Regresion lineal, múltiple y GLM

Santos G

Tabla de contenidos

1 Contexto de proyecto

En esta sección se explora la relación entre variables morfológicas de los pingüinos, en particular entre la longitud del ala (flipper length) y la longitud del pico (bill length). El objetivo es evaluar si existe una asociación lineal entre ambas medidas, lo que permitiría inferir patrones de covariación corporal. Para ello, se aplican métodos de regresión lineal, que asumen una relación lineal y aditiva entre las variables, junto con verificaciones de supuestos estadísticos y exploración de posibles valores atípicos o influyentes.

2 Carga de librerías, dataset y limpieza de los datos

```
# Cargar librerías
library(tidyverse)    # manipulación de datos y ggplot2
library(palmerpenguins) # dataset de pingüinos
library(janitor)      # limpieza de nombres de columnas
library(broom)        # resultados ordenados de modelos
library(car)          # pruebas estadísticas (ej. Levene)

# Cargar dataset
```

```r
df_raw <- penguins %>% as_tibble() # guardo raw para auditoría
df <- df_raw %>% clean_names()
```

3 Visualización inicial + correlación

```r
# Scatter con línea de regresión (usa df ya limpio)
ggplot(df, aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, formula = y ~ x) +
  labs(x = "Flipper length (mm)", y = "Bill length (mm)",
       title = "Relación bill_length_mm ~ flipper_length_mm") +
  theme_minimal()
```

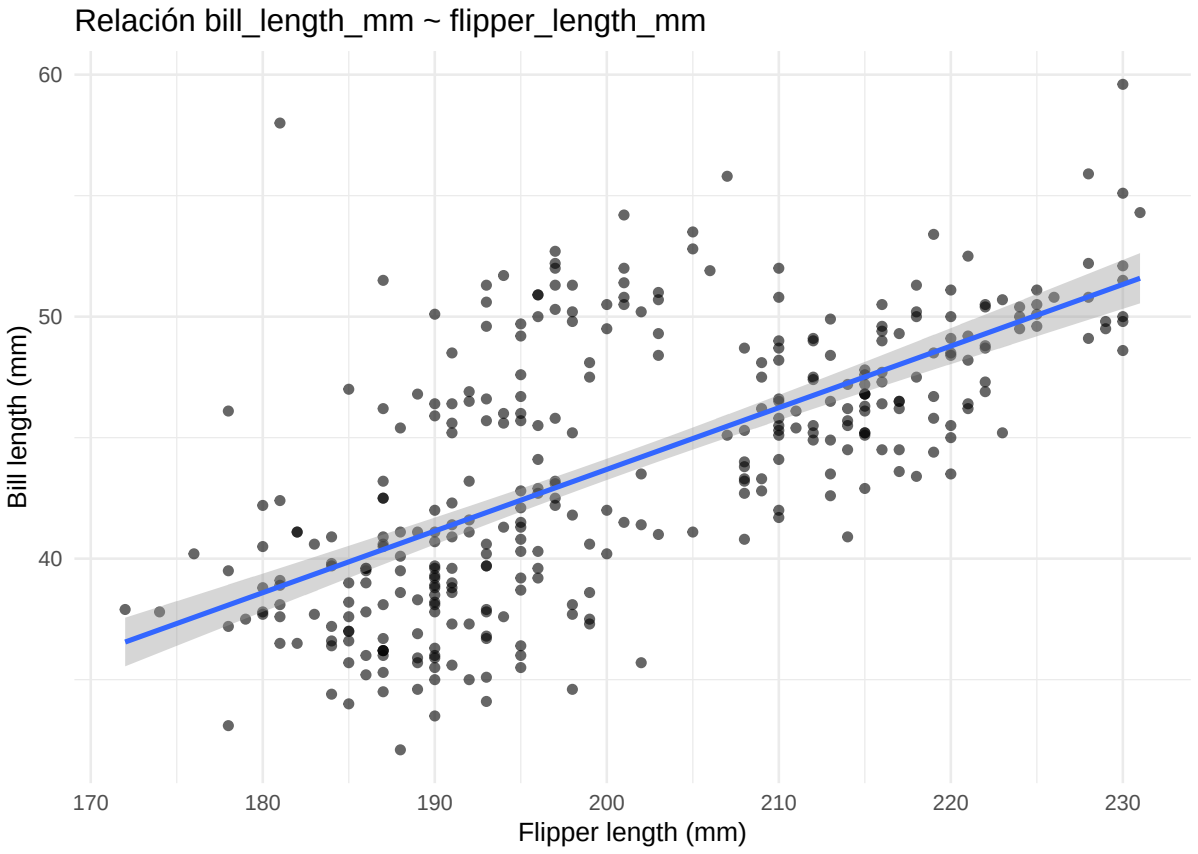

Figura 1: Scatter con línea de regresión entre las variables biil length y flipper length.

La **Figura 1** muestra una tendencia positiva clara: a mayor longitud del ala (flipper length), mayor longitud del pico (bill length). La nube de puntos es relativamente compacta, lo que sugiere una asociación consistente entre ambas variables.

```r
# Correlación Pearson y Spearman (devuelven objetos htest)
cor_pearson <- cor.test(df$flipper_length_mm, df$bill_length_mm,
                        method = "pearson")
```

```r
cor_spearman <- cor.test(df$flipper_length_mm, df$bill_length_mm,
                         method = "spearman")
# Correlación Pearson
pearson_tab <- broom::tidy(cor_pearson) %>%
  select(estimate, statistic, p.value, conf.low, conf.high) %>%
  mutate(
    p.value = ifelse(p.value < 0.001, "< 0.001", round(p.value, 3)),
    across(where(is.numeric), round, 3)
  )

knitr::kable(pearson_tab, caption = "Correlación de Pearson
             entre bill_length y flipper_length")
```

Tabla 1: Correlación de Pearson entre bill_length y flipper_length

estimate	statistic	p.value	conf.low	conf.high
0.656	16.034	< 0.001	0.591	0.713

```r
# Correlación Spearman
spearman_tab <- broom::tidy(cor_spearman) %>%
  select(estimate, statistic, p.value) %>%
  mutate(
    p.value = ifelse(p.value < 0.001, "< 0.001", round(p.value, 3)),
    across(where(is.numeric), round, 3)
  )

knitr::kable(spearman_tab, caption = "Correlación de Spearman
             entre bill_length y flipper_length")
```

Tabla 2: Correlación de Spearman entre bill_length y flipper_length

estimate	statistic	p.value
0.673	2181594	< 0.001

Las **Tablas 1** y **2** reflejan que tanto el coeficiente de Pearson ($r = 0.656$, $p < 0.001$) como el de Spearman ($\rho = 0.673$, $p < 0.001$) confirman una correlación positiva, de magnitud moderada a fuerte. Esto indica que el tamaño del pico está relacionado con el tamaño corporal de los pingüinos, lo cual es esperado en términos de allometría: individuos con alas más largas (indicador del tamaño total) tienden a presentar picos más largos.

4 Modelo lineal simple

```r
# Ajuste del modelo lineal simple
modelo_lm <- lm(bill_length_mm ~ flipper_length_mm, data = df)

```

```r
# Coeficientes y resumen del ajuste
tidy_lm <- broom::tidy(modelo_lm)
glance_lm <- broom::glance(modelo_lm)

# Tablas presentables para Quarto
knitr::kable(tidy_lm, caption = "Coeficientes del modelo lineal
             (bill_length ~ flipper_length)")
```

Tabla 3: Coeficientes del modelo lineal (bill_length ~ flipper_length)

term	estimate	std.error	statistic	p.value
(Intercept)	-7.2648678	3.2001568	-2.27016	0.0238233
flipper_length_mm	0.2547682	0.0158891	16.03410	0.0000000

```r
knitr::kable(glance_lm, caption = "Resumen del ajuste (R², AIC, BIC, etc.)")
```

Tabla 4: Resumen del ajuste (R², AIC, BIC, etc.)

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.430574	0.4288992	4.125874	257.0925	0	1	-968.983	1943.966	1955.471	5787.763	340	342

Las **Tablas 3** y **4** presentan los resultados obtenidos en el modelo lineal:

- El intercepto ((beta_0 = -7.26, p = 0.024)) representa la longitud del pico cuando la longitud del ala es cero. Este valor no tiene un significado biológico directo, pero es necesario dentro de la formulación matemática del modelo.
- El coeficiente de la longitud del ala ((beta_1 = 0.255, p < 0.001)) indica que por cada aumento de 1 mm en la longitud del ala, el pico aumenta en promedio 0.25 mm.
- El modelo explica aproximadamente un 43% de la variación en la longitud del pico ((R^2 = 0.431)), lo cual se considera un ajuste moderado en estudios biológicos.

Estos resultados sugieren una relación positiva clara entre el tamaño corporal (longitud del ala) y el tamaño del pico. En términos ecológicos, esto respalda la idea de alometría morfológica: individuos más grandes (alas más largas) tienden a tener picos más largos, lo que puede estar asociado con la necesidad de capturar presas más grandes o diversificadas. En conjunto, el modelo indica que la morfología del pico no es independiente del tamaño general del cuerpo, sino que ambas variables están estrechamente relacionadas.

5 Intervalos de confianza y predicción

```r
# IC 95% para coeficientes directamente con confint
ic_coef <- as.data.frame(confint(modelo_lm)) %>%
  tibble::rownames_to_column("term")

```

```r
knitr::kable(
  ic_coef,
  caption = "Intervalos de confianza (95%)
  para los coeficientes del modelo lineal",
  digits = 3,
  format = "markdown"
)
```

Tabla 5: Intervalos de confianza (95%) para los coeficientes del modelo lineal

term	2.5 %	97.5 %
(Intercept)	-13.559	-0.970
flipper_length_mm	0.224	0.286

En la **Tabla 5** se presentan los intervalos de confianza (95%) para los coeficientes del modelo:

- Intercepto: ([-13.56, -0.97])
- Longitud del ala ((beta_1)): ([0.224, 0.286])

Esto confirma que el efecto de la longitud del ala es positivo y estadísticamente significativo, ya que el intervalo de confianza no incluye el cero.

```r
# Predicción de la media y predicción individual, en formato tabla kable
newdata <- tibble(flipper_length_mm = c(180, 200))

pred_conf <- predict(modelo_lm, newdata, interval = "confidence", level = 0.95)
pred_pred <- predict(modelo_lm, newdata, interval = "prediction", level = 0.95)

tabla_pred <- tibble(
  flipper_length_mm = newdata$flipper_length_mm,
  fit = pred_conf[, "fit"],
  lwr_conf = pred_conf[, "lwr"],
  upr_conf = pred_conf[, "upr"],
  lwr_pred = pred_pred[, "lwr"],
  upr_pred = pred_pred[, "upr"]
)

knitr::kable(
  tabla_pred,
  caption = "Intervalos de confianza y predicción (95%)
  de la longitud del pico para valores de 180 y 200 mm de aleta",
  digits = 2,
  format = "markdown"
)
```

Tabla 6: Intervalos de confianza y predicción (95%) de la longitud del pico para valores de 180 y 200 mm de aleta

flipper_length_mm	fit	lwr_conf	upr_conf	lwr_pred	upr_pred
180	38.59	37.81	39.38	30.44	46.75
200	43.69	43.25	44.13	35.56	51.82

En la **Tabla 6** se muestran los valores predichos de la longitud del pico para longitudes de aleta de 180 mm y 200 mm:

- IC de confianza (95%): refleja la estimación del promedio poblacional esperado para pingüinos con esas longitudes de aleta.
  - Para 180 mm: [37.81, 39.38]
  - Para 200 mm: [43.25, 44.13]
- IC de predicción (95%): refleja el rango esperado para un individuo particular, por lo que son más amplios.
  - Para 180 mm: [30.44, 46.75]
  - Para 200 mm: [35.56, 51.82]

Estos intervalos muestran que, aunque el modelo estima una tendencia lineal clara (picos más largos en individuos con alas más largas), existe una variabilidad considerable a nivel individual.

En términos ecológicos, esto significa que, aunque el tamaño corporal predice el tamaño del pico en promedio, cada pingüino puede desviarse de esa tendencia debido a factores adicionales como edad, sexo, o adaptaciones específicas relacionadas con la dieta y el hábitat.

6 Diagnósticos de supuestos

```r
# Residuos y fitted
residuales <- residuals(modelo_lm)
fittedv    <- fitted(modelo_lm)
n <- nrow(df)
k <- length(coef(modelo_lm)) - 1  # número de predictores


df_diag <- tibble(fitted = fittedv, resid = residuales)
ggplot(df_diag, aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuales vs Fitted", x = "Valores ajustados",
       y = "Residuales")
```

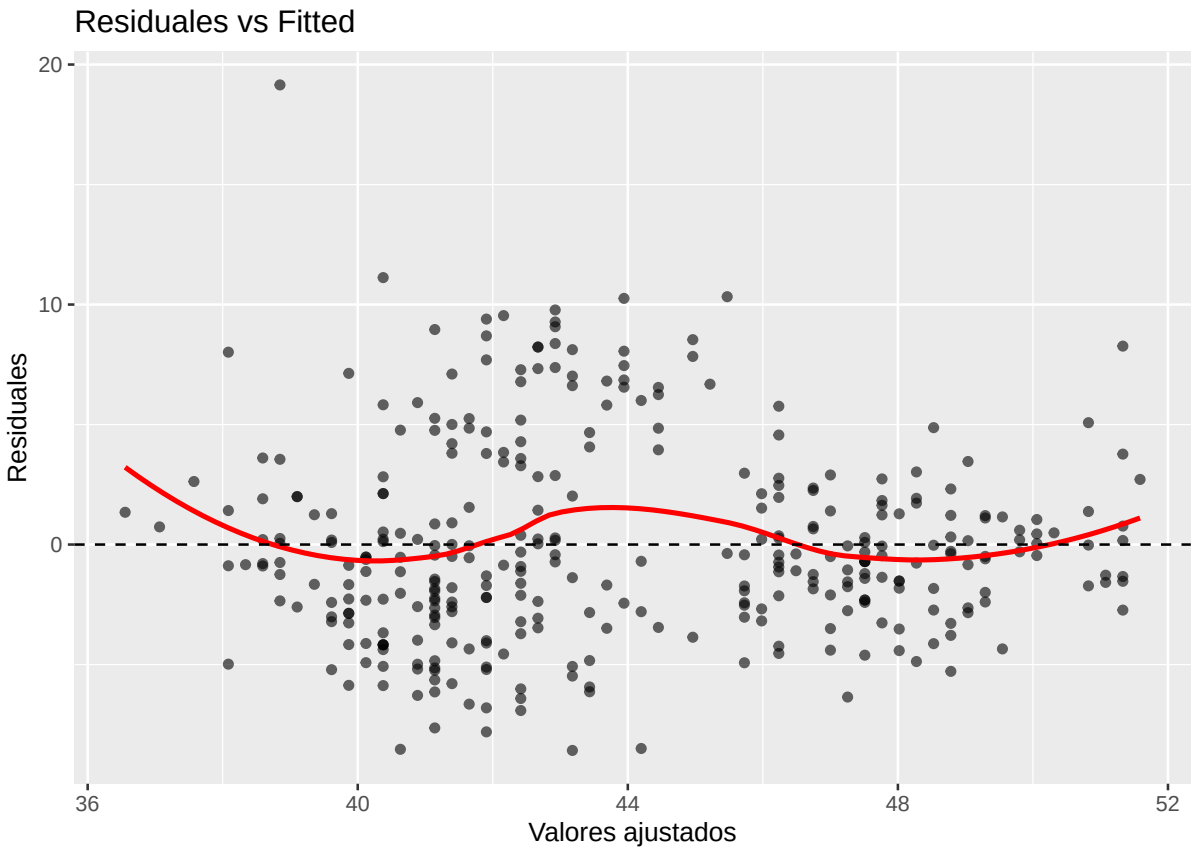

Figura 2: Gráfico de dispersión entre los residuales y los valores ajustados.

La **Figura 2** muestra que los residuos se distribuyen en torno a la línea horizontal de cero, aunque se observa cierta curvatura y dispersión desigual en algunos tramos. Esto indica que la relación entre las variables no es perfectamente lineal y que podría existir cierta heterocedasticidad (varianza no constante de los errores). Sin embargo, no se aprecian patrones extremos que invaliden el modelo de forma inmediata.

```
#QQ-plot (visual)
ggplot(tibble(resid = residuales), aes(sample = resid)) +
  stat_qq() + stat_qq_line() + labs(title = "QQ-plot de residuos")
```

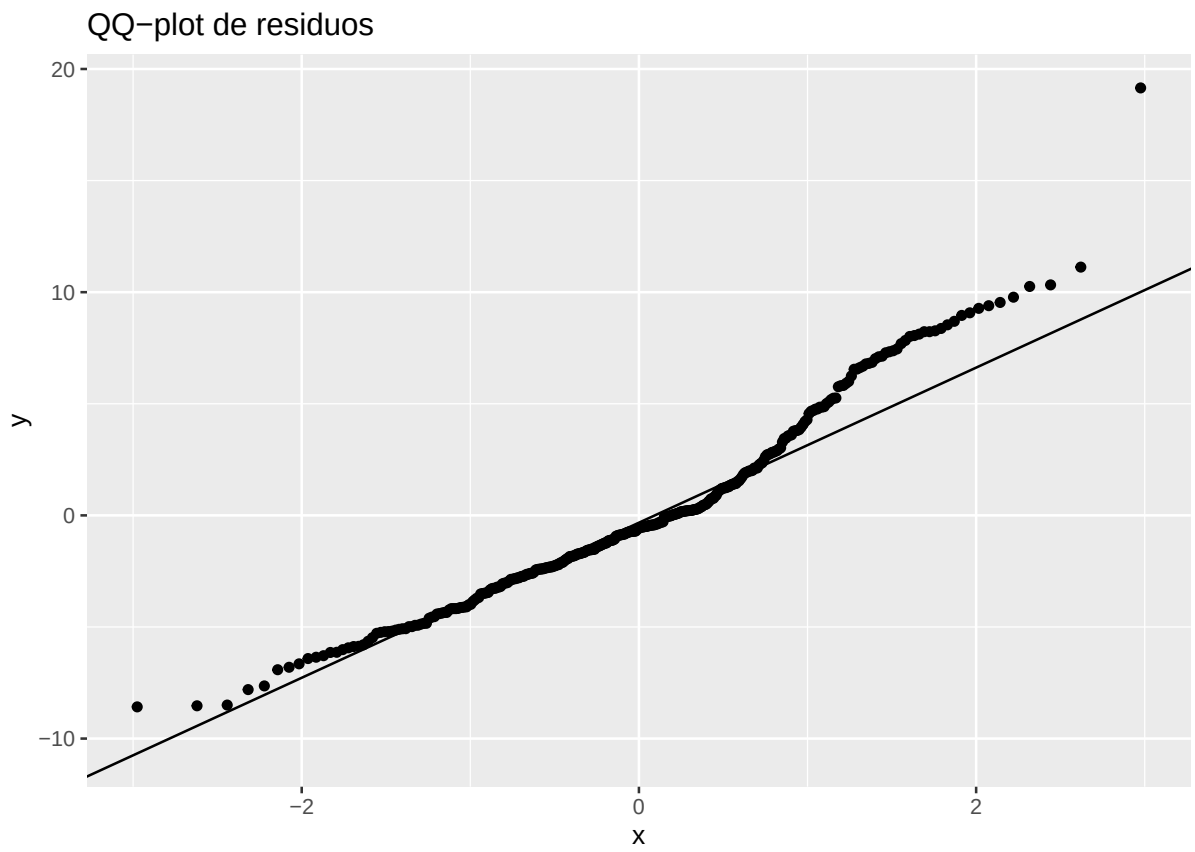

Figura 3: Gráfico QQ-plot residuos.

La **Figura 3** refleja que la mayor parte de los puntos sigue la línea de referencia, lo que sugiere aproximación a la normalidad. No obstante, en las colas se observan desviaciones claras, que rechazarían la hipótesis de normalidad. En la práctica, los modelos lineales son relativamente robustos a esta violación cuando el tamaño de muestra es grande, como en este caso.

```
# Tests de supuestos

# Shapiro-Wilk (normalidad)
shapiro_res <- broom::tidy(shapiro.test(residuales)) %>%
  mutate(test = "Shapiro-Wilk")

# Breusch-Pagan (heterocedasticidad)
bptest_res <- broom::tidy(lmtest::bptest(modelo_lm)) %>%
  mutate(test = "Breusch-Pagan")

# Durbin-Watson (autocorrelación de residuos)
dw_res <- broom::tidy(lmtest::dwtest(modelo_lm)) %>%
  mutate(test = "Durbin-Watson")

# Juntar en tabla y ajustar p-values
tests_table <- bind_rows(shapiro_res, bptest_res, dw_res) %>%
  select(test, statistic, p.value, method) %>%
  mutate(
    p.value = case_when(
```

```
      p.value < 0.001 ~ "< 0.001",
      TRUE ~ as.character(round(p.value, 3))
    ),
    statistic = round(statistic, 3)
  )

knitr::kable(tests_table, caption = "Resultados de los tests
             de supuestos del modelo lineal")
```

Tabla 7: Resultados de los tests de supuestos del modelo lineal

test	statistic	p.value	method
Shapiro-Wilk	0.958	< 0.001	Shapiro-Wilk normality test
Breusch-Pagan	11.043	< 0.001	studentized Breusch-Pagan test
Durbin-Watson	0.939	< 0.001	Durbin-Watson test

La **Tabla 7** resume los resultados de los tests de supuestos aplicados al modelo lineal. El test de Shapiro–Wilk indica desviaciones de la normalidad en los residuos ($p < 0.001$). El test de Breusch–Pagan es significativo ($p < 0.001$), lo que evidencia heterocedasticidad, es decir, que la varianza de los errores no es constante. Finalmente, el test de Durbin–Watson ($DW = 0.939$, $p < 0.001$) muestra autocorrelación positiva de los residuos, lo que sugiere dependencia entre las observaciones, posiblemente por la estructura de especie o colonia.

```
# Calcular medidas de influencia
cooks <- cooks.distance(modelo_lm)
hatv  <- hatvalues(modelo_lm)
rstudent_vals <- rstudent(modelo_lm)

# Crear tabla con observaciones influyentes
influential <- tibble(
  index = seq_along(cooks),
  cooks = cooks,
  hat = hatv,
  rstudent = rstudent_vals
) %>%
  mutate(
    cooks_flag = cooks > (4/length(cooks)),
    rstudent_flag = abs(rstudent) > 3,
    hat_flag = hat > (2*(k+1)/n)
  ) %>%
  filter(cooks_flag | rstudent_flag | hat_flag) %>%
  arrange(desc(cooks)) %>%
  mutate(across(where(is.numeric), round, 3))

knitr::kable(influential, caption = "Observaciones potencialmente
             influyentes (Cook's distance, leverage, residuos studentizados)")
```

Tabla 8: Observaciones potencialmente influyentes (Cook's distance, leverage, residuos studentizados)

index	cooks	hat	rstudent	cooks_flag	rstudent_flag	hat_flag
292	0.097	0.009	4.812	TRUE	TRUE	FALSE
185	0.032	0.015	2.029	TRUE	FALSE	TRUE
323	0.021	0.006	2.729	TRUE	FALSE	FALSE
281	0.021	0.011	1.962	TRUE	FALSE	FALSE
253	0.011	0.014	1.240	FALSE	FALSE	TRUE
267	0.007	0.015	0.920	FALSE	FALSE	TRUE
215	0.004	0.016	0.663	FALSE	FALSE	TRUE
227	0.003	0.015	-0.667	FALSE	FALSE	TRUE
122	0.003	0.012	0.640	FALSE	FALSE	TRUE
255	0.001	0.014	-0.420	FALSE	FALSE	TRUE
219	0.001	0.015	-0.385	FALSE	FALSE	TRUE
217	0.001	0.015	-0.374	FALSE	FALSE	TRUE
28	0.001	0.015	0.328	FALSE	FALSE	TRUE
153	0.001	0.015	-0.325	FALSE	FALSE	TRUE
243	0.001	0.014	0.336	FALSE	FALSE	TRUE
263	0.001	0.015	-0.311	FALSE	FALSE	TRUE
241	0.000	0.015	0.187	FALSE	FALSE	TRUE
20	0.000	0.014	0.179	FALSE	FALSE	TRUE
247	0.000	0.012	0.119	FALSE	FALSE	TRUE
265	0.000	0.015	0.041	FALSE	FALSE	TRUE
237	0.000	0.014	-0.005	FALSE	FALSE	TRUE

La **Tabla 8** presenta las observaciones potencialmente influyentes detectadas mediante Cook's distance, leverage y residuos studentizados. Destaca la observación 292, con un residuo estudentizado muy alto (4.81) y una distancia de Cook por encima del umbral (0.097), lo que indica que afecta fuertemente los parámetros estimados. Las observaciones 185 y 323 también muestran valores de Cook elevados y, en el caso de 185, un leverage alto, lo que refleja un peso excesivo en el ajuste. El resto de observaciones (e.g., 281, 253, 267) presentan leverage relativamente alto, pero con menor impacto individual en el modelo.

Desde una perspectiva biológica, estas observaciones pueden corresponder a individuos atípicos, errores de medición o variabilidad natural de las poblaciones, y su tratamiento debe basarse en criterios ecológicos además de estadísticos.

7 Outliers / puntos influyentes

```
# Crear modelo sin observaciones influyentes y comparar coeficientes
in_idx <- influential$index

if(length(in_idx) > 0){
  modelo_lm_noinf <- lm(bill_length_mm ~ flipper_length_mm,
                        data = df[-in_idx, ])

  compare_coefs <- tibble(
    original = broom::tidy(modelo_lm)$estimate,
    no_influ = broom::tidy(modelo_lm_noinf)$estimate
```

```
  )

  knitr::kable(
    compare_coefs %>%
      mutate(across(where(is.numeric), round, 3)) %>%
      rename("Modelo original" = original,
             "Modelo sin influyentes" = no_influ),
    caption = "Comparación de coeficientes estimados con
    y sin observaciones influyentes."
  )
} else {
  "No se detectaron observaciones influyentes con los umbrales establecidos."
}
```

Tabla 9: Comparación de coeficientes estimados con y sin observaciones influyentes.

Modelo original	Modelo sin influyentes
-7.265	-9.128
0.255	0.263

La **Tabla 9** muestra los coeficientes estimados para el modelo original y para el modelo ajustado sin las observaciones influyentes (n = 21). Se observa que el intercepto cambia de –7.26 a –9.13, y la pendiente de 0.255 a 0.263. Aunque los cambios son moderados, evidencian que las observaciones influyentes, especialmente la 292, tienen un impacto en la magnitud de los parámetros.

8 Conclusiones generales

El modelo lineal confirma una relación positiva fuerte entre el largo del ala y el largo del pico en pingüinos, aunque presenta violaciones a varios supuestos:

- Normalidad: los residuos no siguen una distribución normal (Shapiro-Wilk significativo).
- Homoscedasticidad: la varianza de los errores no es constante (Breusch-Pagan significativo).
- Independencia: los residuos presentan autocorrelación positiva (Durbin-Watson).
- Influencia de outliers: algunas observaciones (p. ej., la 292) afectan de manera importante el ajuste.

Estos resultados sugieren que, si bien el modelo lineal simple ofrece información valiosa, es necesario avanzar hacia enfoques más robustos:

- Evaluar modelos que incluyan variables adicionales (como especie o sexo).
- Considerar modelos lineales generalizados o mixtos que controlen por estructura de datos y agrupamiento.
- Explorar la posibilidad de transformaciones o métodos robustos frente a outliers.
- En síntesis, la relación biológica entre largo de ala y pico es clara y significativa, pero el modelo lineal simple debe interpretarse con cautela debido a las violaciones de supuestos y la influencia de casos extremos.