

Regresiones lineales simples y múltiples

Santos G

Tabla de contenidos

1	Contexto de proyecto	1
2	Carga de librerías, dataset y limpieza de los datos	1
3	Visualización inicial + correlación	2
4	Regresión lineal simple	3
5	Diagnósticos del modelo lineal	6
6	Conclusiones del modelo lineal	11
7	Regresión lineal múltiple	12
8	Diagnósticos del modelo múltiple	14
9	Conclusiones del modelo múltiple	16

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

```
1 # Cargar librerías
2 library(tidyverse)    # manipulación de datos y ggplot2
3 library(palmerpenguins) # dataset de pingüinos
4 library(janitor)      # limpieza de nombres de columnas
5 library(broom)        # resultados ordenados de modelos
6 library(car)          # pruebas estadísticas (ej. Levene)
7
```

```

8 # Cargar dataset
9 df_raw <- penguins %>% as_tibble() # guardo raw para auditoría
10 df <- df_raw %>% clean_names()

```

3 Visualización inicial + correlación

```

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

```

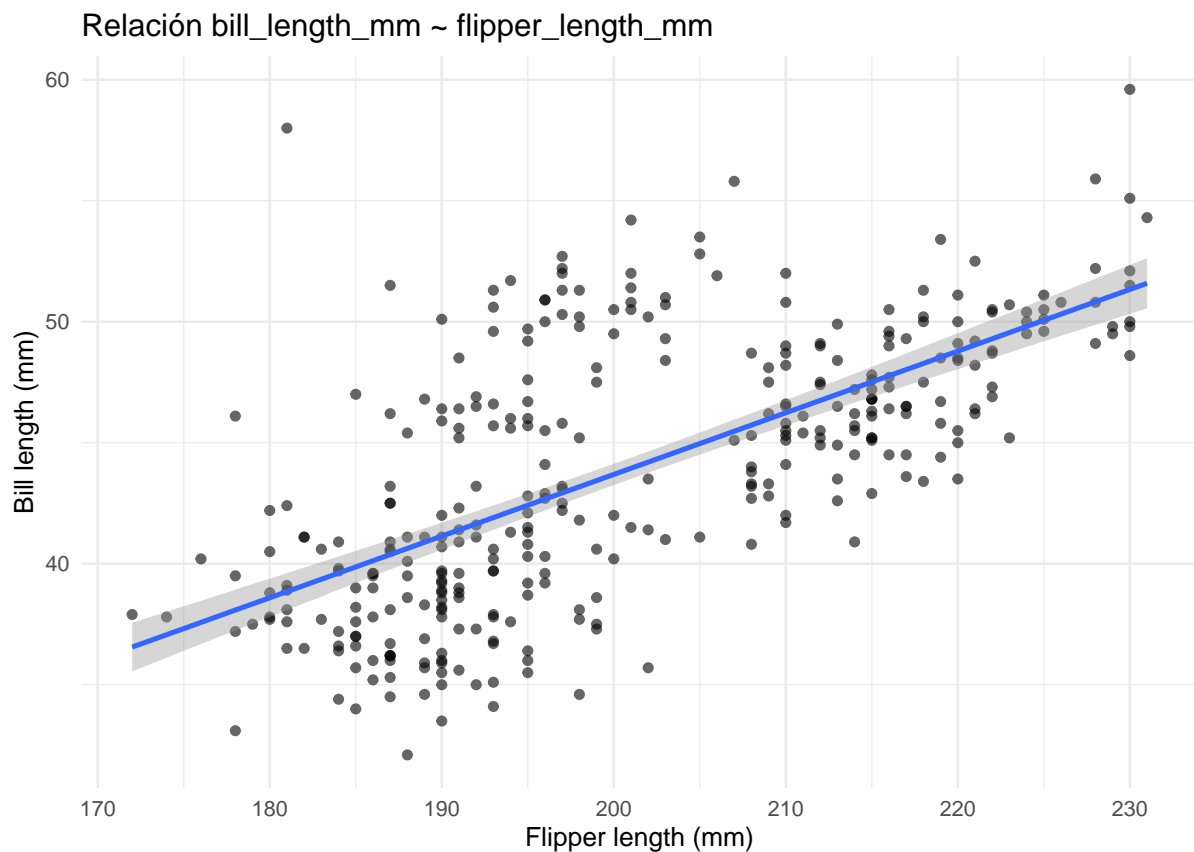


Figura 1: Scatter con línea de regresión entre las variables bill length y flipper length.

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.

```

1 # Correlación Pearson y Spearman (devuelven objetos htest)
2 cor_pearson <- cor.test(df$flipper_length_mm, df$bill_length_mm,

```

```

3         method = "pearson")
4 cor_spearman <- cor.test(df$flipper_length_mm, df$bill_length_mm,
5         method = "spearman")
6 # Correlación Pearson
7 pearson_tab <- broom::tidy(cor_pearson) %>%
8   select(estimate, statistic, p.value, conf.low, conf.high) %>%
9   mutate(
10     p.value = ifelse(p.value < 0.001, "< 0.001", round(p.value, 3)),
11     across(where(is.numeric), round, 3)
12   )
13
14 knitr::kable(pearson_tab, caption = "Correlación de Pearson
15   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

```

1 # Correlación Spearman
2 spearman_tab <- broom::tidy(cor_spearman) %>%
3   select(estimate, statistic, p.value) %>%
4   mutate(
5     p.value = ifelse(p.value < 0.001, "< 0.001", round(p.value, 3)),
6     across(where(is.numeric), round, 3)
7   )
8
9 knitr::kable(spearman_tab, caption = "Correlación de Spearman
10   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 alometría: individuos con alas más largas (indicador del tamaño total) tienden a presentar picos más largos.

4 Regresión lineal simple

```

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

```

```

3
4 # Coeficientes y resumen del ajuste
5 tidy_lm <- broom::tidy(modelo_lm)
6 glance_lm <- broom::glance(modelo_lm)
7
8 # Tablas presentables para Quarto
9 knitr::kable(tidy_lm, caption = "Coeficientes del modelo lineal
10               (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

```

1 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	15787.763	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 ((R² = 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.

```

1 # IC 95% para coeficientes directamente con confint
2 ic_coef <- as.data.frame(confint(modelo_lm)) %>%
3   tibble::rownames_to_column("term")
4
5 knitr::kable(
6   ic_coef,
7   caption = "Intervalos de confianza (95%)")

```

```

8   para los coeficientes del modelo lineal",
9   digits = 3,
10  format = "markdown"
11 )

```

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.

```

1  # Predicción de la media y predicción individual, en formato tabla kable
2  newdata <- tibble(flipper_length_mm = c(180, 200))
3
4  pred_conf <- predict(modelo_lm, newdata, interval = "confidence", level = 0.95)
5  pred_pred <- predict(modelo_lm, newdata, interval = "prediction", level = 0.95)
6
7  tabla_pred <- tibble(
8    flipper_length_mm = newdata$flipper_length_mm,
9    fit = pred_conf[, "fit"],
10   lwr_conf = pred_conf[, "lwr"],
11   upr_conf = pred_conf[, "upr"],
12   lwr_pred = pred_pred[, "lwr"],
13   upr_pred = pred_pred[, "upr"]
14 )
15
16 knitr::kable(
17   tabla_pred,
18   caption = "Intervalos de confianza y predicción (95%)
19   de la longitud del pico para valores de 180 y 200 mm de aleta",
20   digits = 2,
21   format = "markdown"
22 )

```

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.

5 Diagnósticos del modelo lineal

```
1 # Residuos y fitted
2 residuales <- residuals(modelo_lm)
3 fittedv    <- fitted(modelo_lm)
4 n <- nrow(df)
5 k <- length(coef(modelo_lm)) - 1 # número de predictores
6
7
8 df_diag <- tibble(fitted = fittedv, resid = residuales)
9 ggplot(df_diag, aes(x = fitted, y = resid)) +
10   geom_point(alpha = 0.6) +
11   geom_smooth(method = "loess", se = FALSE, color = "red") +
12   geom_hline(yintercept = 0, linetype = "dashed") +
13   labs(title = "Residuales vs Fitted", x = "Valores ajustados",
14         y = "Residuales")
```

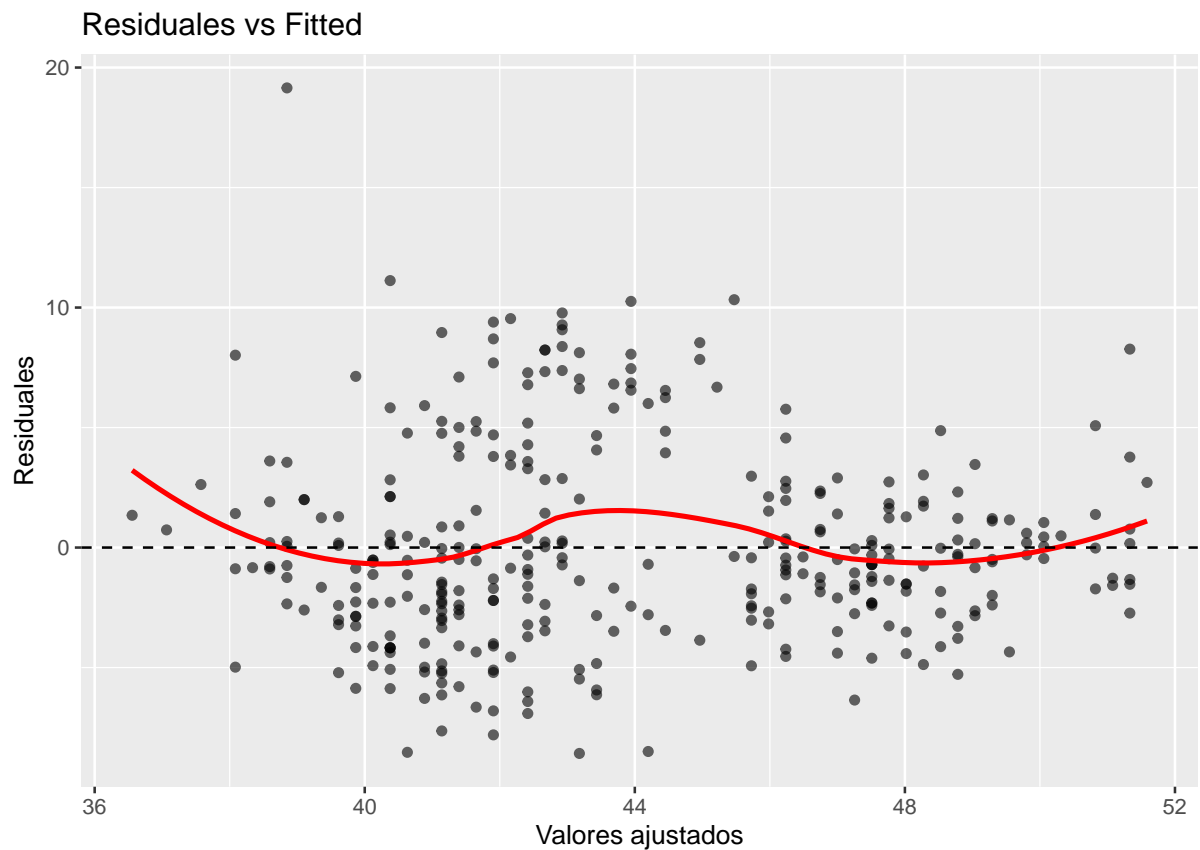


Figura 2: Gráfico de dispersión entre los residuales y los valores ajustados.

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.

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

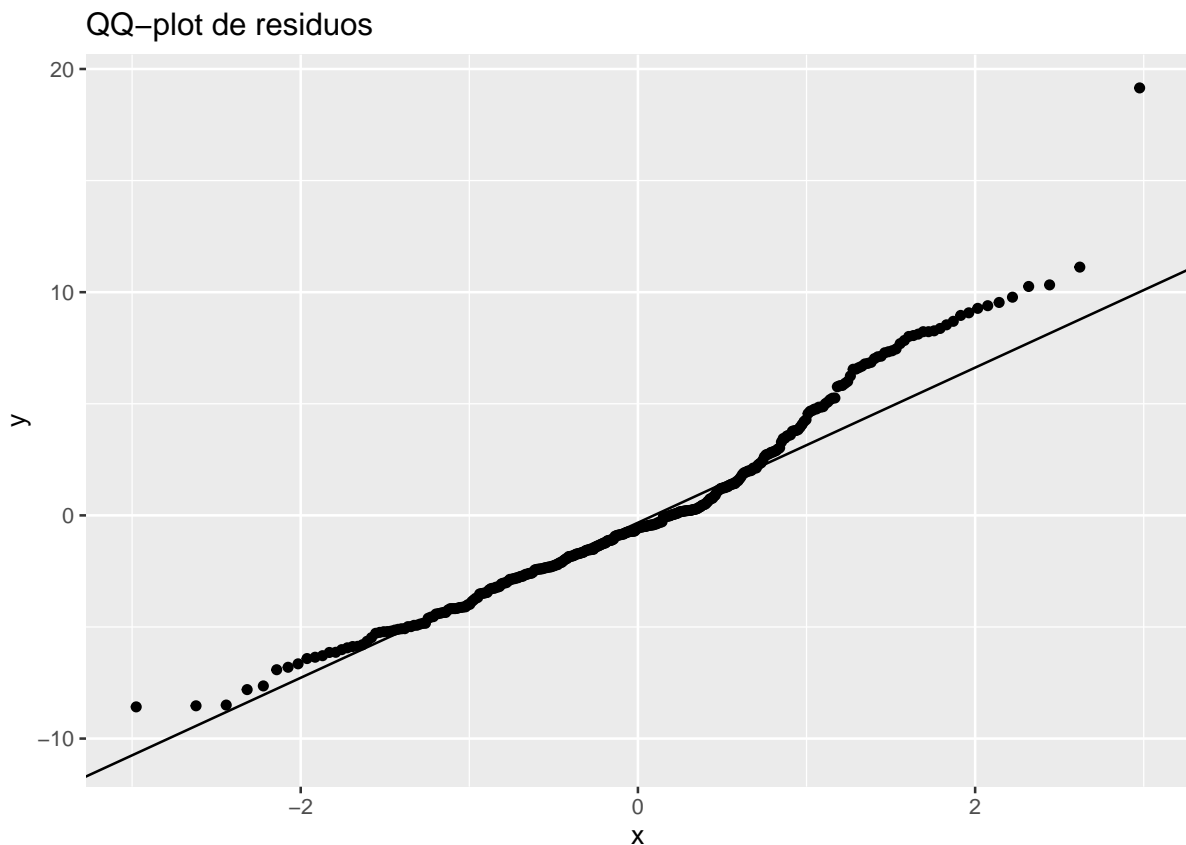


Figura 3: Gráfico QQ-plot residuos.

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.

```

1 # Tests de supuestos
2
3 # Shapiro-Wilk (normalidad)
4 shapiro_res <- broom::tidy(shapiro.test(residuales)) %>%
5   mutate(test = "Shapiro-Wilk")
6
7 # Breusch-Pagan (heterocedasticidad)
8 bptest_res <- broom::tidy(lmtest::bptest(modelo_lm)) %>%
9   mutate(test = "Breusch-Pagan")
10
11 # Durbin-Watson (autocorrelación de residuos)
12 dw_res <- broom::tidy(lmtest::dwtest(modelo_lm)) %>%
13   mutate(test = "Durbin-Watson")
14
15 # Juntar en tabla y ajustar p-values
16 tests_table <- bind_rows(shapiro_res, bptest_res, dw_res) %>%
17   select(test, statistic, p.value, method) %>%
18   mutate(
19     p.value = case_when(

```



```

20   p.value < 0.001 ~ "< 0.001",
21   TRUE ~ as.character(round(p.value, 3))
22 ),
23   statistic = round(statistic, 3)
24 )
25
26 knitr::kable(tests_table, caption = "Resultados de los tests
27   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.

```

1  # Calcular medidas de influencia
2  cooks <- cooks.distance(modelo_lm)
3  hatv <- hatvalues(modelo_lm)
4  rstudent_vals <- rstudent(modelo_lm)
5
6  # Crear tabla con observaciones influyentes
7  influential <- tibble(
8    index = seq_along(cooks),
9    cooks = cooks,
10   hat = hatv,
11   rstudent = rstudent_vals
12 ) %>%
13   mutate(
14     cooks_flag = cooks > (4/length(cooks)),
15     rstudent_flag = abs(rstudent) > 3,
16     hat_flag = hat > (2*(k+1)/n)
17 ) %>%
18   filter(cooks_flag | rstudent_flag | hat_flag) %>%
19   arrange(desc(cooks)) %>%
20   mutate(across(where(is.numeric), round, 3))
21
22 knitr::kable(influential, caption = "Observaciones potencialmente
23   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.

```

1 # Crear modelo sin observaciones influyentes y comparar coeficientes
2 in_idx <- influential$index
3
4 if(length(in_idx) > 0){
5   modelo_lm_noinf <- lm(bill_length_mm ~ flipper_length_mm,
6                         data = df[-in_idx, ])
7
8   compare_coefs <- tibble(
9     original = broom::tidy(modelo_lm)$estimate,
10    no_influ = broom::tidy(modelo_lm_noinf)$estimate
11  )
12
13  knitr::kable(
14    compare_coefs %>%
15    mutate(across(where(is.numeric), round, 3)) %>%

```

```

16     rename("Modelo original" = original,
17            "Modelo sin influyentes" = no_influ),
18     caption = "Comparación de coeficientes estimados con
19     y sin observaciones influyentes."
20 )
21 } else {
22     "No se detectaron observaciones influyentes con los umbrales establecidos."
23 }

```

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.

6 Conclusiones del modelo lineal

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.

7 Regresión lineal múltiple

En vez de usar solo `flipper_length_mm` como predictor del `bill_length_mm`, ahora incluiremos la especie como variable categórica. Esto nos permite responder:

- ¿Sigue siendo significativa la relación entre longitud de ala y pico, una vez controlamos por especie?
- ¿Existen diferencias en la longitud del pico atribuibles directamente a la especie?

```
1 # Modelo de regresión múltiple: bill_length ~ flipper_length + species
2 modelo_lm_multi <- lm(bill_length_mm ~ flipper_length_mm + species, data = df)
3
4 # Resumen de coeficientes
5 tabla_multi <- broom::tidy(modelo_lm_multi, conf.int = TRUE)
6
7 knitr::kable(
8   tabla_multi %>%
9     mutate(across(where(is.numeric), round, 3)),
10   caption = "Coeficientes estimados del modelo lineal múltiple
11   (longitud del pico como variable respuesta)."
```

Tabla 10: Coeficientes estimados del modelo lineal múltiple (longitud del pico como variable respuesta).

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.059	4.039	-0.510	0.611	-10.002	5.885
flipper_length_mm	0.215	0.021	10.129	0.000	0.173	0.257
speciesChinstrap	8.780	0.399	21.998	0.000	7.995	9.565
speciesGentoo	2.857	0.659	4.338	0.000	1.561	4.152

La **Tabla 10** muestra los coeficientes del modelo múltiple que relaciona la longitud del pico (`bill_length_mm`) con la longitud del ala (`flipper_length_mm`) y la especie.

- El intercepto ($\beta_0 = -2.06$, $p = 0.611$) no es significativo y representa la longitud promedio del pico para pingüinos *Adelie* (categoría de referencia) con longitud de ala igual a cero. Este valor carece de interpretación biológica directa, pero es necesario para el modelo.
- El coeficiente de `flipper_length_mm` ($\beta_1 = 0.215$, $p < 0.001$) indica que, controlando por especie, un aumento de 1 mm en la longitud del ala se asocia con un incremento promedio de 0.215 mm en la longitud del pico. Esto confirma una relación alométrica positiva entre el tamaño corporal y la morfología del pico dentro de cada especie.
- La especie *Chinstrap* presenta un pico en promedio 8.78 mm más largo que los *Adelie* ($p < 0.001$), mientras que los *Gentoo* presentan picos 2.86 mm más largos que los *Adelie* ($p < 0.001$), ajustando por longitud del ala. Estas diferencias son significativas y reflejan divergencias morfológicas marcadas entre especies.

```

1 # Métricas generales del modelo
2 tabla_multi_glance <- broom::glance(modelo_lm_multi)
3
4 knitr::kable(
5   tabla_multi_glance %>%
6     mutate(across(where(is.numeric), round, 3)),
7   caption = "Estadísticos de ajuste del modelo múltiple."
8 )

```

Tabla 11: Estadísticos de ajuste del modelo múltiple.

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.776	0.774	2.596	389.966	0	3	- 809.56	1629.119	1648.294	2278.337	338	342

La **Tabla 11** muestra los estadísticos globales del ajuste. El modelo explica un 77.6% de la variabilidad en la longitud del pico ($R^2 = 0.776$), lo cual representa una mejora sustancial respecto al modelo simple ($R^2 \approx 0.43$). El error residual ($\sigma = 2.60$ mm) es considerablemente menor que en el modelo simple, indicando un mejor ajuste.

En conjunto, el modelo múltiple sugiere que tanto el tamaño corporal (longitud de ala) como la especie son predictores clave de la longitud del pico. Desde un punto de vista ecológico, esto refleja que la morfología del pico no solo está asociada al tamaño general del cuerpo, sino que también depende de adaptaciones específicas entre especies, posiblemente vinculadas a diferencias en dieta, ecología trófica o hábitat.

```

1 # Comparar modelo simple vs múltiple
2 modelo_simple <- lm(bill_length_mm ~ flipper_length_mm, data = df)
3 modelo_multiple <- lm(bill_length_mm ~ flipper_length_mm + species, data = df)
4
5 # Extraer métricas de ajuste
6 comp_models <- bind_rows(
7   broom::glance(modelo_simple) %>% mutate(modelo = "Simple"),
8   broom::glance(modelo_multiple) %>% mutate(modelo = "Múltiple")
9 ) %>%
10   select(modelo, r.squared, adj.r.squared, AIC, BIC, sigma, p.value) %>%
11   mutate(across(where(is.numeric), ~round(., 3)))
12
13 knitr::kable(comp_models,
14   caption = "Comparación entre el modelo
15   lineal simple y múltiple")

```

Tabla 12: Comparación entre el modelo lineal simple y múltiple

modelo	r.squared	adj.r.squared	AIC	BIC	sigma	p.value
Simple	0.431	0.429	1943.966	1955.471	4.126	0
Múltiple	0.776	0.774	1629.119	1648.294	2.596	0

La **Tabla 12** compara el modelo lineal simple y el múltiple. El modelo simple (solo con la longitud del ala) explica alrededor del 43% de la variación en la longitud del pico ($R^2 = 0.431$), mientras que el modelo múltiple (incluyendo la especie) aumenta notablemente la capacidad explicativa hasta un 77% ($R^2 = 0.776$).

Además, los criterios de información (AIC y BIC) son considerablemente menores en el modelo múltiple, lo que indica un mejor ajuste penalizado por complejidad. En conjunto, la comparación sugiere que la variable especie aporta información fundamental para explicar las diferencias en la longitud del pico, más allá del tamaño corporal.

8 Diagnósticos del modelo múltiple

```

1 # Residuos del modelo múltiple
2 residuales_m <- resid(modelo_multiple)
3
4 # Pruebas de supuestos
5 shapiro_res_m <- shapiro.test(residuales_m)
6 bptest_res_m <- lmtest::bptest(modelo_multiple)
7 dw_res_m <- lmtest::dwtest(modelo_multiple)
8
9 # Resumen en tabla
10 tests_table_m <- tibble(
11   Test = c("Shapiro-Wilk (normalidad)",
12            "Breusch-Pagan (homocedasticidad)",
13            "Durbin-Watson (independencia)"),
14   Estadístico = c(round(shapiro_res_m$statistic, 3),
15                   round(bptest_res_m$statistic, 3),
16                   round(dw_res_m$statistic, 3)),
17   p_value = c(ifelse(shapiro_res_m$p.value < 0.001, "<0.001",
18                       round(shapiro_res_m$p.value, 3)),
19               ifelse(bptest_res_m$p.value < 0.001, "<0.001",
20                       round(bptest_res_m$p.value, 3)),
21               ifelse(dw_res_m$p.value < 0.001, "<0.001",
22                       round(dw_res_m$p.value, 3)))
23 )
24
25 knitr::kable(tests_table_m,
26               caption = "Resultados de las pruebas de
27               supuestos del modelo múltiple")

```

Tabla 13: Resultados de las pruebas de supuestos del modelo múltiple

Test	Estadístico	p_value
Shapiro-Wilk (normalidad)	0.985	0.001
Breusch-Pagan (homocedasticidad)	2.923	0.404
Durbin-Watson (independencia)	2.450	1.000

La **Tabla 13** presenta las pruebas de supuestos del modelo lineal múltiple. El test de Shapiro-Wilk indica un valor $p = 0.001$, lo que sugiere que los residuos se desvían de una distribución normal perfecta. Sin

embargo, dado el tamaño de la muestra, pequeñas desviaciones pueden detectarse como significativas sin afectar sustancialmente la validez del modelo. El test de Breusch-Pagan ($p = 0.404$) no rechaza la hipótesis de homocedasticidad, por lo que no hay evidencia de heterocedasticidad. Por su parte, el test de Durbin-Watson muestra un valor cercano a 2 ($p = 1.000$), lo que respalda la ausencia de autocorrelación en los residuos. En conjunto, estos resultados sugieren que los supuestos principales de homogeneidad e independencia se cumplen, mientras que la normalidad de los residuos presenta una ligera desviación.

```

1 # Influencia en el modelo múltiple
2 cooks_m <- cooks.distance(modelo_multiple)
3 hatv_m <- hatvalues(modelo_multiple)
4 rstud_m <- rstudent(modelo_multiple)
5
6 influential_m <- tibble(
7   index = seq_along(cooks_m),
8   cooks = cooks_m,
9   leverage = hatv_m,
10  rstudent = rstud_m
11 ) %>%
12   filter(cooks > 4/length(cooks_m) | abs(rstudent) > 3) %>%
13   mutate(across(where(is.numeric), round, 3))
14
15 knitr::kable(influential_m,
16               caption = "Observaciones potencialmente
17               influyentes en el modelo múltiple")

```

Tabla 14: Observaciones potencialmente influyentes en el modelo múltiple

index	cooks	leverage	rstudent
14	0.015	0.011	-2.308
37	0.016	0.013	2.163
90	0.020	0.016	-2.219
122	0.015	0.020	1.720
185	0.064	0.019	3.700
215	0.012	0.021	1.491
250	0.013	0.022	1.540
253	0.023	0.016	2.373
267	0.017	0.019	1.889
292	0.177	0.029	4.998
295	0.012	0.029	-1.270
305	0.028	0.020	-2.364
307	0.015	0.020	-1.731
322	0.014	0.032	-1.298
323	0.016	0.020	1.781
329	0.015	0.020	-1.731
338	0.019	0.023	1.784
339	0.029	0.017	-2.611

La **Tabla 14** muestra las observaciones potencialmente influyentes según Cook's distance, leverage y residuos studentizados. En particular, los casos 185 y 292 destacan por presentar residuos estandarizados mayores a $|3|$, junto con valores elevados de influencia. Esto significa que dichas observaciones pueden

ejercer un efecto desproporcionado sobre los coeficientes estimados del modelo. Desde una perspectiva biológica, estos valores atípicos pueden reflejar errores de medición, individuos inusuales o variabilidad real dentro de las especies analizadas. Por tanto, más que eliminarlos automáticamente, conviene evaluarlos en función del contexto ecológico y de la pregunta de investigación.

9 Conclusiones del modelo múltiple

El modelo lineal múltiple, que incluye la longitud de la aleta y la especie como predictores de la longitud del pico, explica de manera sustancial la variabilidad en los datos (R^2 ajustado ≈ 0.77). Esto indica que más de tres cuartas partes de la variación en la longitud del pico se pueden atribuir a estas variables, lo que refuerza la relevancia de la especie y la morfología corporal en la determinación del tamaño del pico.

En términos de supuestos, el modelo cumple adecuadamente con la homocedasticidad y la independencia de los residuos, aunque se detecta una ligera desviación de la normalidad. Asimismo, se identificaron observaciones influyentes (e.g., casos 185 y 292), lo que implica que los resultados deben interpretarse con cautela, considerando el contexto biológico antes de excluir dichos individuos.

En conjunto, el modelo múltiple representa una mejora clara respecto al modelo simple, al capturar mejor la estructura de los datos y aportar evidencia de la importancia de la variabilidad entre especies.