Tercera Evaluación

Modelos estadísticos lineales

Garcia Rios Santiago

Víctor Jesús Martínez Contreras

Alejandra Castillo García

2024-04-28

Table of contents

## 1 Tilapias

En un ejido en Veracruz un grupo de familias ha montado un sistema de crianza de tilapias y desean ***conocer el efecto de la densidad de peces en el encierro y de la estación del año en el crecimiento*** de los individuos. Para probar el efecto de estos factores, realizan un experimento en encierros en donde colocan 10, 18 o 24 individuos (niveles dentro del factor densidad). Pesan 9 peces antes y después del experimento (marcándolos para recapturar al mismo pez) y registran el incremento en peso al cabo de dos semanas. Realizan estas mediciones en verano y en primavera (factor estación del año). ¿Cuál es el efecto de la época del año y de la densidad de peces en el encierro en el crecimiento de las tilapias? Utiliza todos los recursos vistos en clase para presentar los resultados de este ejercicio. NOTA: Las mediciones en cada estación vienen de peces distintos.

1. Primero, cambiamos a factores los datos y exploramos nuestro dataframe:

Data frame: tilapias

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | Label | missings | Values | Value Labels | Freq. | % |
| 1 | densidad |  | 0 (0.00%) |  | 10 18 24 | 18 18 18 | 33.33 33.33 33.33 |
| 2 | estacion |  | 0 (0.00%) |  | primavera verano | 27 27 | 50.00 50.00 |
| 3 | aumento |  | 0 (0.00%) | *range: 60.0-402.0* | |  |  |

1. Ahora, para determinar si hay un efecto significativo de nuestros dos factores (densidad y estación), podemos utilizar un ANOVA de dos vías:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: ANOVA de dos vías  aumento~densidad\*estación   |  |  | | --- | --- | |  | aumento | | Predictors | p | | densidad | **<0.001** | | estacion | **<0.001** | | densidad:estacion | **<0.001** | | Residuals |  | | Observations | 54 | | R2 / R2 adjusted | 0.936 / 0.929 | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | |

El ANOVA (fórmula aumento ~ densidad \* estación) sugiere que:

* El efecto principal de la **densidad** es estadísticamente significativo y grande (F(2, 48) = 119.20, p < .001; Eta2 (partial) = 0.83, 95% CI [0.76, 1.00]).
* El efecto principal de **estación** es estadísticamente significativo y grande (F(1, 48) = 432.63, p < .001; Eta2 (partial) = 0.90, 95% CI [0.86, 1.00]).
* La interacción entre ***densidad y estación*** es estadísticamente significativa y grande (F(2, 48) = 16.20, p < .001; Eta2 (partial) = 0.40, 95% CI [0.22, 1.00]).

El siguiente gráfico se realizó para ejemplificar la diferencia de medias (comparasiones múltiples ajustada con corrección de *Holm*) :

|  |
| --- |
| Figure 1: Diferencia de medias |

Supuestos del modelo:

|  |
| --- |
| Figure 2: Supuestos del ANOVA |

Parce ser que no tenemos violaciones graves a nuestro modelo.

## 2 Tilapias por separado

Analiza los mismos datos del ejercicio anterior pero probando el efecto de la estación y de la densidad por separado. ¿Qué diferencia encuentras en la interpretación de los resultados? Dado el diseño experimental y la pregunta central, ¿qué tipo de análisis es el más conveniente y por qué? No te olvides de revisar los supuestos.

1. **Densidad**:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: ANOVA para densidad  aumento ~ densidad   |  |  | | --- | --- | |  | aumento | | Predictors | p | | densidad | **<0.001** | | Residuals |  | | Observations | 54 | | R2 / R2 adjusted | 0.317 / 0.290 | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | |

|  |
| --- |
| Figure 3: Diferencia de medias densidad |

El ANOVA de densidad (formula: aumento ~ densidad) sugiere que el efecto principal de la densidad es estadísticamente significativo y grande (F(2, 51) = 11.85, p < .001; Eta2 = 0.32, 95% CI [0.14, 1.00]),.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: ANOVA para estación  aumento~estacion   |  |  | | --- | --- | |  | aumento | | Predictors | p | | estacion | **<0.001** | | Residuals |  | | Observations | 54 | | R2 / R2 adjusted | 0.576 / 0.568 | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | |

|  |
| --- |
| Figure 4: Diferencia de medias estación |

El ANOVA de estación (formula: aumento ~ estación) sugiere que el efecto principal de la estación es estadísticamente significativo y grande (F(1, 52) = 70.57, p < .001; Eta2 = 0.58, 95% CI [0.43, 1.00]).

1. Revisar supuestos (parece que el modelo con interacción cumple mejor con los supuestos).

|  |
| --- |
| Figure 5: ANOVA densidad supuestos de modelo |

|  |
| --- |
| Figure 6: ANOVA estación supuestos de modelo |

**Conclusión**

Si analizamos cada factor por separado, solo podemos concluir sobre el efecto individual de cada variable predictora, sin considerar una interacción entre estas variables. Es decir, no podríamos modelar correctamente si la densidad influye sobre el crecimiento de las tilapias dependiendo de la estación del año (o viceversa).

Nuestro diseño experimental sugiere que tanto la densidad como la estación podrían interaccionar en su efecto sobre el crecimiento de las tilapias. En este caso, el ANOVA de dos vías permite evaluar tanto los efectos principales de cada factor como su interacción. Este análisis permite comprender mejor el fenómeno y puede servir mejor en la practica (por ejemplo, para gestionar la crianza de tilapias en diferentes condiciones).

|  |  |
| --- | --- |
|  | **Con Dummy variables**  NOTA, también se puede hacer el ejercicio como un modelo lineal con variables *dummy*:  [1] 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 18 18 18 18 18 18 18 [26] 18 18 18 18 18 18 18 18 18 18 18 24 24 24 24 24 24 24 24 24 24 24 24 24 24 [51] 24 24 24 24 Levels: 10 18 24  # A tibble: 3 × 2  densidad n  <fct> <int> 1 10 18 2 18 18 3 24 18  Call: lm(formula = aumento ~ 1 + dummy\_densidad\_b + dummy\_densidad\_c,   data = tilapias\_dummy)  Residuals:  Min 1Q Median 3Q Max  -154.100 -79.550 -6.222 82.250 131.200   Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 270.80 20.78 13.029 < 2e-16 \*\*\* dummy\_densidad\_b -81.80 29.39 -2.783 0.00754 \*\*  dummy\_densidad\_c -142.58 29.39 -4.851 1.2e-05 \*\*\* --- Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 88.18 on 51 degrees of freedom Multiple R-squared: 0.3173, Adjusted R-squared: 0.2905  F-statistic: 11.85 on 2 and 51 DF, p-value: 5.937e-05  Call: lm(formula = aumento ~ densidad, data = tilapias)  Residuals:  Min 1Q Median 3Q Max  -154.100 -79.550 -6.222 82.250 131.200   Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 270.80 20.78 13.029 < 2e-16 \*\*\* densidad18 -81.80 29.39 -2.783 0.00754 \*\*  densidad24 -142.58 29.39 -4.851 1.2e-05 \*\*\* --- Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  Residual standard error: 88.18 on 51 degrees of freedom Multiple R-squared: 0.3173, Adjusted R-squared: 0.2905  F-statistic: 11.85 on 2 and 51 DF, p-value: 5.937e-05  Como se observa, ambos enfoques dan el mismo resultado. |

## 3 Producción Maíz 2009

Estos datos provienen de un estudio del maíz en un área periurbana. Aunque hay más variables en la base, para este ejercicio queremos construir un modelo que permita predecir la producción de 2009 (variable produccion\_2009) a partir de las hectáreas sembradas (ha\_maiz). ¿Qué tan bueno es este modelo para realizar predicciones? Utiliza todos los recursos vistos en clase para contestar. NOTA. Elimina el dato con producción de 2009 igual a cero, pues el entrevistador ha reportado que se trata de un error de captura

1. Primero eliminamos datos con error (produccion\_2009 = 0):
2. Ahora, realizamos el *modelado estadístico*:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Producción de 2009 a partir de hectáreas de maíz  produccion\_2009 ~ ha\_maiz   |  |  |  |  | | --- | --- | --- | --- | |  | produccion 2009 | | | | Predictors | Estimates | CI | p | | (Intercept) | 3.41 | -1070.72 – 1077.53 | 0.995 | | ha maiz | 1969.97 \*\*\* | 1592.37 – 2347.57 | **<0.001** | | Observations | 82 | | | | R2 / R2 adjusted | 0.574 / 0.569 | | | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | | | |

Ajustamos un modelo lineal (estimado usando OLS) para predecir la producción produccion de 2009 a partir de las hectáreas de maíz (fórmula: produccion\_2009 ~ ha\_maiz). El modelo explica una proporción de varianza estadísticamente significativa y sustancial (R2 = 0.57, F(1, 80) = 107.79, p < .001, R2 ajustado = 0.57).

* El intercepto del modelo, correspondiente a ha\_maiz = 0, está en 3.41 (IC del 95% [-1070.72, 1077.53], t(80) = 6.31e-03, p = 0.995).

Dentro de este modelo:

* El efecto de ha maiz es estadísticamente significativo y positivo (beta = 1969.97, IC del 95% [1592.37, 2347.57], t(80) = 10.38, p < .001; beta estandarizada = 0.76, IC del 95% [0.61, 0.90])

Los parámetros estandarizados se obtuvieron ajustando el modelo en una versión estandarizada del conjunto de datos. Los intervalos de confianza del 95% (IC) y los *p-values* se calcularon utilizando una aproximación de la distribución t de Wald.

Ahora, revisamos los supuestos del modelo:

|  |
| --- |
| Figure 7: Supuestos del modelo lineal con un predictor |

## 4 Maíz II

Utilizando de nuevo los datos del estudio de maíz periurbano, realiza una comparación entre la producción del 2009 entre las diferentes comunidades. Sabemos que la producción está asociada con el área sembrada, por lo que quisiéramos tomar esto en cuenta en nuestro análisis al comparar las comunidades, i.e. queremos incorporar la covariable hectáreas sembradas de maíz (ha\_maiz). ¿Hay diferencias en producción entre los sitios tomando en cuenta las hectáreas sembradas de maíz? Aunque no se detecte una diferencia significativa entre comunidades, ¿cuál sería la ecuación de regresión para cada comunidad (la finalidad es ejercitar el uso de las variables dummy)? Genera la gráfica correspondiente con tres rectas, una para cada comunidad a partir de las ecuaciones que generaste para cada comunidad.

R automáticamente toma una de las categorías como referencia, por lo que necesitamos generar manualmente las variables *dummy*.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5: Modelo lineal con dos predictores  produccion\_2009 ~ ha\_maiz + comunidad   |  |  |  |  | | --- | --- | --- | --- | |  | produccion 2009 | | | | Predictors | Estimates | CI | p | | (Intercept) | 121.76 | -1568.54 – 1812.05 | 0.886 | | ha maiz | 1936.41 \*\*\* | 1491.20 – 2381.63 | **<0.001** | | comunidad [paredon] | 154.83 | -2252.40 – 2562.07 | 0.898 | | comunidad [sanfrancisco] | -220.45 | -2212.24 – 1771.33 | 0.826 | | Observations | 82 | | | | R2 / R2 adjusted | 0.575 / 0.558 | | | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | | | |

Sin embargo, también podemos hacer las variables *dummy* manualmente (a manera de un *ANCOVA*):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 6: Modelo lineal con predictores dummy  produccion\_2009 ~ ha\_maiz + dummy\_comunidad\_b + dummy\_comunidad\_c   |  |  |  |  | | --- | --- | --- | --- | |  | produccion 2009 | | | | Predictors | Estimates | CI | p | | (Intercept) | 121.76 | -1568.54 – 1812.05 | 0.886 | | ha maiz | 1936.41 \*\*\* | 1491.20 – 2381.63 | **<0.001** | | dummy comunidad b | 154.83 | -2252.40 – 2562.07 | 0.898 | | dummy comunidad c | -220.45 | -2212.24 – 1771.33 | 0.826 | | Observations | 82 | | | | R2 / R2 adjusted | 0.575 / 0.558 | | | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | | | |

Los coeficientes de los modelos corresponden a las hectáreas sembradas y para las diferencias entre las comunidades (ajustadas por las hectáreas). Esto nos permite interpretar cómo difiere la producción entre comunidades, controlando por el tamaño del área sembrada.

Nuestro modelo sugiere que añadir la comunidad al modelo no explica los cambios en la producción de 2009.

* Las ecuaciones de regresión para nuestros modelos son:

Como se observa, ambos enfoques dan exactamente el mismo resultado.

* Visualización del modelo con las Rectas de Regresión para cada comunidad:

|  |
| --- |
| Figure 8: Regresión lineal del modelo con comunidades |

## 5 Aves de Australia

Entender qué aspectos del ambiente y de las actividades humanas afectan la abundancia de organismos es un aspecto muy importante para la conservación de áreas naturales. En 1987 se colectaron datos de aves en 56 parches de vegetación natural en Australia. En este estudio se registró la abundancia de aves (abundancia) y algunas variables que serán utilizadas como predictoras de esta abundancia, incluyendo el área del parche medido (area), el tiempo en años en que dicho parche ha quedado aislado del resto de la vegetación natural (anos.aislam), la distancia al parche de vegetación más cercano (dist), la distancia al parche más grande de vegetación en el área (dist.parche.grande), la cantidad de ganado presente en el parche (ganado, medido de 1 a 5 donde 1 es poco ganado y 5 es abundante ganado), y la altitud. Se tienen en total seis variables predictoras. OJO: Revisa que exista una relación lineal entre la variable de respuesta y las predictoras. Es posible que algunas requieran una transformación.

### 5.1 Relación Lineal y transformaciones

1. Hacemos un gráfico con las 6 variables predictoras y añadimos una correlación de Pearson para evaluar la relación lineal.

|  |
| --- |
| Figure 9: Relación lineal de variables predictoras |

* altitud no se ve mal, pero tal vez podría mejorar con log?
* anos.aislam no se ve mal, pero tal vez podría mejorar con log?
* area, remover outliers o log?
* dist, remover outliers o log?
* dist.parche.grande tal vez podría mejorar con log?
* ganado parece bien

|  |
| --- |
| Figure 10: Relación lineal de variables predictoras log transformadas |

En comparación con [Figure 9](#fig-linealA), la transformación log ([Figure 10](#fig-linealB)):

* ahora parce que area mejoró
* altitud y anos.aislam están mejor sin log
* ganado queda igual
* ver si dist mejora raiz/cuadrada o elevar al cuadrado / estandarizar (rescalar) / quitar outliers
* ver si dist\_parche\_grande mejora mejora raiz/cuadrada o elevar al cuadrado / estandarizar (rescalar) / quitar outliers

|  |
| --- |
| Figure 11: Relación lineal de variables predictoras quitando outliers |

* En la [Figure 11](#fig-linealC), parece que dist sin outliers mejoró bastante y dist.parche.grande no parece mejorar con transformaciones (log, sqrt, ^2, ni eliminando outliers, ni centrando la media).

### 5.2 lm con todos los predictores

Ya con los datos transformados, hacemos el modelo lineal para predecir abundancia a partir de las 6 variables:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 7: Modelo lineal con 6 predictores  abund ~ altitud + anos.aislam + area\_log + dist + dist.parche.grande + ganado\_factor   |  |  |  |  | | --- | --- | --- | --- | |  | abund | | | | Predictors | Estimates | CI | p | | (Intercept) | 52.71 | -175.10 – 280.51 | 0.643 | | altitud | -0.00 | -0.05 – 0.05 | 0.905 | | anos.aislam | -0.02 | -0.13 – 0.10 | 0.747 | | area\_log | 3.26 \*\*\* | 1.99 – 4.53 | **<0.001** | | dist | -0.00 | -0.02 – 0.01 | 0.608 | | dist.parche.grande | -0.00 | -0.00 – 0.00 | 0.829 | | ganado\_factor2 | 0.85 | -5.64 – 7.35 | 0.793 | | ganado\_factor3 | 0.00 | -5.84 – 5.84 | 0.999 | | ganado\_factor4 | -1.20 | -7.69 – 5.29 | 0.712 | | ganado\_factor5 | -11.97 \* | -21.18 – -2.75 | **0.012** | | Observations | 55 | | | | R2 / R2 adjusted | 0.736 / 0.683 | | | | \* p<0.05   \*\* p<0.01   \*\*\* p<0.001 | | | | |

También podemos ver todas las comparaciones (contrastes) de nuestros factores:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 8: Contrastes del Modelo lineal con 6 predictores  **Table** **:** Contrastes del Modelo lineal con 6 predictores   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | altitud | 0.00 | -0.05, 0.05 | 0.905 | | anos.aislam | -0.02 | -0.13, 0.10 | 0.747 | | area\_log | 3.3 | 2.0, 4.5 | **<0.001** | | dist | 0.00 | -0.02, 0.01 | 0.608 | | dist.parche.grande | 0.00 | 0.00, 0.00 | 0.829 | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | 0.85 | -8.3, 10 | 0.999 | | ganado\_factor3 - ganado\_factor1 | 0.00 | -8.2, 8.2 | >0.999 | | ganado\_factor3 - ganado\_factor2 | -0.85 | -8.8, 7.1 | 0.998 | | ganado\_factor4 - ganado\_factor1 | -1.2 | -10, 8.0 | 0.996 | | ganado\_factor4 - ganado\_factor2 | -2.1 | -11, 6.8 | 0.964 | | ganado\_factor4 - ganado\_factor3 | -1.2 | -9.2, 6.8 | 0.993 | | ganado\_factor5 - ganado\_factor1 | -12 | -25, 1.0 | 0.085 | | ganado\_factor5 - ganado\_factor2 | -13 | -26, 0.57 | 0.066 | | ganado\_factor5 - ganado\_factor3 | -12 | -23, -1.4 | **0.020** | | ganado\_factor5 - ganado\_factor4 | -11 | -24, 2.4 | 0.156 | | 1CI = Confidence Interval | | | | |

Fórmula de regresión:

Ajustamos un modelo lineal (estimado usando OLS) para predecir abundancia a partir de altitud, anos.aislam, area\_log, dist, dist.parche.grande y ganado (fórmula: abund ~ altitud + anos.aislam + area\_log + dist + dist.parche.grande + ganado). El modelo explica una proporción de varianza estadísticamente significativa y sustancial (R2 = 0.70, F(6, 48) = 18.36, p < .001, R2 ajustado = 0.66). El intercepto del modelo, correspondiente a altitud = 0, anos.aislam = 0, area\_log = 0, dist = 0, dist.parche.grande = 0 y ganado = 0, está en -115.32 (IC del 95% [-286.79, 56.16], t(48) = -1.35, p = 0.183). Dentro de este modelo:

* El efecto de la ***altitud*** es estadísticamente **no** significativo y positivo (beta = 1.39e-03, IC del 95% [-0.05, 0.05], t(48) = 0.06, p = 0.955; beta estandarizada = 5.80e-03, IC del 95% [-0.20, 0.21])
* El efecto de ***anos aislam*** es estadísticamente **no** significativo y positivo (beta = 0.07, IC del 95% [-0.02, 0.15], t(48) = 1.63, p = 0.110; beta estandarizada = 0.17, IC del 95% [-0.04, 0.38])
* El efecto de ***area log*** es estadísticamente significativo y positivo (beta = 3.53, IC del 95% [2.30, 4.76], t(48) = 5.78, p < .001; beta estandarizada = 0.63, IC del 95% [0.41, 0.85])
* El efecto de ***dist*** es estadísticamente **no** significativo y negativo (beta = -0.01, IC del 95% [-0.02, 2.93e-03], t(48) = -1.56, p = 0.125; beta estandarizada = -0.15, IC del 95% [-0.33, 0.04])
* El efecto de ***dist parche grande*** es estadísticamente **no** significativo y negativo (beta = -9.26e-04, IC del 95% [-3.13e-03, 1.27e-03], t(48) = -0.85, p = 0.402; beta estandarizada = -0.08, IC del 95% [-0.27, 0.11])
* El efecto de ***ganado*** es estadísticamente **no** significativo y negativo (beta = -1.72, IC del 95% [-3.53, 0.09], t(48) = -1.91, p = 0.062; beta estandarizada = -0.24, IC del 95% [-0.49, 0.01])

Los parámetros estandarizados se obtuvieron ajustando el modelo en una versión estandarizada del conjunto de datos. Los intervalos de confianza del 95% (IC) y los valores p se calcularon utilizando una aproximación de la distribución t de Wald.

**Grafico de los estimados del modelo con IC**

|  |
| --- |
| Figure 12 |

* Podemos observar cómo el logarítmo del área del parche medido tiene un efecto positivo y significativo sobre la abundancia.
* También se observa cómo el nivel 5 del factor de ganado tiene un efecto negativo y significativo sobre la abundancia.

Adicionalmente, podemos graficar los efectos de cada predictor sobre la variable de respuesta:

|  |
| --- |
| Figure 13 |

* Igual se observa cómo area\_log y el nivel 5 de ganado tienen una mayor influencia sobre la abundancia.

|  |  |
| --- | --- |
|  | **Valores predichos para los efectos de cada variable**  Los valores predichos para los efectos de [Figure 13](#fig-efectos) son los siguientes:  $altitud # Predicted values of abund  altitud | Predicted | 95% CI ------------------------------------  60 | 19.40 | [14.91, 23.89]  90 | 19.31 | [16.15, 22.47]  120 | 19.22 | [17.19, 21.26]  140 | 19.17 | [17.54, 20.79]  150 | 19.14 | [17.52, 20.75]  165 | 19.09 | [17.23, 20.95]  175 | 19.06 | [16.91, 21.22]  260 | 18.82 | [12.99, 24.64]  $anos.aislam # Predicted values of abund  anos.aislam | Predicted | 95% CI ----------------------------------------  1890 | 20.25 | [13.23, 27.28]  1900 | 20.07 | [14.16, 25.97]  1910 | 19.88 | [15.07, 24.69]  1920 | 19.69 | [15.95, 23.44]  1940 | 19.32 | [17.39, 21.25]  1950 | 19.14 | [17.53, 20.74]  1960 | 18.95 | [16.93, 20.97]  1980 | 18.58 | [14.70, 22.46]  $area\_log # Predicted values of abund  area\_log | Predicted | 95% CI -------------------------------------  -3 | 2.49 | [-4.19, 9.18]  -2 | 5.75 | [ 0.29, 11.21]  0 | 12.26 | [ 9.14, 15.39]  1 | 15.52 | [13.38, 17.66]  2 | 18.78 | [17.17, 20.38]  4 | 25.29 | [22.41, 28.17]  5 | 28.55 | [24.55, 32.55]  8 | 38.32 | [30.68, 45.96]  $dist # Predicted values of abund  dist | Predicted | 95% CI ---------------------------------  25 | 19.85 | [16.68, 23.02]  95 | 19.60 | [17.23, 21.96]  170 | 19.33 | [17.58, 21.07]  240 | 19.07 | [17.44, 20.70]  310 | 18.82 | [16.77, 20.87]  380 | 18.57 | [15.79, 21.35]  450 | 18.31 | [14.68, 21.95]  595 | 17.79 | [12.25, 23.33]  $dist.parche.grande # Predicted values of abund  dist.parche.grande | Predicted | 95% CI -----------------------------------------------  0 | 19.32 | [17.06, 21.58]  550 | 19.19 | [17.55, 20.83]  1100 | 19.06 | [17.25, 20.87]  1650 | 18.93 | [16.31, 21.54]  2250 | 18.78 | [15.03, 22.54]  2800 | 18.65 | [13.77, 23.54]  3350 | 18.52 | [12.47, 24.57]  4450 | 18.26 | [ 9.83, 26.68]  $ganado\_factor # Predicted values of abund  ganado\_factor | Predicted | 95% CI ------------------------------------------ 1 | 22.01 | [17.31, 26.70] 2 | 22.86 | [17.81, 27.90] 3 | 22.01 | [18.74, 25.27] 4 | 20.81 | [15.77, 25.84] 5 | 10.04 | [ 3.77, 16.30]  attr(,"class") [1] "ggalleffects" "list"  attr(,"model.name") [1] "lm\_aves\_todos" |

Ahora, vamos a ver el desempeño de nuestro modelo:

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
---------------------------------------------------------------  
362.193 | 368.332 | 384.273 | 0.736 | 0.683 | 5.332 | 5.895

|  |
| --- |
| Figure 14: Supuestos del modelo con 6 predictores |

* Parece que nuestro modelo cumple con los supuestos de manera razonable.

### 5.3 Backward

Utiliza un procedimiento de eliminación backward comenzando con el modelo saturado (con las seis variables) para proponer un modelo múltiple que prediga la abundancia a partir de un subconjunto de las seis variables predictoras.

1. Modelo Saturado:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 9: Modelo backward saturado  **Table** **:** Modelo backward saturado   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | altitud | 0.00 | -0.05, 0.05 | 0.905 | | anos.aislam | -0.02 | -0.13, 0.10 | 0.747 | | area\_log | 3.3 | 2.0, 4.5 | **<0.001** | | dist | 0.00 | -0.02, 0.01 | 0.608 | | dist.parche.grande | 0.00 | 0.00, 0.00 | 0.829 | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | 0.85 | -8.3, 10 | 0.999 | | ganado\_factor3 - ganado\_factor1 | 0.00 | -8.2, 8.2 | >0.999 | | ganado\_factor3 - ganado\_factor2 | -0.85 | -8.8, 7.1 | 0.998 | | ganado\_factor4 - ganado\_factor1 | -1.2 | -10, 8.0 | 0.996 | | ganado\_factor4 - ganado\_factor2 | -2.1 | -11, 6.8 | 0.964 | | ganado\_factor4 - ganado\_factor3 | -1.2 | -9.2, 6.8 | 0.993 | | ganado\_factor5 - ganado\_factor1 | -12 | -25, 1.0 | 0.085 | | ganado\_factor5 - ganado\_factor2 | -13 | -26, 0.57 | 0.066 | | ganado\_factor5 - ganado\_factor3 | -12 | -23, -1.4 | **0.020** | | ganado\_factor5 - ganado\_factor4 | -11 | -24, 2.4 | 0.156 | | 1CI = Confidence Interval | | | | |

1. Hacemos la eliminación *Backward* utilizando el criterio AIC utilizando la función step()
2. La ecuación con el mejor criterio AIC por el método *Backward* es la siguiente (abund ~ area\_log + ganado\_factor):
3. La tabla de regresión de nuestro modelo óptimo seleccionado por el método *Backward* es el siguiente:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 10: Modelo backward óptimo  **Table** **:** Modelo backward óptimo   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | area\_log | 3.2 | 2.1, 4.2 | **<0.001** | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | 1.4 | -6.7, 9.5 | 0.988 | | ganado\_factor3 - ganado\_factor1 | 0.83 | -6.3, 7.9 | 0.997 | | ganado\_factor3 - ganado\_factor2 | -0.57 | -7.6, 6.5 | >0.999 | | ganado\_factor4 - ganado\_factor1 | -0.58 | -8.8, 7.7 | >0.999 | | ganado\_factor4 - ganado\_factor2 | -2.0 | -10, 6.4 | 0.962 | | ganado\_factor4 - ganado\_factor3 | -1.4 | -8.8, 6.0 | 0.983 | | ganado\_factor5 - ganado\_factor1 | -11 | -19, -2.7 | **0.004** | | ganado\_factor5 - ganado\_factor2 | -12 | -20, -4.8 | **<0.001** | | ganado\_factor5 - ganado\_factor3 | -12 | -18, -5.4 | **<0.001** | | ganado\_factor5 - ganado\_factor4 | -10 | -18, -2.5 | **0.004** | | 1CI = Confidence Interval | | | | |

1. Evaluar el modelo *Backward*:

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
---------------------------------------------------------------  
354.893 | 357.276 | 368.945 | 0.733 | 0.706 | 5.366 | 5.685

1. Validar supuestos del Modelo *Backward*

|  |
| --- |
| Figure 15: Supuestos del modelo backward óptimo |

### 5.4 Forward

Utiliza un procedimiento de selección forward comenzando con un modelo que solamente contenga al intercepto para proponer un modelo múltiple que prediga la abundancia a partir de un subconjunto de las seis variables predictoras.

1. Modelo Nulo (solo intercepto "abund ~ 1")
2. Selección *Forward*
3. La ecuación con el mejor criterio AIC por el método *Backward* es la siguiente (abund ~ area\_log + ganado\_factor). Ambos modelos (*Forward* y *Backward*) dieron el mismo resultado.
4. La tabla de regresión de nuestro modelo óptimo seleccionado por el método *Forward* es el siguiente:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 11: Modelo forward óptimo  **Table** **:** Modelo forward óptimo   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | area\_log | 3.2 | 2.1, 4.2 | **<0.001** | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | 1.4 | -6.7, 9.5 | 0.988 | | ganado\_factor3 - ganado\_factor1 | 0.83 | -6.3, 7.9 | 0.997 | | ganado\_factor3 - ganado\_factor2 | -0.57 | -7.6, 6.5 | >0.999 | | ganado\_factor4 - ganado\_factor1 | -0.58 | -8.8, 7.7 | >0.999 | | ganado\_factor4 - ganado\_factor2 | -2.0 | -10, 6.4 | 0.962 | | ganado\_factor4 - ganado\_factor3 | -1.4 | -8.8, 6.0 | 0.983 | | ganado\_factor5 - ganado\_factor1 | -11 | -19, -2.7 | **0.004** | | ganado\_factor5 - ganado\_factor2 | -12 | -20, -4.8 | **<0.001** | | ganado\_factor5 - ganado\_factor3 | -12 | -18, -5.4 | **<0.001** | | ganado\_factor5 - ganado\_factor4 | -10 | -18, -2.5 | **0.004** | | 1CI = Confidence Interval | | | | |

1. Evaluar el modelo *Forward*:

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
---------------------------------------------------------------  
354.893 | 357.276 | 368.945 | 0.733 | 0.706 | 5.366 | 5.685

1. Validar supuestos del modelo *Forward*:

|  |
| --- |
| Figure 16: Supuestos del modelo forward óptimo |

### 5.5 Modelado II

Utiliza el procedimiento de selección de modelos basado en el AIC para seleccionar un grupo de modelos que predigan la abundancia a partir de un subconjunto de las variables predictoras. Puedes acudir a Burnham et al. 2011 y a Symonds & Moussalli 2011 y al ejercicio que realizamos en clase para revisar de nuevo esta aproximación. Para hacer el ejercicio más sencillo (es decir, para reducir el número de modelos a explorar), utiliza solamente el siguiente conjunto de cuatro variables predictoras en lugar de las seis de los modelos de los incisos previos: *área, ganado, distancia al parche más grande y altitud*. ***Explora todos los modelos posibles*** para estas cuatro variables y realiza las comparaciones que sean necesarias a través del valor de AIC de los modelos. ¿Cuál sería el conjunto que propondrías como buenos modelos? ¿Cuál es la importancia relativa de las diferentes variables tomando en cuenta su aparición en los diferentes modelos candidatos?

Para explorar TODOS los modelos (incluidas las interacciones), podemos usar el siguiente código a partir del paquete [glmulti](https://www.jstatsoft.org/article/view/v034i12) (Calcagno and Mazancourt 2010).

Podemos revisar cuántos modelos óptimos obtuvimos (abajo de la línea roja representa los mejores modelos de acuerdo al criterio de *Akaike information criterion*):

|  |
| --- |
| Figure 17: Módelos óptimos basados en AIC |

Los mejores modelos (por criterio de AIC) propuestos son:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 12: Módelos óptimos basados en AIC  **Table** **:** Módelos óptimos basados en AIC   | model | aic | weights | | --- | --- | --- | | abund ~ 1 + ganado\_factor + area\_log + ganado\_factor:area\_log | 353.5732 | 0.07385706 | | abund ~ 1 + altitud + area\_log + dist.parche.grande + area\_log:altitud + ganado\_factor:altitud | 354.5491 | 0.04534102 | | abund ~ 1 + ganado\_factor + area\_log | 354.8935 | 0.03816841 | | abund ~ 1 + ganado\_factor + altitud + area\_log + ganado\_factor:area\_log | 355.2517 | 0.03190876 | | abund ~ 1 + altitud + area\_log + area\_log:altitud + ganado\_factor:altitud + ganado\_factor:area\_log | 355.3074 | 0.03103274 | | abund ~ 1 + ganado\_factor + area\_log + dist.parche.grande + ganado\_factor:area\_log | 355.4698 | 0.02861325 | |

Como vemos, el modelo óptimo propuesto es (“abund ~ 1 + ganado\_factor + area\_log + ganado\_factor:area\_log)

La tabla de regresión de este modelo es la siguiente:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 13: Resumen del Modelo con interacción  **Table** **:** Resumen del Modelo con interacción   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | -1.1 | -9.7, 7.5 | 0.996 | | ganado\_factor3 - ganado\_factor1 | -1.6 | -9.4, 6.2 | 0.979 | | ganado\_factor3 - ganado\_factor2 | -0.48 | -7.4, 6.4 | >0.999 | | ganado\_factor4 - ganado\_factor1 | -2.8 | -12, 6.0 | 0.890 | | ganado\_factor4 - ganado\_factor2 | -1.8 | -9.8, 6.3 | 0.971 | | ganado\_factor4 - ganado\_factor3 | -1.3 | -8.5, 5.9 | 0.986 | | ganado\_factor5 - ganado\_factor1 | -14 | -23, -4.9 | **<0.001** | | ganado\_factor5 - ganado\_factor2 | -13 | -22, -4.6 | **<0.001** | | ganado\_factor5 - ganado\_factor3 | -13 | -20, -5.0 | **<0.001** | | ganado\_factor5 - ganado\_factor4 | -11 | -20, -2.6 | **0.005** | | area\_log | 1.9 | 0.13, 3.6 | **0.036** | | ganado\_factor \* area\_log |  |  |  | | 2 \* area\_log | 1.8 | -0.78, 4.4 | 0.165 | | 3 \* area\_log | 2.1 | -0.87, 5.1 | 0.161 | | 4 \* area\_log | 6.6 | 1.6, 12 | **0.011** | | 5 \* area\_log | 0.80 | -2.3, 3.9 | 0.601 | | 1CI = Confidence Interval | | | | |

La importancia relativa de los términos (evidencia para cada variable en todos los modelos evaluados) es la siguiente:

|  |
| --- |
| Figure 18: Importancia relativa de los términos |

|  |  |
| --- | --- |
|  | **Con efectos Mixtos**  También podríamos evaluarlo con efectos mixtos.   * Podríamos considerar altitud como un efecto aletorio; si la influencia de esta variable sobre la abundancia es variada, podría ser un efecto aleatorio. * Tal vez ganado podría considerarse. Aunque suena más a un efecto fijo, si tenemos varios registros por nivel de ganado o si nos interesa la variabilidad de la respuesta de la abundancia de aves a diferentes niveles de manejo de ganado en varios sitios, podríamos tomarlo como un efecto aleatorio.   Initialization... TASK: Exhaustive screening of candidate set. Fitting...  After 50 models: Best model: abund~1+ganado\_factor+area\_log+ganado\_factor:area\_log Crit= 330.904738800118 Mean crit= 385.487585313527 Completed.  0.86 sec elapsed  glmulti.analysis Method: h / Fitting: glmer.glmulti / IC used: aic Level: 2 / Marginality: FALSE From 36 models: Best IC: 330.904738800118 Best model: [1] "abund ~ 1 + ganado\_factor + area\_log + ganado\_factor:area\_log" Evidence weight: 0.994012070053268 Worst IC: 454.757938515849 1 models within 2 IC units. 0 models to reach 95% of evidence weight.    Como vemos, parece ser que en este caso llegamos a la misma fórmula y no es necesario considerar los efectos mixtos para mejorar nuestro modelo.  abund ~ 1 + ganado\_factor + area\_log + ganado\_factor:area\_log |

Ahora comparamos nuestro modelo óptimo con los modelos *backward* y *forward*:

Podemos comparar los modelos *backward* y *forward* con la función anova:

Analysis of Variance Table  
  
Model 1: abund ~ area\_log + ganado\_factor  
Model 2: abund ~ area\_log + ganado\_factor  
 Res.Df RSS Df Sum of Sq Pr(>Chi)  
1 49 1583.7   
2 49 1583.7 0 0

Ambos modelos convergen en la misma fórmula:

lm(formula = abund ~ area\_log + ganado\_factor, data = aves\_final)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 14: Modelo forward y backward óptimo  **Table** **:** Modelo forward y backward óptimo   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | area\_log | 3.2 | 2.1, 4.2 | **<0.001** | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | 1.4 | -6.7, 9.5 | 0.988 | | ganado\_factor3 - ganado\_factor1 | 0.83 | -6.3, 7.9 | 0.997 | | ganado\_factor3 - ganado\_factor2 | -0.57 | -7.6, 6.5 | >0.999 | | ganado\_factor4 - ganado\_factor1 | -0.58 | -8.8, 7.7 | >0.999 | | ganado\_factor4 - ganado\_factor2 | -2.0 | -10, 6.4 | 0.962 | | ganado\_factor4 - ganado\_factor3 | -1.4 | -8.8, 6.0 | 0.983 | | ganado\_factor5 - ganado\_factor1 | -11 | -19, -2.7 | **0.004** | | ganado\_factor5 - ganado\_factor2 | -12 | -20, -4.8 | **<0.001** | | ganado\_factor5 - ganado\_factor3 | -12 | -18, -5.4 | **<0.001** | | ganado\_factor5 - ganado\_factor4 | -10 | -18, -2.5 | **0.004** | | 1CI = Confidence Interval | | | | |

Ahora, comparamos estos métodos (*forward* y *backward*) con nuestro modelo (“abund ~ 1 + ganado\_factor + area\_log + ganado\_factor:area\_log”):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 15: Modelo con interacción  **Table** **:** Modelo con interacción   | **Characteristic** | **Beta** | **95% CI**1 | **p-value** | | --- | --- | --- | --- | | area\_log | 3.2 | 2.1, 4.2 | **<0.001** | | ganado\_factor |  |  |  | | ganado\_factor2 - ganado\_factor1 | 1.4 | -6.7, 9.5 | 0.988 | | ganado\_factor3 - ganado\_factor1 | 0.83 | -6.3, 7.9 | 0.997 | | ganado\_factor3 - ganado\_factor2 | -0.57 | -7.6, 6.5 | >0.999 | | ganado\_factor4 - ganado\_factor1 | -0.58 | -8.8, 7.7 | >0.999 | | ganado\_factor4 - ganado\_factor2 | -2.0 | -10, 6.4 | 0.962 | | ganado\_factor4 - ganado\_factor3 | -1.4 | -8.8, 6.0 | 0.983 | | ganado\_factor5 - ganado\_factor1 | -11 | -19, -2.7 | **0.004** | | ganado\_factor5 - ganado\_factor2 | -12 | -20, -4.8 | **<0.001** | | ganado\_factor5 - ganado\_factor3 | -12 | -18, -5.4 | **<0.001** | | ganado\_factor5 - ganado\_factor4 | -10 | -18, -2.5 | **0.004** | | 1CI = Confidence Interval | | | | |

| Name | AIC | AICc | BIC | R2 | R2\_adjusted | RMSE |
| --- | --- | --- | --- | --- | --- | --- |
| forward\_optimo.2 | 354.8935 | 357.2765 | 368.9448 | 0.7328060 | 0.7055414 | 5.366056 |
| lm\_interaccion\_optimo | 353.5732 | 359.7128 | 375.6539 | 0.7744559 | 0.7293471 | 4.930122 |

Podemos evaluarlos con base en su rendimiento (0-100%). Para esto, el cálculo se basa en normalizar todos los índices (i.e. escalarlos en un rango de 0 a 1) y tomar el valor medio de todos los índices para cada modelo.

| Name | R2 | R2\_adjusted | RMSE | Performance\_Score |
| --- | --- | --- | --- | --- |
| lm\_interaccion\_optimo | 0.7744559 | 0.7293471 | 4.930122 | 0.7142857 |
| forward\_optimo.2 | 0.7328060 | 0.7055414 | 5.366056 | 0.2857143 |

Podemos graficar este rendimiento en un gráfico spiderweb donde los índices son normalizados (mayores valores indican mejor rendimiento).

|  |
| --- |
| Figure 19: Spiderweb comparando modelos |

### 5.6 Comparar modelos

Compara los resultados obtenidos con los diferentes métodos de selección de modelos. ¿A qué conclusión llegamos que conteste la pregunta principal de investigación.

La principal diferencia entre los dos modelos lineales (obtenido por método forward/backward y nuestro modelado) es la inclusión de un término de interacción en el segundo modelo.

Modelo *forward/backward* **lm(formula = abund ~ area + ganado)**

Este modelo es un modelo lineal simple que incluye dos predictores: **area** y **ganado**. Esta fórmula modela la abundancia de aves como una función lineal del área del parche y la cantidad de ganado presente, sin considerar cómo la combinación de estos dos factores podría afectar de manera conjunta la abundancia de aves. En este modelo, los coeficientes para **area** y **ganado** representan el cambio esperado en la abundancia de aves por cada unidad de cambio en **area** y **ganado**, respectivamente, asumiendo que el otro predictor se mantiene constante.

Modelo 2 **lm(formula = abund ~ 1 + ganado + area + ganado:area)**

Este modelo también incluye una interacción entre **ganado** y **area**. El término de interacción **ganado:area** examina cómo el efecto de una variable (por ejemplo, **ganado**) sobre la abundancia de aves cambia dependiendo del nivel de la otra variable (**area**), y viceversa.

* **Efectos principales**: Los coeficientes de **ganado** y **area** todavía representan el efecto de cada uno de estos predictores sobre la abundancia de aves, pero ahora estos efectos son condicionales al nivel de la otra variable siendo en su nivel base (el nivel más bajo).
* **Término de interacción**: El coeficiente de **ganado:area** indica cuánto cambia el efecto de **ganado** sobre la abundancia por cada unidad de incremento en **area**, y viceversa. Si este coeficiente es significativamente diferente de cero, implica que el efecto de una variable sobre la abundancia de aves depende del nivel de la otra variable.

Si bien, al comparar las métricas de ambos modelos (AIC, R^2) no hay mucha diferencia, podemos pensar que el modelo con interacción nos es útil si creemos que la relación entre las variables predictoras y la respuesta no es simplemente aditiva, sino que una variable podría modificar el efecto de la otra. Podría ser que el efecto del ganado sobre la abundancia de aves sea más pronunciado en áreas más grandes o más pequeñas, lo cual no podría captarse en el primer modelo.. Además, la significancia estadística del término de interacción nos sugiere que es necesario incluirla en nuestro modelo.

Podemos ver la diferencia de los estimados en nuestros dos modelos de la siguiente manera (abundancia 2 en rojo corresponde al modelo con interacción, abundancia 1 en azul al modelo sin interacción):

|  |
| --- |
| Figure 20: Comparación de estimados de los 2 modelos |

|  |
| --- |
| Figure 21: Comparación de estimados de los 2 modelos 2 |

El modelo con interacción es muy similar al aditivo en casi todos los efectos principales. Sin embargo, algunos factores tienen estimados diferentes al considerar la interacción en el modelo. Por ejemplo, se puede observar que el nivel de ganado 4 es más fuerte con la interacción. Además, permite ver cómo interaccionan de manera positiva los términos de interacción.

Por último, podemos visualizar los efectos sobre la variable de respuesta de nuestro modelo de interacción:

|  |
| --- |
| Figure 22: Efectos sobre la variable de respuesta con modelo de interacción |

|  |
| --- |
| Figure 23: Efectos marginales de los términos de interacción |

Calcagno, Vincent, and Claire de Mazancourt. 2010. “**Glmulti**: An*R*Package for Easy Automated Model Selection with (Generalized) Linear Models.” *Journal of Statistical Software* 34 (12). <https://doi.org/10.18637/jss.v034.i12>.