MODELOS LOG-LINEALES

1. Variables categóricas. Una breve introducción

La mayoría de modelos de la asignatura *Modelos Lineales Generalizados* se centran en el comportamiento de una variable numérica respuesta. Por ejemplo, el objetivo de los modelos de regresión logística es entender el comportamiento de los parámetros que definen a una variable binomial (número de éxitos en un experimento). Lo mismo para la regresión de Poisson, donde el objetivo se centra en los conteos obtenidos en un trabajo.

Sin embargo, los modelos log-lineales buscan obtener información y realizar contrastes sobre las variables categóricas de un modelo. Este cambio de paradigma es sutil, pero importante de comprender a la hora de entender los modelos log-lineales.

A continuación, se definen las variables categóricas y se muestra un método para contrastes simples empleado ampliamente en el mundo científico para su análisis.

1.1. ¿Qué son las variables categóricas?

Las variables categóricas son aquellas que definen la aparición o la ausencia de una característica sobre un individuo. Por ejemplo, si se está realizando un experimento sobre un grupo concreto de animales, la variable *especie* sería una variable categórica. Cada individuo pertenecerá a una especie concreta, es decir, estará dentro de esa categoría concreta.

Este concepto es equivalente a los factores que se trataron durante la asignatura de planificación. Lo que llamamos variable categórica es equivalente a un factor definido para un experimento, y sus categorías se tratarían de los niveles de ese factor.

Por lo tanto, las variables categóricas ya son un concepto familiar. Pero es importante comprender que en modelos log-lineales, nuestro interés se centra en estas variables categóricas y cómo se relacionan entre ellas.

1.2. Tablas de contingencia

Se trata de un método ampliamente usado para evaluar variables categóricas, o factores. Estos métodos analizan la frecuencia de individuos en distintas combinaciones de categorías para extraer conclusiones sobre los factores estudiados. Se suelen emplear cuando se quiere evaluar 1 o 2 factores. A continuación, se presentan un par de ejemplos.

Imaginemos un trabajo donde se cuentan las ballenas de 4 subpoblaciones diferentes. Se apunta la especie a la que pertenece cada uno de los individuos avistados, es decir, la variable categórica es la *especie* a la que pertenece cada individuo. Las subpoblaciones no son categorías de otra variable categórica porque no estamos interesados en analizar este factor o su interacción con la especie.

Cuando se tiene tan solo una variable categórica, se construye una tabla como la siguiente:

Especie	Subpoblación 1	Subpoblación 2	Subpoblación 3	Subpoblación 4
Jorobada	9	7	10	8
Gris	3	5	3	4
Azul	0	1	2	0

Tabla 1. Tabla de contingencia de frecuencias observadas. Se cuentan ballenas en base a su especie a lo largo de 4 subpoblaciones diferentes.

Esta tabla se completa mediante datos observados a la hora de realizar el trabajo. Estos datos son conteos, pero no hay que olvidar que el objetivo del trabajo se centra en la **variable categórica**, no en dichos conteos. Los conteos son solo una herramienta para obtener conclusiones acerca de la variable *Especie*.

Cuando se tiene un trabajo con una única variable categórica se puede realizar un **contraste de homogeneidad**. Este contraste permite saber si las tres especies se distribuyen igual en las 4 subpoblaciones o si existe alguna subpoblación que tenga una distribución distinta. Es decir, se plantean las siguientes hipótesis:

- H_0: La distribución de especies es homogénea entre subpoblaciones.
- H_1: Al menos una subpoblación presenta una distribución distinta.

El estadístico de contraste que se emplea en estas situaciones es el χ^2 (Chi-cuadrado). Para poder emplear este estadístico, es necesario calcular una **tabla de valores esperados**, cuyos valores se obtienen mediante la siguiente expresión:

$$E_{ij} = \frac{O_{i\cdot} \cdot O_{\cdot j}}{N}$$

Donde:

- $O_{i\cdot}$ son los valores observados totales por especie (fila),
- $O_{\cdot j}$ son los valores observados totales por subpoblación (columna),
- N es el total de observaciones.

Al aplicar esta fórmula a cada celda, se obtiene la siguiente tabla de valores esperados:

Especie	Subpoblación 1	Subpoblación 2	Subpoblación 3	Subpoblación 4
Jorobada	8	9	10	8
Gris	3	4	4	3
Azul	1	1	1	1

Tabla 2. Tabla de contingencia de frecuencias esperadas.

Una vez calculados los valores esperados, se procede al contraste con el estadístico Chi-cuadrado:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:

- r es el número de categorías de especies,
- k es el número de subpoblaciones.

Para los datos muestreados, se obtiene un valor $\chi^2 = 4.37$. Este resultado es **inferior al valor crítico** de χ^2 con 6 grados de libertad (calculados como $(r-1)(k-1)$) y un nivel de significancia $\alpha = 0.05$.

Conclusión: No se rechaza la hipótesis nula. Hay **homogeneidad** en la distribución de especies entre las cuatro subpoblaciones de ballenas analizadas.

Esta es una conclusión sobre la **variable categórica**, no sobre los conteos sobre los cuales se han basado los cálculos del contraste. Este es el matiz importante que hay que comprender para entender los modelos log-lineales.

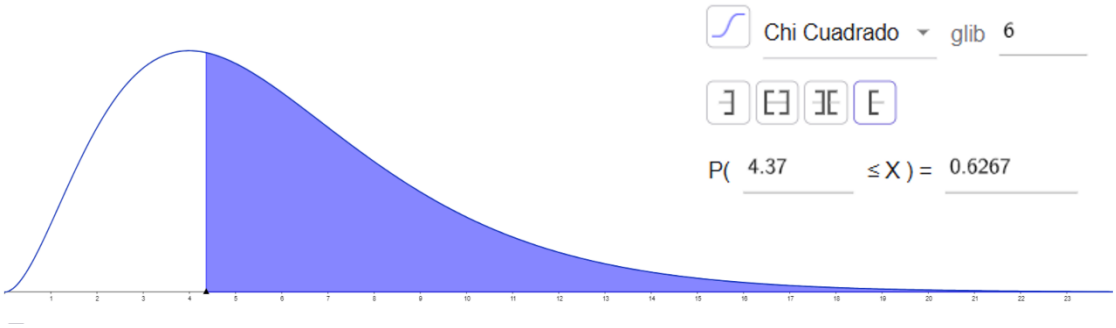

Figure 1: Representación del estadístico Chi Cuadrado para nuestro contraste.

1.3. Tablas de dos vías.

Las **tablas de contingencia** también se pueden emplear cuando aparecen **dos factores o variables categóricas**. Por ejemplo, supongamos que en el ejemplo anterior no solo queremos evaluar el factor *Especie*, sino también el factor *Zona*. Se vuelven a contar e identificar ballenas, pero esta vez en 4 zonas diferentes.

Tabla de frecuencias observadas para los factores Especie y Zona

Especie	Zona A	Zona B	Zona C	Zona D
Jorobada	9	3	10	2
Gris	3	5	6	2
Azul	0	1	4	0

Tabla 3. Frecuencias observadas para los factores *Especie* y *Zona*.

Ahora, con dos factores diferenciados, el interés del estudio se centra en la relación entre estas dos variables. Para el análisis se puede realizar un **contraste de independencia**, es decir, averiguar si ambos factores están relacionados entre sí o no.

Las hipótesis planteadas son:

- H_0: *Especie* y *Zona* son independientes.
- H_A: *Especie* y *Zona* no son independientes.

Se vuelve a emplear el estadístico χ^2 para resolver el contraste, y se utiliza la tabla de frecuencias esperadas (no mostrada en este ejemplo, pero se calcula igual que en el contraste de homogeneidad).

El resultado del test fue:

- $\chi^2 = 6.77$
- Grados de libertad $= 6$
- Nivel de significación $\alpha = 0.05$

Conclusión: El valor del estadístico no supera el valor crítico, por lo tanto, **no hay evidencia suficiente para rechazar la hipótesis nula**, lo cual apoya la independencia entre *Especie* y *Zona*.

Es importante recordar que este análisis evalúa la **asociación entre variables**, no las frecuencias absolutas de los datos.

Tanto el **contraste de homogeneidad** como el de **independencia** siguen los mismos cálculos y metodología. Sin embargo, las hipótesis que se contrastan y el objetivo del análisis **son diferentes**, y este detalle es fundamental al definir el objetivo de un trabajo.

> **Consejo:** Un contraste de independencia estadística comparte una lógica fundamental con la evaluación de interacciones en un ANOVA. En ambos casos, se busca determinar si la relación entre variables o factores es significativa o simplemente aleatoria. Quizá visualizar el contraste desde esta perspectiva ayude a mejorar la comprensión de cómo funciona un modelo log-lineal.

1.4. Tablas de tres vías

Las **tablas de tres vías** se utilizan cuando se dispone de datos clasificados según **tres variables**.

Cada celda de la tabla representa la **frecuencia de una combinación específica** de las categorías de las tres variables.

En este caso, las variables son:

- *Zona*
- *Estación*
- *Especie Observada*

Se presentan las **frecuencias observadas** para cada combinación de estas tres variables.

La siguiente tabla presenta las **frecuencias observadas** para cada combinación de las tres variables categóricas: *Zona*, *Estación* y *Especie Observada*. Este tipo de presentación es habitual en análisis de tablas de **tres vías**, donde el interés puede centrarse en efectos individuales o en posibles **interacciones entre factores**.

Zona / Estación	Ballena Jorobada	Ballena Gris	Total
Zona Norte			
Invierno	6	32	38
Verano	14	49	63
Total	20	81	101
Zona Sur			
Invierno	10	81	91
Verano	8	36	44
Total	18	117	135
Total general	38	198	236

Tabla 4. Frecuencias observadas para los factores *Zona*, *Estación* y *Especie observada*.

1.4.1. Tipos de independencia y asociación en tablas de tres vías Cuando se trabaja con tablas de contingencia que involucran **tres variables categóricas**, es importante distinguir entre varios conceptos de dependencia e independencia. Estos conceptos permiten analizar si las relaciones entre las variables son reales o el resultado de efectos indirectos o confundidos. A continuación, se describen los principales tipos:

Independencia condicional Dos variables son **condicionalmente independientes** respecto a una tercera cuando, al controlar por esta tercera variable, **la asociación entre las dos primeras desaparece**.

Ejemplo: Supongamos que estamos analizando tres variables:
- *Zona* (Norte o Sur)
- *Estación* (Invierno o Verano)
- *Especie observada* (Jorobada o Gris)

Si observamos una asociación entre *Zona* y *Estación*, pero al analizar por separado dentro de cada *Especie* dicha asociación desaparece, entonces decimos que **Zona y Estación son condicionalmente independientes dado Especie**.

Esto es crucial en análisis log-lineales, donde controlamos por una variable para estudiar interacciones entre otras.

Independencia marginal La **independencia marginal** se refiere a que dos variables son independientes sin tener en cuenta ninguna otra variable. Es decir, observamos su relación considerando únicamente los totales marginales.

Sin embargo, esta independencia **no garantiza** la independencia condicional. De hecho, puede inducir a errores de interpretación cuando existe una tercera variable que actúa como factor de confusión.

> **Paradoja de Simpson**: ocurre cuando una relación que se observa en los totales marginales se invierte o desaparece al desagregar por una tercera variable.

Asociación parcial Cuando dos variables no son completamente independientes después de controlar por una tercera, hablamos de **asociación parcial**. Esta puede presentarse de dos formas:

- El patrón de dependencia entre las dos variables **varía según los niveles** de la tercera variable (hay **interacción triple**).
- El patrón de dependencia **se mantiene constante** en todos los niveles de la tercera variable, aunque no desaparece del todo.

Esto indica que existe **cierta relación** entre las variables que no se anula al condicionar por otra.

Independencia múltiple Existe **independencia múltiple** cuando dos variables presentan el mismo grado de asociación en cada subtabla definida por la tercera variable. En este caso, la tercera variable es **independiente de las otras dos**, y no influye en su relación.

Esto se refleja en un patrón de asociación **consistente y estable** a lo largo de los niveles de la tercera variable.

Independencia completa Se dice que hay **independencia completa** cuando las tres variables son independientes entre sí, es decir:

- No hay dependencia directa entre ninguna pareja de variables.
- Tampoco existe dependencia condicional entre ninguna pareja dado la tercera.

Este es el escenario más simple y raro en la práctica. En términos de modelos log-lineales, correspondería a un modelo que solo incluye los términos marginales (sin interacciones).

1.5. Limitaciones de las tablas de contingencia y utilidad de los modelos log-lineales

A la hora de analizar la independencia entre **tres variables categóricas** mediante tablas de contingencia de tres vías, puede parecer que algunas variables son independientes cuando se condiciona respecto a una tercera. Sin embargo, es fundamental tener en cuenta que la **independencia condicional no implica necesariamente independencia marginal**. Es decir, aunque dos variables parezcan no relacionadas cuando se observa su relación sin controlar ninguna otra, pueden mostrar una asociación significativa cuando se analiza su relación **condicionando por una tercera variable**.

Este tipo de discrepancias puede generar confusión en la interpretación de los resultados. De hecho, en algunas tablas es posible que dos variables que inicialmente parecen independientes revelen una relación importante al incorporar el efecto de una tercera. Esta complejidad pone de manifiesto las **limitaciones de las tablas de contingencia tradicionales**, especialmente cuando el número de variables aumenta.

A partir de los **dos factores**, las tablas de contingencia dejan de ser una herramienta práctica. Buscar relaciones entre **tres variables** implica analizar no solo asociaciones marginales y condicionales, sino también **interacciones triples**, lo cual se vuelve difícil de manejar e interpretar de forma manual.

Además, para construir correctamente una tabla de valores esperados con más de dos factores, es necesario considerar **independencias condicionales múltiples** para cada combinación de niveles entre las variables. Aunque este procedimiento es posible, resulta laborioso, extenso y puede generar errores si no se hace con precisión.

1.6. Modelos log-lineales: una solución eficaz

Ante estas limitaciones, se recurre a los **modelos log-lineales**, que permiten ajustar y analizar de forma precisa la estructura de las celdas de una tabla de contingencia de múltiples vías. Estos modelos incorporan de forma explícita las **interacciones** entre variables, permitiendo contrastar hipótesis de independencia (marginal o condicional) y facilitando la interpretación de asociaciones complejas.

Este enfoque es especialmente útil cuando se trabaja con datos de clasificación múltiple y se desea:

- Comprobar independencia entre factores.
- Evaluar la presencia de interacciones dobles o triples.
- Ajustar modelos que expliquen los patrones de frecuencias observadas.

Recordad: los modelos log-lineales están diseñados para estudiar **relaciones entre variables categóricas**, no para analizar una variable respuesta numérica. Su utilidad radica en desentrañar la estructura de dependencias dentro de una tabla de frecuencias.

2. Modelos log-lineales.

Como se ha comentado previamente, el interés ahora no se encuentra en explicar una variable respuesta, sino en detectar la asociación entre variables. En este sentido, los modelos log-lineales nos aportan un procedimiento coherente y estructurado que nos permitirá identificar asociaciones simples y complejas entre dos o más variables.

Como el objetivo no es explicar una variable, estos modelos **no distinguen variables respuestas y predictoras**: ahora **todas las variables del estudio se consideran como variables respuesta**, porque lo que intentamos comprender es la asociación entre las mismas.

¿Cómo se integran estos modelos dentro de la categoría de Modelos Lineales Generalizados?

Como se explicó en el apartado anterior, trabajamos con **tablas de contingencia**. Para entender el funcionamiento de este tipo de modelo, es fundamental tener presente que lo que **se busca modelizar** es **la**

frecuencia esperada en cada celda de la tabla. Es decir, dado que nuestro objetivo es analizar la asociación entre variables categóricas, el modelo se centra en estimar con qué frecuencia se combinarían los distintos niveles o categorías de las variables si existiera una dependencia entre ellas.

¿Cómo se modela la frecuencia de cada combinación de niveles?

La frecuencia esperada en cada celda corresponde a datos de tipo conteo. Como ya se ha tratado en la asignatura, este tipo de datos suele modelizarse mediante una **distribución de Poisson**. Además, para vincular la media de esta distribución con el predictor lineal, se utiliza habitualmente la **función de enlace logarítmica**.

En este caso, estaríamos frente a un modelo lineal de términos aditivos (similar a un ANOVA de dos factores fijos) tal que:

$$Y_{ij} \sim Po(\mu_{ij})$$

$$y_{ij} = log(E_{ij}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

donde:

μ es el efecto general.

α_i es el efecto de la i-ésima categoría de la 1ª variable clasificatoria.

β_j es el efecto de la j-ésima categoría de la 2ª variable clasificatoria.

$\alpha\beta_{ij}$ es el efecto interacción de la i-ésima categoría de la 1ª y la j-ésima categoría de la 2ª variable.

Ahora bien, cuando hay **independencia** entre las dos variables, la expresión de las frecuencias esperadas no incluye la interacción. Es decir, $\alpha\beta_{ij} = 0$. Este es el contraste que nos interesa realizar.

¿Cómo contrastamos si $\alpha\beta_{ij} = 0$?

Como ya se ha comentado, en este contexto no nos interesa modelizar una variable respuesta específica, sino evaluar si existe asociación entre las variables categóricas analizadas. Por tanto, el enfoque no consiste en seleccionar el modelo con mejor ajuste o capacidad predictiva, interpretarlo y utilizarlo para hacer predicciones, como se ha hecho en otros casos. En su lugar, lo que haremos será comparar dos modelos: uno que no incluye el término de interacción $\alpha\beta$ (asumiendo independencia) y otro que sí lo incorpora (permitiendo interacción). Esta comparación se llevará a cabo mediante una técnica ya conocida en el curso: evaluaremos si la diferencia de *deviance* entre ambos modelos es significativa. Veamos esto aplicado en una tabla de dos vías o factores.

2.1. Modelos para tablas de Dos Vías.

Recuerda: cuando hay **independencia** entre las dos variables de interés, la expresión de las frecuencias esperadas no incluye la interacción. Esta será nuestra **hipótesis nula**. La forma de llevar a cabo este contraste es comparar el ajuste del modelo sin ese término de interacción con el modelo que sí lo incluye, siendo en este caso el saturado:

$$\begin{cases} H_0: & \log(E_{ij}) = \mu + \alpha_i + \beta_j \\ H_A: & \log(E_{ij}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \end{cases}$$

- De nuevo, para realizar el contraste, se compara el ajuste del modelo sin ese término (H_0) con el modelo saturado (H_A). Aquí se utiliza la $DEVIANCE$ como medida del ajuste:
  - Si la H_0 es correcta, el ajuste del modelo reducido (es decir, sin el término de interacción) será tan bueno como el saturado, dando lugar a una $DEVIANCE$ próxima a 0.

  - Si es falsa, habrá una diferencia en el ajuste de ambos modelos, y la $DEVIANCE$ será significativamente mayor que 0.

- La estimación de los parámetros se hace de la manera habitual: utilizando *mínimos cuadrados ponderados iterados*.
- Al igual que en χ^2, la $DEVIANCE$ se distribuye como una χ^2 $((I-1)(J-1))$.

Ejemplo de modelo para 2 vías.

Para ilustrar este contraste, retomaremos el ejemplo presentado en el apartado anterior. Se trata de una tabla de contingencia que recoge las frecuencias observadas de ballenas, clasificadas según tres especies distintas y cuatro zonas de muestreo. Las variables categóricas en estudio son, por tanto, *Especie* y *Zona muestreada*, con tres y cuatro niveles respectivamente. La pregunta que nos planteamos puede formularse del siguiente modo: ¿la distribución de especies de ballena varía según la zona muestreada, o bien ambas variables son independientes?

	Zona A	Zona B	Zona C	Zona D
Ballena Jorobada	10	8	12	6
Ballena Gris	2	6	2	6
Ballena Azul	0	2	5	6

Al hilo de lo anterior, podríamos anotar la hipótesis a contrastar tal que:

$$\begin{cases} H_0: & y_{ij} = \log(E_{ij}) = \beta_0 + \beta_1 \cdot \text{Gris} + \beta_2 \cdot \text{Azul} + \beta_3 \cdot \text{Zona}_B \\ & \quad +\beta_4 \cdot \text{Zona}_C + \beta_5 \cdot \text{Zona}_D \\ H_A: & y_{ij} = \log(E_{ij}) = \beta_0 + \beta_1 \cdot \text{Gris} + \beta_2 \cdot \text{Azul} + \beta_3 \cdot \text{Zona}_B \\ & \quad +\beta_4 \cdot \text{Zona}_C + \beta_5 \cdot \text{Zona}_D \\ & \quad +\delta_1 \cdot \text{Gris} : \text{Zona}_B + \delta_2 \cdot \text{Gris} : \text{Zona}_C + \delta_3 \cdot \text{Gris} : \text{Zona}_D \\ & \quad +\delta_4 \cdot \text{Azul} : \text{Zona}_B + \delta_5 \cdot \text{Azul} : \text{Zona}_C + \delta_6 \cdot \text{Azul} : \text{Zona}_D \end{cases}$$

Observa que es en la hipótesis alternativa (H_A) donde se incorporan todas las posibles interacciones entre los niveles de las variables categóricas.

En R, el contraste podría realizarse como sigue:

```
## Preparamos los datos
datos1 <- data.frame(
  conteo = c(10, 8, 12, 6,    # Jorobada
             2, 6, 2, 6,      # Gris
             0, 2, 5, 6),     # Azul
  Zona = factor(c("1", "2", "3","4",
                  "1", "2", "3","4",
                  "1", "2", "3","4"),
                levels = c("1", "2", "3","4")),
  Especie = factor(c("Jorobada", "Jorobada", "Jorobada","Jorobada",
                     "Gris", "Gris", "Gris", "Gris",
                     "Azul", "Azul", "Azul", "Azul"),
                   levels = c("Jorobada", "Gris", "Azul")))

attach(datos1)
```

```r
## Paso 1.
## . . . .
## Aquí ajustamos el modelo H_0 (sin la interacción) que será el que comparemos
## con el saturado en el contraste (Paso 2).
ajuste1 <- glm(conteo ~ Zona + Especie, family=poisson)
summary(ajuste1)
```

```
##
## Call:
## glm(formula = conteo ~ Zona + Especie, family = poisson)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8940     0.3094   6.122 9.26e-10 ***
## Zona2         0.2877     0.3819   0.753  0.45125
## Zona3         0.4595     0.3687   1.246  0.21268
## Zona4         0.4055     0.3727   1.088  0.27661
## EspecieGris  -0.8109     0.3005  -2.699  0.00696 **
## EspecieAzul  -1.0186     0.3236  -3.148  0.00164 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 29.745  on 11  degrees of freedom
## Residual deviance: 14.320  on  6  degrees of freedom
## AIC: 64.703
##
## Number of Fisher Scoring iterations: 5
```

En este caso, la impresión del `summary` tiene un propósito puramente pedagógico, ya que no nos ofrece por sí sola información concluyente sobre la asociación entre variables. Lo que realmente nos permitirá evaluarla es el contraste estadístico. Una vez ajustado el modelo, comparamos las *deviances*.

Dado que el modelo que incluye la interacción corresponde al modelo saturado, la *deviance* del modelo reducido (sin interacción) puede utilizarse directamente como estadístico de contraste. Este estadístico sigue una distribución χ^2 con $(I-1)(J-1)$ grados de libertad, donde I y J son el número de niveles de cada variable. Por tanto, el contraste puede plantearse del siguiente modo:

```r
## Paso 2.
## . . . .
## Ahora realizamos el contraste de:

    # H0: No hay interacción entre los factores (ab = 0). El término de
    # interacción (modelo saturado) no mejora el ajuste. La diferencia entre
    # ambos es pequeña.
    # HA: El modelo saturado mejora significativamente la deviance.
    # Es necesario incluir el(los) término(s) de interacción en el modelo.

1-pchisq(ajuste1$deviance, ajuste1$df.residual)
```

```
## [1] 0.02625963
```

Con un $p - valor = 0.02625$, tenemos evidencia estadísticamente suficiente para rechazar la hipótesis nula. Es decir, las variables *Especie* y *Zona muestreada* no son independientes. Véase la Figura 2 para una representación gráfica de cómo se obtiene este valor.

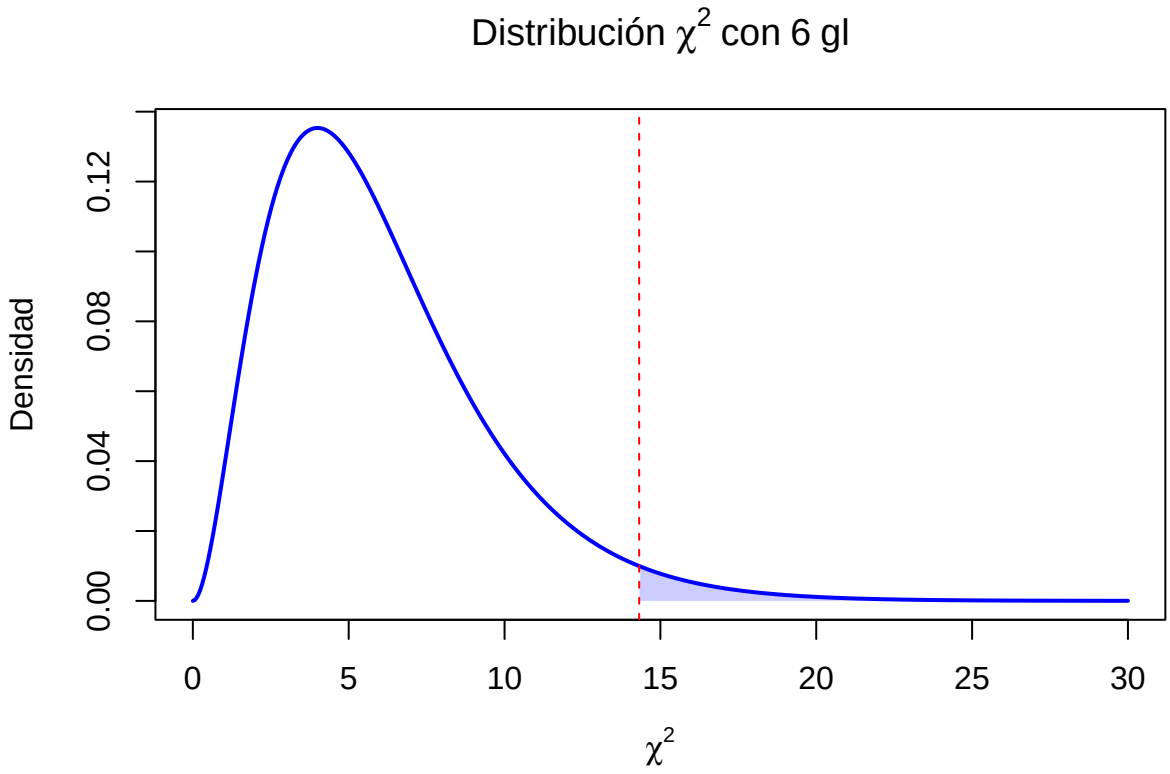

Figura 2: Distribución χ^2 para 6 grados de libertad. La línea roja discontinua marca la DEVIANCE observada en el modelo ejecutado para la hipótesis nula, siendo la DEVIANCE del modelo saturado igual a 0. La sombra azul responde al $p - valor$ observado, es decir, $P(X \geq$ deviance observada$) = 0.0263$.

Si quisiéramos interpretar este $p-valor$ en el contexto de la Figura 2, podríamos decir que es la probabilidad de obtener una DEVIANCE tan grande como la del modelo ajustado, si el modelo con interacción (que en este caso es el saturado con DEVIANCE = 0) no fuera necesario.

2.2. Modelos para Tablas de Tres Vías

Cuando se trabaja con **tablas de contingencia** que involucran **tres variables categóricas**, se pueden examinar las relaciones entre esas tres variables para entender cómo interactúan entre sí. En una tabla de tres vías, se tienen tres factores, cada uno con su propio número de categorías: la primera variable X con I categorías, la segunda variable Y con J categorías, y la tercera variable Z con K categorías.

Este tipo de análisis requiere modelos más complejos, ya que no solo debemos considerar las relaciones entre pares de variables, sino también las interacciones de tres factores a la vez. Esto hace que los modelos sean más detallados y que se puedan comprobar diferentes dependencias entre las variables.

En este caso, el **modelo saturado** es el más general e incluye todos los posibles términos, tanto los efectos principales como las interacciones entre las variables. Este modelo captura completamente la relación entre las tres variables y se expresa mediante la siguiente ecuación:

$$\log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

Donde:

- μ es el **efecto global** o intercepto.
- α_i representa el **efecto principal** de la i-ésima categoría de la primera variable X (por ejemplo, la zona).
- β_j es el **efecto principal** de la j-ésima categoría de la segunda variable Y (por ejemplo, la estación).
- γ_k es el **efecto principal** de la k-ésima categoría de la tercera variable Z (por ejemplo, la especie).
- $\alpha\beta_{ij}$ es el **efecto de la interacción** entre las variables X y Y (por ejemplo, cómo la relación entre zona y estación afecta a la distribución de especies).
- $\alpha\gamma_{ik}$ es el **efecto de la interacción** entre las variables X y Z (por ejemplo, cómo la zona afecta a la distribución de especies dependiendo del clima).
- $\beta\gamma_{jk}$ es el **efecto de la interacción** entre las variables Y y Z (por ejemplo, cómo la estación y la especie están relacionadas).
- $\alpha\beta\gamma_{ijk}$ es el **efecto de la interacción triple**, que refleja cómo las tres variables interactúan simultáneamente y afectan las frecuencias observadas.

Este modelo saturado incluye todos los efectos principales y las interacciones, lo que le permite ajustar perfectamente los datos. Sin embargo, en la práctica, puede ser demasiado complejo, por lo que se suelen considerar modelos más simples que incluyen solo ciertas interacciones, según las hipótesis que se deseen probar.

Inclusión de Términos

En los **modelos log-lineales**, la **inclusión de términos** es esencial para ajustar el modelo a los datos sin sobreajustarlo y sin perder relaciones importantes entre las variables. Comenzamos con un modelo simple, que solo incluye los **efectos principales** de las variables. Luego, podemos agregar términos que representen **interacciones** entre las variables, lo que nos permite capturar cómo una variable influye en otra, dependiendo de la tercera.

Es importante recordar que, en un modelo log-lineal, no estamos prediciendo una variable en función de las demás, sino explorando las relaciones simultáneas entre todas las variables. Es importante tener en cuenta que, si asumimos **márgenes fijos** para alguna o varias de las variables, los términos que representen estos márgenes deben incluirse necesariamente en el modelo. Los **márgenes fijos** se refieren a las restricciones que se imponen sobre los totales marginales de las variables, lo cual limita las posibles distribuciones de los datos. Si estos márgenes no se incorporan en el modelo, no se podrá describir correctamente la relación entre las variables y las frecuencias observadas. Esto garantiza que el modelo refleje adecuadamente las dependencias entre las variables y que los resultados sean consistentes con las restricciones del diseño experimental o los datos observados.

Este proceso de inclusión de términos asegura que el modelo sea lo suficientemente complejo para explicar las relaciones entre las variables sin agregar una complejidad innecesaria.

Comenzando con el Modelo Mínimo

El primer paso es definir el **modelo mínimo**, el cual asume que las variables son independientes entre sí. Este modelo sirve como el punto de referencia para comparar con modelos más complejos. En términos generales, para una tabla de tres vías, el modelo de **independencia completa** se expresa como:

$$\log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$$

En este modelo:

- α_i, β_j, y γ_k son los **efectos principales** de cada una de las variables (en este caso, Zona, Estación y Especie), y el modelo asume que estas variables son independientes entre sí.

Modelos con Interacciones Dobles

Una vez que tenemos el modelo de independencia completa, podemos considerar **interacciones entre pares de variables**. Estas interacciones permiten modelar cómo una variable influye en la relación entre otras dos. Por ejemplo, si consideramos las interacciones entre Zona y Estación, Zona y Especie, o Estación y Especie, podemos obtener una mejor descripción de cómo estas relaciones afectan las frecuencias observadas.

El modelo que incluye **interacciones dobles** se puede expresar como:

$$\log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$$

Aquí:

- $\alpha\beta_{ij}$, $\alpha\gamma_{ik}$, y $\beta\gamma_{jk}$ son los términos de **interacción entre dos variables**.
- Estos términos representan cómo la relación entre dos variables depende de la tercera variable.

Este modelo captura las interacciones entre pares de variables, pero no tiene en cuenta la interacción conjunta de las tres variables, lo que puede ser relevante en algunos casos.

Modelo Saturado

El **modelo saturado** es el modelo más complejo que incluye todas las interacciones posibles entre las tres variables. Esto incluye no solo las interacciones dobles, sino también la **interacción triple**, que muestra cómo la relación entre dos variables puede depender de la tercera. El modelo saturado se expresa como:

$$\log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

En este modelo:

- $\alpha\beta\gamma_{ijk}$ representa la **interacción triple** entre las tres variables. Si este término es significativo, sugiere que la relación entre las tres variables es compleja y no puede ser explicada solo por interacciones binarias.

El modelo saturado es útil para ajustar perfectamente los datos, pero a menudo no se utiliza como modelo final, ya que puede ser demasiado complejo y no siempre aporta información adicional significativa. En la práctica, se prefieren modelos más sencillos que capturen las interacciones más importantes y que sean más fáciles de interpretar.

Interpretaciones según los términos incluidos

- **Si el modelo no incluye la interacción de tercer orden** $\alpha\beta\gamma_{ijk}$:
  El modelo describe una **asociación parcial** entre las tres variables.
  Ejemplo: La relación entre Zona y Estación depende del nivel de Especie, pero no hay una interacción conjunta de las tres variables. Es decir, hay asociaciones entre pares, pero no una dependencia compleja entre Zona, Estación y Especie a la vez.
- **Si el modelo no incluye** $\alpha\beta\gamma_{ijk}$ ni $\alpha\beta_{ij}$:
  El modelo refleja **independencia condicional** entre Zona (X) y Estación (Y), dado Especie (Z).
  Ejemplo: Una vez que se fija el tipo de ballena observada, la Zona en la que se avistó y la Estación del año son independientes entre sí.

- **Si el modelo no incluye algunas interacciones de segundo orden** (como $\alpha\beta_{ij}$ pero sí $\alpha\gamma_{ik}$):
  El modelo describe **asociaciones condicionales** entre variables.
  Ejemplo: El modelo puede indicar que existe relación entre Zona y Especie, y entre Estación y Especie, pero no entre Zona y Estación, ni siquiera controlando por Especie.
- **Si el modelo no incluye ninguna interacción**:
  El modelo asume **independencia completa** entre las tres variables.
  Ejemplo: La zona geográfica, la estación del año y la especie observada son totalmente independientes: no hay ninguna asociación directa ni condicional entre ellas.

Consideraciones al Incluir Términos

Al incluir términos en el modelo, es importante equilibrar el ajuste del modelo con la **parsimonia**, es decir, la simplicidad del modelo. Aunque un modelo más complejo (como el saturado) puede ajustarse perfectamente a los datos, es posible que no proporcione una mejor interpretación o predicción. Por lo tanto, siempre se deben comparar modelos con diferentes niveles de complejidad utilizando estadísticos como la **deviance** o el **AIC**.

Además, al incluir términos, debemos tener en cuenta el tipo de **muestreo** y los **supuestos de independencia** entre las variables, ya que esto afectará a los términos que deben ser incluidos en el modelo. Por ejemplo, si se asume que una variable es marginalmente independiente de las demás, debemos asegurarnos de incluir los efectos principales de esa variable en el modelo.

En resumen, la inclusión de términos en un modelo log-lineal permite explorar las relaciones y dependencias entre las variables categóricas, y es un paso fundamental para construir un modelo que capture adecuadamente las interacciones entre los factores sin caer en el sobreajuste.

Ajuste de Modelos Log-lineales

Para ajustar un modelo log-lineal podemos seguir los pasos siguientes:

- Marcar el **modelo mínimo** del que partimos por el tipo de muestreo realizado.
- Comenzar ajustando el modelo sin la interacción de segundo orden. La **DEVIANCE** de este ajuste es un claro indicador de si existe la dependencia completa.

$$H_0 : \log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$$

$$H_1 : \log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

Si hay dependencia completa:

Si encontramos que el modelo de **independencia completa** es adecuado, ya tenemos el mejor modelo. Cuando no lo es, conviene contrastar (con la diferencia de **DEVIANCE** o **AIC**) el modelo con cualquiera de los que indican **independencia condicional**, por ejemplo:

$$H_0 : \log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik}$$

$$H_1 : \log(E_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$$

Si alguno de ellos resulta apropiado:

Podemos entonces eliminar otra interacción más y comprobar si una variable es independiente de las otras dos.

Finalmente, si alguno de estos modelos es apropiado, podemos compararlo con un modelo de **independencia completa** (sin interacciones).

Procedimiento para ajustar el modelo:

El proceso de ajuste puede realizarse mediante un **procedimiento backward** o **forward**. En el **procedimiento backward** (el que acabamos de explicar), comenzamos con el modelo más complejo (por ejemplo, el modelo saturado) y vamos eliminando términos de interacción hasta encontrar el modelo más parsimonioso. En el **procedimiento forward**, comenzamos con un modelo simple y vamos añadiendo términos de interacción hasta encontrar el modelo adecuado.

Comparación de Modelos:

Con el uso de herramientas en R, como las funciones **step**, **drop1** y **add1**, podemos realizar estos procedimientos de manera eficiente, comparando modelos y ajustándolos según los criterios de **deviance** y **AIC**.

Una alternativa es crear una tabla con las posibles modelizaciones y comparar rápidamente los resultados utilizando el **AIC** y la **deviance** de cada modelo. De esta manera, podemos identificar rápidamente el modelo que mejor se ajusta a los datos sin caer en el sobreajuste.

Ejemplo de modelo para 3 vías.

Para ilustrar la aplicación de los modelos log-lineales en tablas de contingencia de tres vías, vamos a hilar de nuevo con el ejemplo de la distribución de ballenas, esta vez con el añadido del factor *Estación*, pasando a tener tres variables categóricas de 2 niveles cada una:

- Zona: Representando la ubicación geográfica de la observación, constituido por Zona Norte y Zona Sur.
- Estación: La estación del año en la que se realiza la observación, Invierno y Verano.
- Especie: La especie de ballena observada, Ballena Jorobada y Ballena Gris.

De nuevo, nuestro interés radica en conocer si existe interacción o se trata de factores independientes, aunque esta vez evaluamos hasta el tercer grado de interacción.

La tabla a continuación muestra las frecuencias observadas de cada combinación de estas tres variables:

Zona / Estación	Ballena Jorobada	Ballena Gris	Total
Zona Norte			
Invierno	6	32	38
Verano	14	49	63
Total	20	81	101
Zona Sur			
Invierno	10	81	91
Verano	8	36	44
Total	18	117	135

Zona / Estación	Ballena Jorobada	Ballena Gris	Total
Total general	38	198	236

Con esta modificación, nuestro modelo saturado pasaría a ser:

$$log(E_{ijl}) = \beta_0 + \beta_1 \cdot Gris + \beta_2 \cdot Verano + \beta_3 \cdot Sur + \delta_1 \cdot Gris : Verano + \delta_2 \cdot Gris : Sur + \delta_3 \cdot Verano : Sur + \delta_4 \cdot Gris : Verano : Sur$$

Ajustamos los modelos saturado y con interacciones de primer orden:

```
#Modelo con iteraccio segundo orden, saturado
fit.sat <-glm (conteos ~ (estacion + zona + especie )^3,
               family = poisson)
fit.intd <-glm (conteos ~ (estacion + zona + especie )^2,
                family = poisson)
summary(fit.intd)
```

```
##
## Call:
## glm(formula = conteos ~ (estacion + zona + especie)^2, family = poisson)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         1.7408     0.3476   5.008 5.49e-07 ***
## estacionVerano                      0.5911     0.3752   1.575   0.1151
## zonaSur                             0.9193     0.3722   2.470   0.0135 *
## especieBallena Gris                 1.7342     0.3617   4.795 1.63e-06 ***
## estacionVerano:zonaSur             -1.2097     0.2768  -4.371 1.24e-05 ***
## estacionVerano:especieBallena Gris  0.3247     0.3727   0.871   0.3836
## zonaSur:especieBallena Gris        -0.5086     0.3749  -1.357   0.1749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 150.404895  on 7  degrees of freedom
## Residual deviance:   0.049752  on 1  degrees of freedom
## AIC: 52.994
##
## Number of Fisher Scoring iterations: 3
```

No hay diferencia significativa en el ajuste de la DEVIANCE entre el modelo saturado y el modelo con interacciones de primer grado (p-valor 0.2513). Podemos descartar el modelo saturado.

```
anova(fit.intd, fit.sat, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: conteos ~ (estacion + zona + especie)^2
## Model 2: conteos ~ (estacion + zona + especie)^3
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         1   0.049752
## 2         0   0.000000  1 0.049752   0.8235
```

Continuamos entonces comparando el modelo sin interacciones con el modelo de primer grado.

```
fit.int0 <-glm (conteos ~ estacion + zona + especie,
                family = poisson)
summary(fit.int0)
```

```
##
## Call:
## glm(formula = conteos ~ estacion + zona + especie, family = poisson)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.1849     0.1884  11.597   <2e-16 ***
## estacionVerano         0.2902     0.1316   2.205   0.0274 *
## zonaSur               -0.1870     0.1308  -1.430   0.1527
## especieBallena Gris    1.6507     0.1771   9.320   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 150.405  on 7  degrees of freedom
## Residual deviance:  24.589  on 4  degrees of freedom
## AIC: 71.532
##
## Number of Fisher Scoring iterations: 4
```

Existe una diferencia significativa en el ajuste de la DEVIANCE entre el modelo con interacciones de primer grado y sin interacciones (p-valor $2.322e^{-11}$). Además, el modelo con interacciones presenta un AIC: 54.951, significativamente mejor que el modelo sin interacciones con un valor de 101.47.

```
anova(fit.int0, fit.intd, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: conteos ~ estacion + zona + especie
## Model 2: conteos ~ (estacion + zona + especie)^2
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         4    24.5885
## 2         1     0.0498  3   24.539 1.928e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ejecutamos un algoritmo stepwise para buscar un mejor modelo, usando el modelo con todas las interacciones, y obtenemos un modelo con 2 interacciones:

```
step(fit.intd)
```

```
## Start:  AIC=52.99
## conteos ~ (estacion + zona + especie)^2
##
##                     Df Deviance    AIC
```

```
## - estacion:especie  1   0.8084 51.752
## - zona:especie      1   1.9069 52.851
## <none>                  0.0498 52.994
## - estacion:zona     1  19.9462 70.890
##
## Step:  AIC=51.75
## conteos ~ estacion + zona + especie + estacion:zona + zona:especie
##
##                 Df Deviance    AIC
## <none>               0.8084 51.752
## - zona:especie   1   3.6788 52.623
## - estacion:zona  1  21.7181 70.662

##
## Call:  glm(formula = conteos ~ estacion + zona + especie + estacion:zona +
##     zona:especie, family = poisson)
##
## Coefficients:
##                 (Intercept)               estacionVerano
##                      1.5504                       0.8733
##                     zonaSur          especieBallena Gris
##                      1.0110                       1.9548
##      estacionVerano:zonaSur  zonaSur:especieBallena Gris
##                     -1.2322                      -0.6032
##
## Degrees of Freedom: 7 Total (i.e. Null);  2 Residual
## Null Deviance:       150.4
## Residual Deviance: 0.8084    AIC: 51.75
```

```
fit.step <- glm(formula = conteos ~ estacion + zona + especie +
                  estacion:zona + zona:especie, family = poisson)
```

Dado que una de las interacciones del modelo obtenido anteriormente $Zona : Especie$, no es significativa a $\alpha = 0.95$, creamos un nuevo ajuste donde la eliminamos. Este nuevo ajuste además, tiene un AIC de 52.623 frente al 51.752, muy similar al modelo Stepwise original.

```
#summary(fit.step) #zona:especie no es signifcativa
fit.step2 <- glm(formula = conteos ~ estacion + zona +
                   especie + estacion:zona, family = poisson)
summary(fit.step2)
```

```
##
## Call:
## glm(formula = conteos ~ estacion + zona + especie + estacion:zona,
##     family = poisson)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.8113     0.2200   8.234  < 2e-16 ***
## estacionVerano            0.8733     0.1931   4.521 6.14e-06 ***
## zonaSur                   0.5055     0.2054   2.461   0.0138 *
## especieBallena Gris       1.6507     0.1771   9.320  < 2e-16 ***
## estacionVerano:zonaSur   -1.2322     0.2755  -4.473 7.73e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 150.4049  on 7  degrees of freedom
## Residual deviance:   3.6788  on 3  degrees of freedom
## AIC: 52.623
##
## Number of Fisher Scoring iterations: 4
```

Evaluamos que, efectivamente, la diferencia en DEVIANCE no es estadísticamente significativa entre modelos (p-valor = 0.09022).

```
anova(fit.step2, fit.step, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: conteos ~ estacion + zona + especie + estacion:zona
## Model 2: conteos ~ estacion + zona + especie + estacion:zona + zona:especie
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         3     3.6788
## 2         2     0.8084  1   2.8704  0.09022 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 7: Comparativa de resultados

Modelo	Deviance	AIC
Sin interacciones	24.589	71.53
Interacciones 1er orden	0.050	52.99
Saturado (2º orden)	0.000	54.94
Stepwise (modelo final)	3.679	52.62

Dado que no existen diferencias significativas, el AIC es más bajo y su menor número de variables lo hace más parsimonioso, escogemos el modelo Stepwise como modelo que mejora ajusta nuestros datos.

$$log(E_{ijl}) = \beta_0 + \beta_1 \cdot Gris + \beta_2 \cdot Verano + \beta_3 \cdot Sur + \beta_4 \cdot (Verano : Sur)$$

Que con sus correspondientes coeficientes sería:

$$log(E_{ijl}) = 1.8113 + 1.6507 \cdot Gris + 0.8733 \cdot Verano + 0.5055 \cdot Sur + (-1.2322) \cdot (Verano : Sur)$$

Table 8: Comparación entre conteos reales y predicciones del modelo

conteos	estacion	zona	especie	predicho	dif
6	Invierno	Norte	Ballena Jorobada	6	0
32	Invierno	Norte	Ballena Gris	32	0

conteos	estacion	zona	especie	predicho	dif
14	Invierno	Sur	Ballena Jorobada	10	-4
49	Invierno	Sur	Ballena Gris	53	4
10	Verano	Norte	Ballena Jorobada	15	5
81	Verano	Norte	Ballena Gris	76	-5
8	Verano	Sur	Ballena Jorobada	7	-1
36	Verano	Sur	Ballena Gris	37	1

Este modelo implica que la abundancia esperada de ballenas jorobadas en invierno en la zona norte es igual al intercepto; $e^{(1.8113)} = 6.118396 \sim 6$ ballenas, y, en cambio, el número de ballenas grises en verano en la zona norte sería de $e^{(1.8113+0.8733)} = 14.65234 \sim 15$. Graficamos los valores observados frente al ajuste:

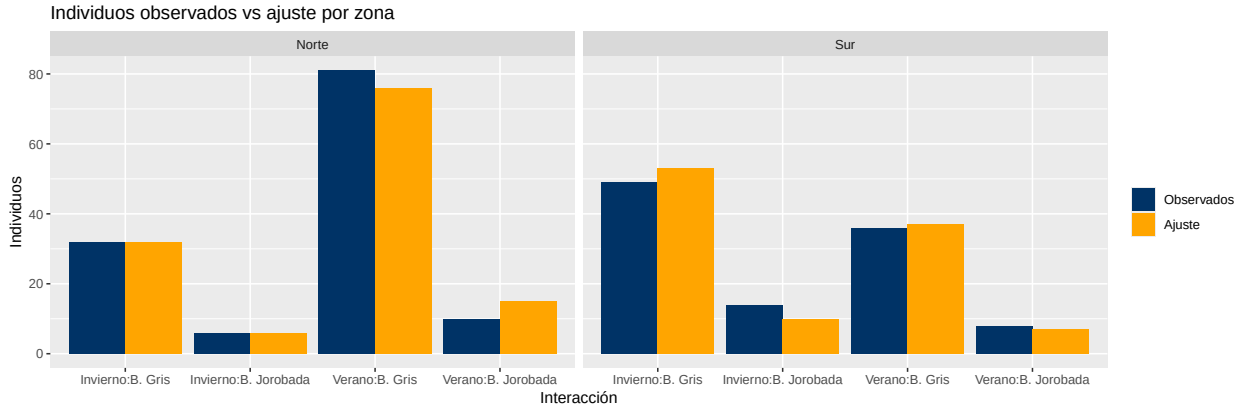

Figura 3:Comparación visual entre los conteos observados y los valores ajustados por el modelo stepwise final con interacción significativa entre estación y zona. Se representa el número de individuos observados (azul) frente al esperado bajo el modelo (amarillo) para cada combinación de estación y especie, separadas por zona. La proximidad entre ambas barras en cada grupo indica un buen ajuste del modelo sin necesidad de incluir interacciones entre especie y los demás factores.