinstrap ($\beta = 6.26$, $p < 0.001$):**
  En comparación con los pingüinos *Adelie*, los *Chinstrap* tienen odds 525 veces mayores de presentar un pico largo. Esto sugiere una diferencia morfológica marcada entre especies.
- **Especie Gentoo ($\beta = 1.95$, $p = 0.031$):**
  Los *Gentoo* también tienen mayor probabilidad de presentar picos largos respecto a los *Adelie*, con odds 7 veces mayores, aunque el efecto es más moderado que en *Chinstrap*.

En conjunto, el modelo logístico indica que tanto el tamaño corporal (longitud del ala) como la especie son predictores significativos de la probabilidad de poseer un pico largo. Las especies *Chinstrap* y *Gentoo* muestran mayor probabilidad de presentar picos largos en comparación con *Adelie*, lo que refleja diferencias alométricas y adaptativas relacionadas con el hábitat y tipo de dieta. La longitud del ala, a su vez, actúa como indicador de tamaño corporal general, reforzando la relación alométrica entre las dimensiones corporales (ver Tabla 2).

5 Resumen del ajuste y verosimilitud

```
# Resumen del ajuste GLM: desviancias, AIC, logLik
glm_glance <- broom::glance(modelo_glm) %>%
  tibble::as_tibble() %>%
  select(null.deviance, df.null, deviance, df.residual, AIC, logLik) %>%
  mutate(across(where(is.numeric), round, 3))

knitr::kable(glm_glance)
```

Tabla 3: Resumen del ajuste GLM (desviancias, AIC, logLik)

null.deviance	df.null	deviance	df.residual	AIC	logLik
473.692	341	166.918	338	174.918	-83.459

- **Comparación de desviancias:** La *null deviance* es 473.7 y la *residual deviance* baja a 166.9. Esa gran reducción (más de 300 unidades) indica que el modelo con predictores mejora muchísimo respecto a un modelo sin ellos. En otras palabras, *flipper_length_mm* y *species* aportan información muy significativa para explicar si un pingüino tiene un pico largo (>45 mm).
- **AIC (174.918):** Este valor por sí solo no tiene un significado absoluto, pero sí se usa para comparar modelos alternativos (por ejemplo, con o sin alguna variable). Si probaras un modelo solo con *flipper_length_mm*, su AIC sería más alto, lo que confirmaría que incluir *species* mejora el ajuste.
- **LogLik (-83.459):** Este valor está relacionado con la verosimilitud cuanto más alto (menos negativo), mejor el ajuste. A medida que agregas predictores relevantes, el logLik debería aumentar.

En conjunto el modelo logra una reducción drástica de la deviance (de 473 a 167), lo cual demuestra que el modelo predice bastante bien la probabilidad de "pico largo". La combinación de *flipper_length_mm* (una medida continua) y *species* (categoría) explica la mayor parte de la variación observada (ver Tabla 3).

6 Diagnósticos GLM: overdispersion y AUC / performance

```
# Overdispersion:ratio deviance/df.residual

dispersion <- with(summary(modelo_glm), deviance / df.residual)

# Predicciones y matriz de confusión simple (umbral 0.5)
```

```r
df_glm <- df_glm %>%
  mutate(
    fitted_prob = predict(modelo_glm, type = "response"),
    fitted_class = ifelse(fitted_prob >= 0.5, 1, 0)
  )

conf_tab <- table(Actual = df_glm$long_bill, Predicho = df_glm$fitted_class)
accuracy <- sum(diag(conf_tab)) / sum(conf_tab)

# AUC / ROC
roc_obj <- pROC::roc(df_glm$long_bill, df_glm$fitted_prob, quiet = TRUE)
auc_val <- round(pROC::auc(roc_obj), 3)

# Tablas resumen
diag_table <- tibble(
  measure = c("Dispersion (deviance/df)", "Accuracy (threshold 0.5)", "AUC"),
  value = c(round(dispersion, 3), round(accuracy, 3), auc_val)
)

knitr::kable(diag_table)
```

Tabla 4: Diagnósticos GLM: overdispersion, accuracy y AUC

measure	value
Dispersion (deviance/df)	0.494
Accuracy (threshold 0.5)	0.904
AUC	0.958

```r
knitr::kable(as.data.frame(conf_tab))
```

Tabla 5: Matriz de confusión (GLM, umbral 0.5)

Actual	Predicho	Freq
0	0	148
1	0	4
0	1	29
1	1	161

Se muestra los indicadores de desempeño del modelo logístico ajustado. La razón de dispersión (0.494) indica ausencia de sobre-dispersión, lo que confirma que el modelo binomial es apropiado. La precisión del 90.4% y el valor AUC de 0.958 evidencian una alta capacidad predictiva y discriminatoria del modelo, reflejando que la longitud de las aletas y la especie son buenos predictores de la probabilidad de presentar un pico largo (ver Tabla 4) . Además se observa que el número de aciertos fue sustancialmente mayor que los errores, reforzando la validez del modelo (ver Tabla 5) .

7 Predicciones: tabla presentable y ejemplo de interpretación

```r
# Predicción para valores representativos de flipper por especie
newdata <- expand.grid(
  flipper_length_mm = c(170, 190, 210),
  species = unique(df_glm$species)
) %>%
  as_tibble()

preds <- newdata %>%
  mutate(
    fit_prob = predict(modelo_glm, newdata = newdata, type = "response"),
    lwr = NA_real_, upr = NA_real_
  )

# obtener IC por link (logit) y transformar a prob (opcional)
pred_link <- predict(modelo_glm, newdata = newdata, se.fit = TRUE,
                     type = "link")
crit <- qnorm(0.975)
preds <- preds %>%
  mutate(
    fit = round(plogis(pred_link$fit), 3),
    lwr = round(plogis(pred_link$fit - crit * pred_link$se.fit), 3),
    upr = round(plogis(pred_link$fit + crit * pred_link$se.fit), 3)
  )

knitr::kable(preds)
```

Tabla 6: Predicciones (probabilidad de pico largo) por flipper_length y especie

flipper_length_mm	species	fit_prob	lwr	upr	fit
170	Adelie	0.0005840	0.000	0.005	0.001
190	Adelie	0.0123920	0.004	0.042	0.012
210	Adelie	0.2122326	0.056	0.552	0.212
170	Gentoo	0.0041077	0.000	0.089	0.004
190	Gentoo	0.0813552	0.015	0.341	0.081
210	Gentoo	0.6553492	0.522	0.768	0.655
170	Chinstrap	0.2349829	0.053	0.628	0.235
190	Chinstrap	0.8683359	0.731	0.941	0.868
210	Chinstrap	0.9929876	0.965	0.999	0.993

Se presentan las probabilidades predichas de presentar un pico largo en función de la longitud de las aletas y la especie. Se observa una relación positiva: al aumentar la longitud de las aletas, la probabilidad de tener un pico largo también aumenta. Sin embargo, esta relación varía entre especies. En particular, los pingüinos *Chinstrap* presentan la mayor probabilidad de tener picos largos, seguidos por *Gentoo*, mientras que *Adelie* muestra consistentemente bajas probabilidades. Estos resultados confirman que tanto el tamaño corporal como la especie influyen significativamente en la morfología del pico (ver Tabla 6) .

8 Ajuste GLMM (efecto aleatorio por isla)

```r
# Ajuste GLMM (binomial) con intercepto aleatorio por isla
modelo_glmm <- glmer(long_bill ~ flipper_length_mm + species + (1 | island),
            data = df_glm,
            family = binomial(link = "logit"),
            control = glmerControl(optimizer = "bobyqa",
                        optCtrl = list(maxfun = 200000)))

# Coeficientes fijados (fixed effects) y efecto aleatorio (varianza)
glmm_tidy <- broom.mixed::tidy(modelo_glmm, effects = "fixed",
          conf.int = TRUE) %>%
  mutate(
    estimate = round(estimate, 4),
    OR = round(exp(estimate), 3),
    p.value = ifelse(p.value < 0.001, "< 0.001", round(p.value, 3))
  )

rand_eff <- as.data.frame(VarCorr(modelo_glmm))
rand_eff_tbl <- tibble(
  term = rand_eff$grp,
  variance = round(rand_eff$vcov, 4)
)

knitr::kable(glmm_tidy)
```

Tabla 7: efectos fijos (coeficientes y OR)

effect	term	estimate	std.error	statistic	p.value	conf.low	conf.high	OR
fixed	(Intercept)	-33.5120	7.2003934	-4.654188	< 0.001	-47.6244935	-19.399470	0.000
fixed	flipper_length_mm	0.1533	0.0365621	4.193833	< 0.001	0.0816750	0.224996	1.166
fixed	speciesChinstrap	6.2646	0.7747393	8.086017	< 0.001	4.7460947	7.783017	525.631
fixed	speciesGentoo	1.9542	0.9102029	2.146946	0.032	0.1701916	3.738122	7.058

```r
knitr::kable(rand_eff_tbl)
```

Tabla 8: Varianza del efecto aleatorio (island)

term	variance
island	0

El modelo mixto (GLMM) que incluyó un intercepto aleatorio por isla no mostró variación atribuible a este factor (varianza ≈ 0) (ver Tabla 8), lo que indica que la probabilidad de presentar un pico largo no difiere significativamente entre islas.

En cuanto a los efectos fijos ver (Tabla 7) , todos resultaron estadísticamente significativos. La longitud de las aletas tuvo un efecto positivo sobre la probabilidad de presentar un pico largo: por cada milímetro adicional en la longitud del *flipper*, las probabilidades de tener un pico largo aumentan aproximadamente un 16.6% (OR = 1.17; $p < 0.001$).

Asimismo, se observaron diferencias claras entre especies. Los pingüinos *Chinstrap* mostraron una probabilidad de presentar pico largo más de 500 veces superior a la de los *Adelie* (OR = 525.6; $p < 0.001$), mientras que los *Gentoo* presentaron una probabilidad aproximadamente 7 veces mayor (OR = 7.06; $p = 0.032$).

En conjunto, estos resultados sugieren que el tamaño de las aletas es un predictor importante de la presencia de picos largos y que existen diferencias marcadas entre especies, aunque no entre islas.

9 Diagnóstico GLMM y comparación con GLM

```r
# Extract AIC and logLik for comparison
model_comp <- tibble(
  model = c("GLM", "GLMM"),
  AIC = c(AIC(modelo_glm), AIC(modelo_glmm)),
  logLik = c(logLik(modelo_glm), logLik(modelo_glmm))
) %>%
  mutate(across(where(is.numeric), ~ round(., 3)))

knitr::kable(model_comp)
```

Tabla 9: Comparación AIC/logLik entre GLM y GLMM

model	AIC	logLik
GLM	174.918	-83.459
GLMM	176.918	-83.459

```r
# ICC aproximado para GLMM (logistic)
var_island <- as.data.frame(VarCorr(modelo_glmm))$vcov[1]
icc <- round(var_island / (var_island + (pi^2 / 3)), 3)

tibble(stat = c("ICC (aprox.)"), value = c(icc)) %>% knitr::kable()
```

Tabla 10: Proporción de variación entre islas (ICC aprox.)

stat	value
ICC (aprox.)	0

La comparación entre el modelo lineal generalizado (GLM) y el modelo mixto (GLMM) mostró valores prácticamente idénticos de log-verosimilitud (-83.46), con un AIC ligeramente mayor en el GLMM (176.9) respecto al GLM (174.9) (ver Tabla 9).

Este resultado indica que la incorporación del efecto aleatorio por isla no mejora el ajuste del modelo, sino que introduce un parámetro adicional sin aportar ganancia en verosimilitud. Por tanto, el modelo más

parsimonioso y adecuado es el GLM simple, que explica la variabilidad en la probabilidad de presentar pico largo sin necesidad de considerar diferencias entre islas.

En concordancia, el coeficiente de correlación intraclase (ICC) calculado para el GLMM fue ≈ 0 (ver Tabla 10), lo que confirma que no existe una proporción relevante de la variación atribuible al nivel de agrupamiento "isla".

En resumen, tanto el análisis del AIC como el ICC sugieren que el efecto del contexto espacial (isla) es despreciable, y que la variabilidad en la presencia de picos largos está determinada principalmente por la especie y la longitud del flipper.

10 Conclusiones y recomendaciones prácticas (GLM / GLMM)

El análisis realizado con el conjunto de datos de pingüinos de Palmer permitió modelar la probabilidad de que un individuo presente un pico largo (> 45 mm) en función de variables morfométricas y taxonómicas, aplicando tanto modelos lineales generalizados (GLM) como modelos mixtos (GLMM).

Los resultados del modelo GLM binomial con enlace logit mostraron que la longitud del flipper es un predictor significativo: por cada milímetro adicional, las probabilidades de tener un pico largo aumentan aproximadamente un 16–17 % (OR ≈ 1.17). Asimismo, se detectaron diferencias claras entre especies: *Chinstrap* es la especie con mayor probabilidad de presentar picos largos, seguida por *Gentoo*, mientras que *Adelie* presenta la menor.

El modelo demostró un ajuste muy adecuado, sin indicios de sobre-dispersión (dispersion ≈ 0.49), con una exactitud de clasificación del 90 % y una curva ROC de AUC = 0.96, lo que evidencia una excelente capacidad predictiva y discriminante.

Por otro lado, el modelo mixto (GLMM) que incorporó un intercepto aleatorio por isla no mejoró sustancialmente el ajuste (AIC ≈ 176.9 frente a 174.9 del GLM) y presentó una varianza del efecto aleatorio prácticamente nula (0). El ICC ≈ 0 confirmó que las diferencias entre islas no aportan variabilidad adicional al fenómeno modelado. En consecuencia, el GLMM no ofrece ventajas sobre el GLM en este contexto, y el modelo simple resulta más parsimonioso y eficiente.