Desafio 3

Situación 1

La viña Chateau Latour es ampliamente reconocida como una de las mejores en el mundo, y posee una rica historia que nace en el año 1638. La calidad de la cosecha tiene un gran impacto en el valor de comercialización del vino, pudiendo llegar a costar miles de dólares. Muchos han investigado el efecto del momento en que se realiza la cosecha sobre la calidad del vino, pero un aspecto menos explorado es el efecto que tienen las lluvias durante el tiempo de la cosecha sobre la misma. Este problema se refiere al último punto. Los datos se encuentran en el archivo Latour.txt y corresponden a:

Variable	Descripcion
cosecha	año de la cosecha
calidad	puntaje asignado a la calidad de la cosecha (1: peor,5: mejor)
dias	número de días, a partir del 31 de agosto, en que finalizó la cosecha
lluvia	1: lluvia durante la cosecha, 0: sin lluvia durante la cosecha

a. Obtenga un gráfico de la calidad de la cosecha versus el momento en que ella finalizó, diferenciando con colores o figuras las cosechas en que se produjo algún tipo de lluvia de aquellas en las que no. Comente sobre el efecto del momento de cosecha sobre su calidad, comparando ambos grupos. No olvide etiquetar correctamente la figura.
b. Se propone ajustar un modelo que permita tanto diferentes pendientes como interceptos a ambos grupos. Obtenga un intervalo de 95% de confianza para el número de días después del 31 de agosto en que debe finalizar una cosecha para que el valor medio estimado de su calidad disminuya en una unidad, cuando se observa lluvia durante la cosecha. Interprételo. Hint: Note que se le pide inferencia sobre un parámetro de la forma $\frac{1}{x_0^t\beta}$. Puede obtenerlo a partir de un intervalo para $x_0^t\beta$.

Archivo Latour:

cosecha	calidad	dias	lluvia
1961	5	28	0
1962	4	50	0
1963	1	53	1
1964	3	38	0
1965	1	46	1
1966	4	40	0
1967	3	35	1
1968	2	38	1
1969	2	45	1
1970	4	47	0
1971	3	45	1
1972	1	54	1
1973	2	39	1
1974	1	45	1
1975	4	40	1
1976	3	32	0
1977	2	47	0
1978	4	50	0
1979	3	48	0
1980	1	54	1
1981	3	39	1
1982	5	30	0

cosecha	calidad	dias	lluvia
1983	3	41	0
1984	1	44	1
1985	4	41	0
1986	4	46	0
1987	1	47	1
1988	4	40	0
1989	4	21	0
1990	5	32	0
1991	3	40	1
1992	1	39	1
1993	3	36	1
1994	3.5	29	1
1995	4	27	0
1996	5	32	0
1997	4	25	0
1998	3.5	35	1
1999	3.5	30	0
2000	5	41	0
2001	3.5	43	0
2002	4	47	0
2003	5	30	0
2004	4	49	0

Entendiendo el problema

La ideaa es analizar que pasa con la calidad del vino y la influencia de la presencia de lluvias en el periodo de la cosecha.

Para hacer este analisis debemos crear ungráfico para poder onservar la relación entre calidad y la isidencia de la lluvia.

Crear un modelo lineal en el cual podamos tener diferentes dependientes e interceptos para los dos grupos (cony sin lluvia) y hacerlo mediante la inclusión de una interacciónentre el número de dias y la variable de lluvia.

Debemos determinar para el grupo de las cosechas con lluvia cuantos dias adicionales sonnecesarios para que la calidad del vino disminuya en una unidad productiva, luego es necesario cuantificar la incertidmbre de la estimación obtenida para un intervalo de 95% de confianza para la cantidad obtenida.

Visualizar: Primero el grafico exploratorio que vamos ha hacer nos va ha permitir diferenciar las condiciones de lluvia o no y el cambio de la calidad.

Modelar: Vamos a identificar si el efecto de retrasar la cosecha segun las condiciones climáticas va ha hacer diferencia o no, por eso vamos ha incluir el término principal y el término de interaccion para que el modelo que vamos ha crear permita la variación entre la tasa de cambio de la calidad respecto a los dias y el intercepto varien segun la funcion de si hubo o no lluvia.

Interpretar: Bueno, en el grupo de lluvia la pendiente del modelo es la suma de los coeficientes asociados a los dias y a la interaccion de $\beta_1 + \beta_2$. Segun esa pediente vamos a calular el número de dias adicionales que se requieren para que la calidad disminuya 1 unidad usando la formula delta:

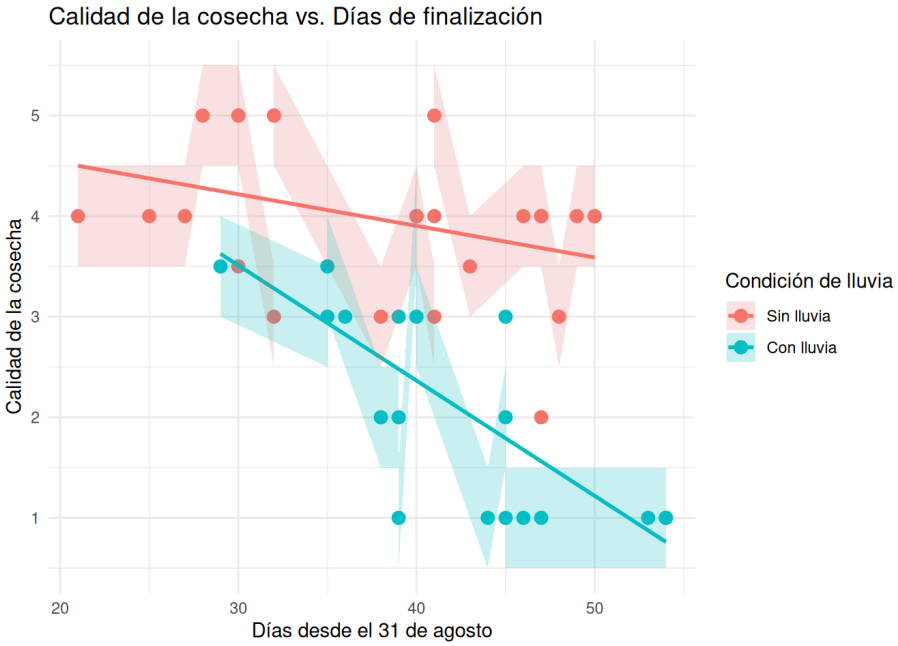
$$d = -\frac{1}{\beta_1 + \beta_3}$$

y obtenemos además el intpervalo de confianza que estamos buscando.

Solucion

```
library(ggplot2)
library(knitr)
library(broom)

# --- Definir los datos manualmente ---
datos <- data.frame(
  cosecha = c(1961:2004),
  calidad = c(5,4,1,3,1,4,3,2,2,4,3,1,2,1,4,3,2,4,3,1,3,5,3,1,4,4,1,4,4,5,3,1,3,3.5,4,5,4,3.5,3.5,5,3.5,4,5,4),
  dias = c(28,50,53,38,46,40,35,38,45,47,45,54,39,45,40,32,47,50,48,54,39,30,41,44,41,46,47,40,21,32,40,39,36,29,
27,32,25,35,30,41,43,47,30,49),
  lluvia = c(0,0,1,0,1,0,1,1,1,0,1,1,1,1,1,0,0,0,0,1,1,0,0,1,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,0,0,0)
)

# Convertir la variable 'lluvia' en factor
datos$lluvia <- factor(datos$lluvia, levels = c(0,1), labels = c("Sin lluvia", "Con lluvia"))

# Parte (a): Presentar el gráfico de dispersión con área sombreada para la intersección y la línea de regresion p
ara las dos condiciones,a si podemos identificar como se comportan los datos

ggplot(datos, aes(x = dias, y = calidad, color = lluvia, fill = lluvia)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_ribbon(aes(ymin = calidad - 0.5, ymax = calidad + 0.5), alpha = 0.2, color = NA) +
  labs(title = "Calidad de la cosecha vs. Días de finalización",
       x = "Días desde el 31 de agosto",
       y = "Calidad de la cosecha",
       color = "Condición de lluvia",
       fill = "Condición de lluvia") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
formulas <- data.frame(
  Condición = c("Sin lluvia", "Con lluvia"),
  Fórmula = c("$$y = \\beta_0 + \\beta_1 \\times dias$$",
              "$$y = (\\beta_0 + \\beta_2) + (\\beta_1 + \\beta_3) \\times dias$$")
)
kable(formulas, escape = FALSE, caption = "Fórmulas de regresión para cada condición de lluvia")
```

Fórmulas de regresión para cada condición de lluvia

Condición	Fórmula
Sin lluvia	$y = \beta_0 + \beta_1 \times dias$

Condición	Fórmula
Con lluvia	$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times dias$$

```
# Parte (b): Ajuste del modelo con interacción ---


modelo <- lm(calidad ~ dias * lluvia, data = datos)
summary(modelo)
```

```
##
## Call:
## lm(formula = calidad ~ dias * lluvia, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6833 -0.5703  0.1265  0.4385  1.6354
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.16122    0.68917   7.489 3.95e-09 ***
## dias                  -0.03145    0.01760  -1.787   0.0816 .
## lluviaCon lluvia       1.78670    1.31740   1.356   0.1826
## dias:lluviaCon lluvia -0.08314    0.03160  -2.631   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7578 on 40 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6612
## F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10
```

```
# Extraer coeficientes
coeficientes <- coef(modelo)
beta1 <- coeficientes["dias"]
beta3 <- coeficientes["dias:lluviaCon lluvia"]

# Calcular la pendiente para el grupo con lluvia
pendiente_lluvia <- beta1 + beta3

# Calcular el número de días para que la calidad disminuya en 1 unidad en presencia de lluvia
d_est <- -1 / pendiente_lluvia

# --- Cálculo del error estándar para el intervalo de confianza ---
vcov_mat <- vcov(modelo)
var_beta1 <- vcov_mat["dias", "dias"]
var_beta3 <- vcov_mat["dias:lluviaCon lluvia", "dias:lluviaCon lluvia"]
cov_beta1_beta3 <- vcov_mat["dias", "dias:lluviaCon lluvia"]

# Varianza de (beta1+beta3)
var_t <- var_beta1 + var_beta3 + 2 * cov_beta1_beta3
se_t <- sqrt(var_t)

# Método delta para calcular el error estándar de d = -1/t
se_d <- se_t / (pendiente_lluvia^2)

# Intervalo de confianza del 95%
z <- qnorm(0.975)
LI <- d_est - z * se_d
LS <- d_est + z * se_d

# Presentar resultados con knitr

resultado <- data.frame(
  Estimación = d_est,
  `Límite Inferior` = LI,
  `Límite Superior` = LS
)
kable(resultado, caption = "Intervalo de confianza para la disminución de 1 unidad en calidad con lluvia")
```

Intervalo de confianza para la disminución de 1 unidad en calidad con lluvia

	Estimación	Límite.Inferior	Límite.Superior

	Estimación	Límite.Inferior	Límite.Superior
dias	8.727273	4.809318	12.64523

Analisis de resultados:

a. gráfica:

A primera vista podemo notar que para los dos grupos la calidad de la cosecha disminuye a medida que transcurren los dias desde el inicio de la cosecha “31 de agosto”.

Pero a tomar en cuenta:

1. En el grupo de sin lluvia que está representado en rojo, la pendiente es relativamente es baja egun transcurren los dias o sea si hay una disminución de la calidad pero en terminos generales de una baja cantidad, sin tomar en cuenta los outliers. Ahora en el grupo de la lluvia representada en azul, la pendiente es mucho mayor y con lo cual entre mas tarde inicie la cosecha a partir del “31 de agosto” el impacto sobre la calidad del vino es bastante notable.
2. Podemos observar que la dispersion de los datos en ambos grupos (con y sin lluvia) eso nos puede indicar que definitivamente la fecha de finalizacion de la cosecha **No Es El Único Factor Que Explica La Calidad**, aún asi la tendencia general que se muestra en la líneas muestra una relación negativa entre el retraso de la cosecha y la calidad del producto final.
3. He sombreado en un área el intervalo o rango de calidad en torno a la dispersion de los datos, esto nos permite apreciar claramente que las cosechas que finalizan en fechas tardías son significativamente de peor calidad que aquellas en el mismo tiempo pero en temporada sin lluvias.
4. En términos generales la gráfica refuerza la hipótesis de que retrasar la cosecha puede llegar a disminuir la calidad el vino y se remarca más aún cuando las condiciones climáticas son adversas (me refiero a las lluvias, porque este estudio no incluye huracanes, nevadas, tornados etc).

En definitiva planificar la cosecha de forma que se minimice la insidencia ó el riesgo de las lluvias ayudará a mejorar una calida mas alta en el producto final.

b. modelo:

el modelo de la cosecha está en funcion de las dos variables:

- “días”: desde el 31 de agosto hasta la finalizacion de la cosecha.
- “lluvia”: es una variable categórica que está convertida a factor e indica si hubo llubia o nó durante la cosecha.
- Interacción: este término permite que la pendiente (el transcurrir de los dias) varíe según la presencia o ausencia de lluvia.
- Este modelo parece adecuado en los términos de ajuste porque el **R²** ajustado es de 0.6612 y esto indica que por lo menos el 66% de la variabilidad en la calidad del vino se puede explicar por la presencia de lluvia, además el **p-valor** de todo el modelo es realmente bajo = 4.017×10^{-10} y esto nos sugiere que el modelo es significativo.

Entonces:

- $\beta_0 = 5.16122$ Intercepto del grupo de referencia o sin lluvia, paara una cosecha sin lluvia y siendo $dias \approx 0$ y aunque realmente no debe existir una cosecha a 0 dias, este valor nos va ha servir como una referencia y construir la ecuación.
- $\beta_1 = -0.03145$ Efecto de cada dia adicional en la calidad del grupo sin lluvia, en el grupo sin lluvia la calidad en promedio disminuye 0.031 unidades.
- $\beta_2 = 1.78970$ Cambio del intercepto ó efecto base de cuando se presentan lluvias, con este coeficiente vemos el cambio en el intercepto para las cosechas con lluvia comparado a las sin lluvia, para 0 dias la calidad de las cosechas con lluvia seria 1.78670u mas que en las cosechas sin lluvia., aunque el p-valor de 0.1826 demuestra que la diferencia no es estadísticamente importante.
- $\beta_3 = -0.08314$ Cambio de la pendiente respecto a los días cuando se presenta lluvia ó sea el efecto de la interacción, este valor es realmente la pendiente respeco en los dias de lluvia se hace negativa,asi que este efecto de los dias de lluvia sobre calidad es $\beta_1 + \beta_3 = -0.03245 + (-0.08314) \approx -0.11459$, lo que significa que cada dia de lluvia adicional en la cosecha causa una reduccion de calidad de 0.115unidades, ete efecto si es estadisticamente significatico ya que $p = 0.0120$

Como es necesario establecer la variable factor es la categoría “sin lluvia”, entonces:

Para el grupo **“sin lluvia”** podemos deducir que la ecuación es:

$$calidad = \beta_0 + \beta_1 \, dias$$

, en este caso β_0 es la calidad estimada cuando los dias = 0 y β_1 es el cambio en la calidad por cada dia adicional.

Para el grupo **“con lluvia”**, en este caso la variable lluvia toma un valor y el modelo queda de esta forma:

$$calidad = (\beta_0 + \beta_1) + (\beta_1 + \beta_3) \, dias$$

y para este caso nos encontramos que el intercepto o l calidad base en el caso que los dias sean igual a cero y hay presencia de lluvias lo definimos como $\beta_0 + \beta_1$, mientras la pendiente del grupo **“con lluvia”** es $\beta_1 + \beta_3$ e indica el cambio en la calidad por cada dia adiciona en presencia e lluvias.

Es de esperar que el coeficiente $\beta_1 + \beta_3$ sea negativo ya que comocemos que la calidad disminuye a medida que el tiempo aumenta, de esta forma el modelo del ejercicio podemos definirlo como:

$$Calidad = \beta_0 + \beta_1\ dias + \beta_2\ lluvia + \beta_3(dias\ x\ lluvia) + \epsilon$$

$$Calidad = (5.16122 + 1.78670) + (-0.03145 - 0.08314)\ dias \approx 6.94792 - 0.11459\ dias)$$

Para determinar el número de dias necesarios para disminuir una unidad la calidad necesitamos conocer si la pendiente $\beta_1 + \beta_3$ aumente d dias, entonces:

$$\Delta\ calidad = (\beta_1 + \beta_3)d$$

$$\Delta\ calidad = (-0.03145 + -0.08314)d$$

Una ves ya definimos la funcion y conociendo que necesitamos conocer el número de dias para un delta de -1 dia, asi que:

$$-1 = (-0.03145 + -0.08314)d\ \Rightarrow\ -\frac{1}{(-0.03145 + -0.08314)}\ \Rightarrow d = -\frac{1}{-0.11459}d = 8.72676498821887$$

y podemos observar rápidamente que $(\beta_1 + \beta_3 = -0.11459) < 0$ el numero de dias es positivo y esto concuerda con la lógica observada del ejercicio.

Para conocer el intervalo de confianza para $d = -\frac{1}{(\beta_1+\beta_3)}$ tendremos que aplicar **el método delta** o sea, que $g(t) = -\frac{1}{t}\ \therefore\ t = \beta_1 + \beta_3 = -0.11459$, entonces, la derivada seria: $g'(t)= = $ y sabemos que $SE(t)$ corresponde al error estándar de t y entonces el error estándar de d se puede aproximar a:

$$SE(d) \approx \frac{SE(t)}{t^2}$$

Para calcular $SE(t)$, se utiliza la matriz de varianzas y covarianzas del modelo;

$$Var(t) = Var(\beta_1) + Var(\beta_3) + 2Cov(\beta_1, \beta 3),\ y\ SE(t) = *\sqrt{\overline{Var(t)}}$$

Situación 2

Sea:

$$Y_1 = \theta + \epsilon_1$$

$$Y_2 = 2\theta - \phi + \epsilon_2$$

$$Y_3 = \theta + 2\phi + \epsilon_3$$

donde $E[\epsilon_i] = 0$ para $i = 1,2,3 $. Encuentre las estimaciones de los parámetros θ, ϕ por mínimos cuadrados.

Solucion

Para poder calcular los estimadores por el método de los minimos cuadrados debemos escribir el sistema matricial:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \theta \\ \phi \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.$$

Ahora definamos que:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \theta \\ \phi \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

Y entones podemos decir que el estimador de minimos cuadrados es:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

1. Calculamos el producto de la transpuesta de la matriz X por la matriz X, $X^\top X$

$$X^\top = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix}$$

$$X^\top X = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} = \left(\begin{pmatrix} 1^2 + 2^2 + 1^2 \\ 0 \cdot 1 + (-1) \cdot 2 + 2 \cdot 1 \end{pmatrix} \begin{pmatrix} 1 \cdot 0 + 2 \cdot (-1) + 1 \cdot 2 \\ 0^2 + (-1)^2 + 2^2 \end{pmatrix} \right) \Rightarrow \left(\begin{pmatrix} 1 + 4 + 1 \\ 0 + -2 + 2 \end{pmatrix} \begin{pmatrix} 0 + -2 + 2 \\ 0 + 1 + 4 \end{pmatrix} \right) \Rightarrow$$

2. Cálculo de $(X^\top X)^{-1}$

$$(X^\top X)^{-1} = \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{5} \end{pmatrix}$$

3. Cálculo de $(X^\top Y)$, sea:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \Rightarrow X^\top Y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} Y_1 + 2Y_2 + Y_3 \\ -Y_2 + 2Y_3 \end{pmatrix}$$

4. Calular el estimador $\hat{\beta}$

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} Y_1 + 2Y_2 + Y_3 \\ -Y_2 + 2Y_3 \end{pmatrix} = \begin{pmatrix} \frac{Y_1 + 2Y_2 + Y_3}{6} \\ \frac{-Y_2 + 2Y_3}{5} \end{pmatrix}$$

Entonces, las estimaciones por mínimos cuadrados son:

- Estimacion de θ

$$\hat{\theta} = \frac{Y_1 + 2Y_2 + Y_3}{6}$$

- Estimacion de ϕ

$$\phi = \frac{-Y_2 + 2Y_3}{5}$$

Análisis

La situación propuesta es un conjunto de modelos que corresponden a las ecuaciones de medición para las observaciones $Y_1, Y_2 \; y \; Y_3$, no presenta grupos o variables categóricas, preenta 3 ecuaciones que pueden relacionar dos parámetros $\theta \; y \; \phi$ con varias combinaciones:

- $Y_1 = \theta + \epsilon_1, , \quad \therefore \quad Y_1$ es una medida directa de θ mas el error.
- $Y_2 = 2\theta - \phi + \epsilon_2, \quad \therefore \quad Y_2$ es una combinación lineal de $\theta \; y \; \phi$
- $Y_3 = \theta + 2\phi + \epsilon_, \quad \therefore \quad Y_3$ es otra combinación lineal de $\theta \; y \; \phi$

```r
# hacer el test si el resultado es correcto, para ello calculamos las 2 estimaciones usando R, estimamos valores para Y1,Y2 y Y3
Y <- c(Y1 = 10, Y2 = 15, Y3 = 12)

# hagamos la matriz
X <- matrix(c(1, 0,   # Para Y1: theta + 0*phi
              2, -1,  # Para Y2: 2theta - phi
              1, 2),  # Para Y3: theta + 2phi
            nrow = 3, byrow = TRUE)

# Calculemos el estimador usando R
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y

# calculemos las estimaciones de theta y de phi
theta_hat <- unname(beta_hat[1])
phi_hat   <- unname(beta_hat[2])

# comprobemos:
# Calculamos theta manualmente según la fórmula: theta=(Y1+2Y2+Y3)/6
theta_hat_manual <- (Y[1] + 2 * Y[2] + Y[3]) / 6
# Calculamos phi manualmente según la fórmula: (-Y2 + 2*Y3) / 5
phi_hat_manual <- (-Y[2] + 2 * Y[3]) / 5

# Crear una tabla de resultados:
resultados <- data.frame(
  Parámetro = c("θ", "φ"),
  Método = rep("Mínimos cuadrados", 2),
  `Est_R` = c(theta_hat, phi_hat),
  `Est_manual` = c(theta_hat_manual, phi_hat_manual)
)

# Asegurarse de que no existan nombres de fila
rownames(resultados) <- NULL

# Mostrar la tabla con knitr::kable
library(knitr)
kable(resultados, digits = 4,
      caption = "Comparación de estimaciones de θ y φ")
```

Comparación de estimaciones de θ y φ

Parámetro	Método	Est_R	Est_manual
θ	Mínimos cuadrados	8.6667	8.6667
φ	Mínimos cuadrados	1.8000	1.8000

Situación 3

Los datos que siguen corresponden al registro de nuevos casos de melanoma informados en Estados Unidos durante $1969 - 1991$ en hombres blancos clasificados por edad y región. La última columna es el tamaño de la población registrada en el censo de EE.UU.

Región	Edad	Casos	Población
Norte	0-35	61	2880262
Norte	35-44	76	564535
Norte	45-54	98	592983
Norte	55-64	104	450740
Norte	65-74	63	270908
Norte	75+	80	161850
Sur	0-35	64	1074246
Sur	35-44	75	220407
Sur	45-54	68	198119
Sur	55-64	63	134084
Sur	65-74	45	70708
Sur	75+	27	34233

a. Grafique el log de la tasa observada (número de casos sobre tamaño de la población) versus la edad. ¿Qué observa en este gráfico?

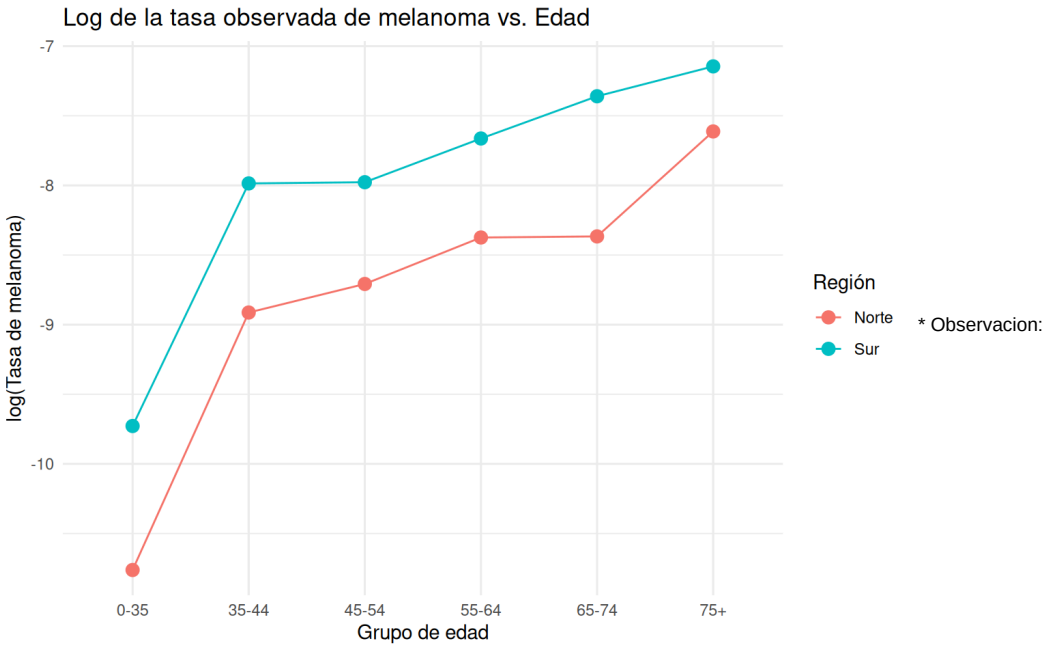
b. Ajuste un modelo de Poisson incorporando un intercepto, una variable dummy para la región, variables dummies para distinguir los grupos de edad (tome como referencia el grupo de menor edad) y el **offset** que le parezca adecuado. ¿El ajuste parece adecuado?

c. Verifique si la función de varianza es adecuada. ¿Qué puede estar ocurriendo? ¿Le parece que el problema puede ser solucionado utilizando un parámetro de escala o cambiando a un modelo binomial-negativo?

Solucion:

A: Gráfica:

```
# creamos los vectores para los datos
datos <- data.frame(
  Región = rep(c("Norte", "Sur"), each = 6),
  Edad = rep(c("0-35", "35-44", "45-54", "55-64", "65-74", "75+"), times = 2),
  Casos = c(61, 76, 98, 104, 63, 80,
            64, 75, 68, 63, 45, 27),
  Población = c(2880262, 564535, 592983, 450740, 270908, 161850,
                1074246, 220407, 198119, 134084, 70708, 34233)
)

# Hagamos el cálculo de tasa y logaritmo
datos$Tasa <- datos$Casos / datos$Población
datos$logTasa <- log(datos$Tasa)

# creamos el factor de edad para conservar el orden,
datos$Edad <- factor(datos$Edad, levels = c("0-35", "35-44", "45-54", "55-64", "65-74", "75+"))

# creamos la gGrafica para el logTasa vs Edad diferenciados por colores

library(ggplot2)
ggplot(datos, aes(x = Edad, y = logTasa, group = Región, color = Región)) +
  geom_point(size = 3) +
  geom_line(aes(group = Región)) +
  labs(title = "Log de la tasa observada de melanoma vs. Edad",
       x = "Grupo de edad",
       y = "log(Tasa de melanoma)") +
  theme_minimal()
```



En la gŕafica podemos observar como el log de la tasa ó la tasa observada en una escala logaritmica, varía con la edad.

Lo observado coincide con la hipótesis común de que el riesgo a contraer melanoma aumenta con la edad.

Podemos observar que aunque la tendencia es similar si hay una diferencia entre la region norte y la región sur lo cual tambien coincide cn la hipótesis comun de la relacion entre el melanoma y el nivel de exposición solar, es de esperar que en la region norte de estados unidos el nivel de exposición solar es menor debido al cambio de estación mientras en el sur el impacto de estos cambios es menor y por ende el riesgo a melanoma es mayor.

B: Ajuste del modelo de Poisson Vamos a ajustar un modelo de Poisson para poder explicar el número de casos:

- usamos un intercepto.

- usamos una variable dummy para la región (por ejemplo para usar el norte como referencia).
- usamos otra variable dummy para los grupos de edad y tomamos al grupo entre $0 - 35$ como referencia.
- creamos un **offset** logaritmico de la población (tasa = casos/poblacion)

La ecuacion de un modelo como este es:

$$log(E[casos]) = \beta_0 + \beta_1 I(region = sur) + \sum_{j=2}^{6} \beta_j I(Edad = grupo_j + log(población)$$

```
# llevamos la region al factor y aseguramos un orden con el Norte como referencia
datos$Región <- factor(datos$Región, levels = c("Norte", "Sur"))

# asjustemos el modelo de poisson con un offset(poblacion)
modelo <- glm(Casos ~ Región + Edad,
              offset = log(Población),
              family = poisson(link = "log"),
              data = datos)
summary(modelo)
```

```
##
## Call:
## glm(formula = Casos ~ Región + Edad, family = poisson(link = "log"),
##     data = datos, offset = log(Población))
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.65831    0.09518 -111.97   <2e-16 ***
## RegiónSur     0.81948    0.07103   11.54   <2e-16 ***
## Edad35-44     1.79737    0.12093   14.86   <2e-16 ***
## Edad45-54     1.91309    0.11844   16.15   <2e-16 ***
## Edad55-64     2.24180    0.11834   18.94   <2e-16 ***
## Edad65-74     2.36572    0.13152   17.99   <2e-16 ***
## Edad75+       2.94468    0.13205   22.30   <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: 895.8197  on 11  degrees of freedom
## Residual deviance:   6.2149  on  5  degrees of freedom
## AIC: 92.44
##
## Number of Fisher Scoring iterations: 4
```

Análisis del modelo:

El modelo Poisson parece adecuado porque los coeficientes son significativos (todos los p-valor son menores a (0.05), y la desviación residual es baja (6.2149) y esto indica unbuen ajuste y adicional, tambien el AIC (Criterio de Información de Akaike) es relativamente bajo (92.44) y esto tambien nos sugiere que el modelo es eficiente.

No se observa en el modelo una significativa sobre-dispersión y esto definitivamente indica que el modelo de Poisson es adecuado para el caso.

EN el caso que la dispersión fuera mayor a 1 podriamos considerar usar otros modelos como quasi-Poisson o binomial-negativo.

Tambien observamos en el modelo que la región Sur tiene una tasa de insidencia de melanoma significativamente mayor que la region Norte, a una razón de ≈ 2.27, lo que coincide con lo observado en la gráfica.

Tambien encontramos que hay un efecto positivo y creciente con la edad ya que a medida que los grupos de edad aumentan el riesgo del melanoma tambien aumenta considerablemente en comparación con el grupo de $0 - 35$ años ya que los grupos de mayor edad va ha tener log-tasas mas elevadas de riesgo a melanoma.

C: Verificamos a varianza y la presuncion de posible sobre-dispersión.

Analizando el modelo de Poisson se asume que la varianza es igual a la media, si obtenemos que el coeficiente de la desviación por los grados de libertad ó desviación es mucho mayor que 1 nos indicará que puede existir sobre-dispersión.

```
# Calculo de la rata de la desviación residual a los grados de libertad
dispersion <- sum(residuals(modelo, type = "pearson")^2) / modelo$df.residual
cat("Ratio de dispersión =", dispersion, "\n")
```

```
## Ratio de dispersión = 1.223019
```

Análisis:

El ratio obtenido es 1.223019 que es un poco mayor a 1, podemos considerar que existe una sobre-dispersión moderada y el modelo de Poisson con este bajo nivel de dispersión funciona bien y sigue siendo adecuado, aunque por prudencia siempre es adecuado revisar los gráficos de residuos y otros diagnósticos adicionales para confirmar que el modelo no presenta otros problemas, tambien es conveniente ajustar otros modelos tal como el quasi-poisson o el binomial-negativo si se considera que la sobre-dispersion existente afecta las inferencias (error estándar, intervalos de confianza, etc.)

Situación 4

En un estudio se desean investigar los factores que influencian la elección de alimento primario de los caimanes. El estudio se basa en la información obtenida de 219 caimanes capturados en cuatro lagos de Florida. La variable respuesta de escala nominal es el tipo primario de alimento, en volumen, encontrado al interior del estómago de un caimán. En la tabla a continuación se presenta la clasificación de la elección de alimento primario según lago y tamaño del caimán.

Lago	Tamaño (m)	Peces	Invertebrados	Reptiles	Aves	Otros
Hancock	≤ 2.3	23	4	2	2	8
Hancock	> 2.3	7	0	1	3	5
Ocklawaha	≤ 2.3	5	11	1	0	3
Ocklawaha	> 2.3	13	8	6	1	0
Trafford	≤ 2.3	5	11	2	1	5
Trafford	> 2.3	8	7	6	3	5
George	≤ 2.3	16	19	1	2	3
George	> 2.3	17	1	0	1	3

a. Fije la categoría de referencia en: Peces. Ajuste el modelo.
b. Calcule las razones de chances para los lagos y el tamaño.
c. Presente de manera ordenada dichas razones en una tabla.
d. ¿Tiene algún efecto el tamaño del caimán sobre la elección del alimento primario?. Justifique e interprete los posibles efectos en términos de razones de chances.Interprete el efecto del tamaño del caimán sobre la elección primaria de alimentos invertebrados.
e. ¿Tiene algún efecto el lago sobre la elección del alimento primario? Justi que e interprete los posibles efectos en términos de razones de chances.

Entendiendo el problema:

Tenemos la información consolidadade 219 caimanes de cuatro lagos de Florida US, la variable de respuesta representa al alimento primario encontrado en el estomago de los sujetos de estudio, este se clasifica por su volumen en 5 categorías:
$Peces, Invertebrados, Reptiles, AvesyOtrasespecies$ y contamos con la información de dos variables explicativas o predictoras:

- **Lago**: Los sujeos de estudio fueron encontrados en los lagos: **Hancock, Ocklawaha, Trafford, George**
- **Tamaño del caimán**: los sujetos estan clasificados en dos grupos dependiendo de la longitud: ≤ 2.3 y ≥ 2.3 metros.

El objetivo es investigar los factores que influyen en la eleccion del alimento primario y eb específico:

- Fijar como categoría de referencia en la respuesta la opción “Peces”.
- Ajustar un modelo multinomial (modelo de regresión logística multinomial) usando dummies para el lago y el tamaño (tomando el grupo de menor tamaño como referencia).
- Calcular las razones de chances (odds ratios) asociadas a cada factor y presentarlas en una tabla.
- Evaluar si el tamaño del caimán tiene efecto sobre la elección y, en particular, interpretar el efecto sobre la elección de alimentos invertebrados.
- Evaluar el efecto del lago sobre la elección del alimento primario.

```
# vamos a usar VGAM, se necesita instalar el paquete en R-studio
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```r
# cramos los vectores para el análisis y configuramos las variables Lago y tamaño como factores, esto le indica a R que cuando lo use en un modelo la librería VGRAM con vglm() cree de forma automática variables dummy para cada uno de susu niveles, similar a un one-hot pero a diferencia dummy descarta una columna y de esta manera los coeficientes se interpretan en comparacion con ese nivel de referencia.

datos <- data.frame(
  Lago = rep(c("Hancock", "Ocklawaha", "Trafford", "George"), each = 2),
  Tamaño = rep(c("≤ 2.3", "> 2.3"), times = 4),
  Peces = c(23, 7, 5, 13, 5, 8, 16, 17),
  Invertebrados = c(4, 0, 11, 8, 11, 7, 19, 1),
  Reptiles = c(2, 1, 1, 6, 2, 6, 1, 0),
  Aves = c(2, 3, 0, 1, 1, 3, 2, 1),
  Otros = c(8, 5, 3, 0, 5, 5, 3, 3)
)

# Aseguremonos de reconfigurar el factor de Tamaño para que "≤ 2.3" sea el nivel de referencia como lo solicita el ejercicio
datos$Tamaño <- factor(datos$Tamaño, levels = c("≤ 2.3", "> 2.3"))

# A.
#Ajustamos el modelo utilizando VGAM vglm(), en la función multinomial() y fijamos la categoría de referencia para la respuesta solicitada y para ello usamos la primera columna cbind que corresponde a los Peces como se pide en el ejercicio.
# vamos a obtener los coeficientes para cada caegoria en función de las variables predictoras exceptuando a los Peces que es la referencia.

modelo <- vglm(cbind(Peces, Invertebrados, Reptiles, Aves, Otros) ~ Lago + Tamaño,
               family = multinomial(refLevel = 1),  # refLevel = 1 esto nos indica que "Peces" es la variable de referencia
               data = datos)

# Resumen del modelo creado
summary(modelo)
```

```
##
## Call:
## vglm(formula = cbind(Peces, Invertebrados, Reptiles, Aves, Otros) ~
##     Lago + Tamaño, family = multinomial(refLevel = 1), data = datos)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   -0.090814   0.308039  -0.295 0.768136
## (Intercept):2   -3.665796   1.058973  -3.462 0.000537 ***
## (Intercept):3   -2.723736   0.710395  -3.834 0.000126 ***
## (Intercept):4   -1.572721   0.474821  -3.312 0.000926 ***
## LagoHancock:1   -1.658359   0.612877  -2.706 0.006813 **
## LagoHancock:2    1.242777   1.185432   1.048 0.294466
## LagoHancock:3    0.695118   0.781263   0.890 0.373608
## LagoHancock:4    0.826196   0.557540   1.482 0.138378
## LagoOcklawaha:1  0.937219   0.471906   1.986 0.047030 *
## LagoOcklawaha:2  2.458872   1.118128   2.199 0.027871 *
## LagoOcklawaha:3 -0.653208   1.202098  -0.543 0.586861
## LagoOcklawaha:4  0.005653   0.776513   0.007 0.994191
## LagoTrafford:1   1.121985   0.490513   2.287 0.022174 *
## LagoTrafford:2   2.935253   1.116409   2.629 0.008559 **
## LagoTrafford:3   1.087767   0.841669   1.292 0.196221
## LagoTrafford:4   1.516369   0.621435   2.440 0.014683 *
## Tamaño> 2.3:1   -1.458205   0.395945  -3.683 0.000231 ***
## Tamaño> 2.3:2    0.351263   0.580033   0.606 0.544786
## Tamaño> 2.3:3    0.630660   0.642480   0.982 0.326296
## Tamaño> 2.3:4   -0.331550   0.448247  -0.740 0.459506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1]),
## log(mu[,4]/mu[,1]), log(mu[,5]/mu[,1])
##
## Residual deviance: 17.0798 on 12 degrees of freedom
##
## Log-likelihood: -47.5138 on 12 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', '(Intercept):3'
##
##
## Reference group is level  1  of the response
```

```
print('coeficientes del modelo')
```

```
## [1] "coeficientes del modelo"
```

```
str(coef(modelo))
```

```
##  Named num [1:20] -0.0908 -3.6658 -2.7237 -1.5727 -1.6584 ...
##  - attr(*, "names")= chr [1:20] "(Intercept):1" "(Intercept):2" "(Intercept):3" "(Intercept):4" ...
```

```
# Extraer los coeficientes del modelo
coef_model <- coef(modelo)
# Ver la estructura para confirmar
str(coef_model)
```

```
##  Named num [1:20] -0.0908 -3.6658 -2.7237 -1.5727 -1.6584 ...
##  - attr(*, "names")= chr [1:20] "(Intercept):1" "(Intercept):2" "(Intercept):3" "(Intercept):4" ...
```

```
# Los nombres tienen la forma "(Predictor):(Nivel)".
# Por ejemplo: "(Intercept):1", "LagoHancock:1", "Tamaño≤ 2.3:1", etc.
coef_names <- names(coef_model)

# Dividir los nombres en dos partes utilizando ':' como separador
split_names <- strsplit(coef_names, split = ":")

# Extraer la primera parte (Predictor) y la segunda parte (Nivel de respuesta)
Predictor <- sapply(split_names, `[`, 1)
NivelRespuesta <- sapply(split_names, `[`, 2)

# Crear un data frame con la información extraída y los coeficientes
resultados <- data.frame(
  Predictor = Predictor,
  NivelRespuesta = NivelRespuesta,
  Estimate = coef_model,
  OR = exp(coef_model)
)

# Quitar nombres de fila para que no aparezcan etiquetas no deseadas
rownames(resultados) <- NULL

# Mostrar la tabla usando knitr::kable
library(knitr)
kable(resultados, digits = 4,
      caption = "Razones de chances (Odds Ratios) para cada categoría de alimento (referencia: Peces)")
```

Razones de chances (Odds Ratios) para cada categoría de alimento (referencia: Peces)

Predictor	NivelRespuesta	Estimate	OR
(Intercept)	1	-0.0908	0.9132
(Intercept)	2	-3.6658	0.0256
(Intercept)	3	-2.7237	0.0656
(Intercept)	4	-1.5727	0.2075
LagoHancock	1	-1.6584	0.1905
LagoHancock	2	1.2428	3.4652
LagoHancock	3	0.6951	2.0039
LagoHancock	4	0.8262	2.2846
LagoOcklawaha	1	0.9372	2.5529
LagoOcklawaha	2	2.4589	11.6916
LagoOcklawaha	3	-0.6532	0.5204
LagoOcklawaha	4	0.0057	1.0057
LagoTrafford	1	1.1220	3.0709
LagoTrafford	2	2.9353	18.8263
LagoTrafford	3	1.0878	2.9676
LagoTrafford	4	1.5164	4.5557
Tamaño> 2.3	1	-1.4582	0.2327
Tamaño> 2.3	2	0.3513	1.4209
Tamaño> 2.3	3	0.6307	1.8788
Tamaño> 2.3	4	-0.3316	0.7178

Analisis

Sobre el modelo

> modelo <- vglm(cbind(Peces, Invertebrados, Reptiles, Aves, Otros) ~ Lago + Tamaño, family = multinomial(refLevel = 1), data = datos)

El modelo está ajustado para formar una matriz en el siguiente orden: 1. Peces, 2. Invertebrados, 3. Reptiles, 4. Aves, 5. Otros

Al usar en el modelo el *reflevel = 1*, fijamos como categoría de referencia a la primera serie que es nuestro caso es "*Peces*", esta funciona como la referencia y las demás categorias se van ha comparar contra esta, entonces las ecuaciones lineales que se van ha usar el modelo corresponde a las siguientes:

Ecuacion lineal	Categoría	Coeficientes
$log(\frac{\mu_2}{\mu_1})$	Invertebrados	Nivel 1
$log(\frac{\mu_3}{\mu_1})$	Reptiles	Nivel 2
$log(\frac{\mu_4}{\mu_1})$	Aves	Nivel 3
$log(\frac{\mu_5}{\mu_1})$	Otros	Nivel 4

$\therefore$ μ es el valor esperado o la media de la categoria correspondiente

Por tanto el modelo está estimando los logaritmos de las razones de las medias esperadas de cada categoría frente a la referencia (Peces), asi que el modelo compara el logaritmo de la razón de probabilidades (log-dds) de cada categoría con la categoría de referencia ó base.

La matriz μ representada en la tabla contiene las probabilidades (en el caso de datos de conteo los valores esperados), de que se observe en cada categoría dads las variables predictoras y el **offset** si lo hay. Básicamente los μ describen como se relacionan las variables predictoras con la probabilidad relativa ó la cuenta esperada de cada resultado.

```
# Verificamos el orden en el modelo

# hacemos una matriz de respuesta con cbind() para ver el orden
response_matrix <- cbind(Peces = datos$Peces,
                         Invertebrados = datos$Invertebrados,
                         Reptiles = datos$Reptiles,
                         Aves = datos$Aves,
                         Otros = datos$Otros)

# Extraemos los nombres de las columnas de la matriz de respuesta
response_names <- colnames(response_matrix)

# Imprimimimos los nombres para confirmar el orden
print(response_names)
```

```
## [1] "Peces"         "Invertebrados" "Reptiles"      "Aves"
## [5] "Otros"
```

Sobre lo que nos muestra la tabla de razones

En la tabla de Razones de chances (Odds Ratios) para cada categoría de alimento (referencia: Peces), el coeficiente estimado (en la escala de log-odds) podemos observar el oeficiente estimado (en la escala de log-odds) y la razón de oportunidades (OR que significa la exponencial del coeficiente).

Encontramos que cada fila nos indica cuanto es el nivel de oportunidad al aumentar en 1 unidad en el predictor y el logaritmo de la razón de probabilidades de elegir ese alimento especifico en vez de peces.

Intercepto:

Las filas de intercepto corresponden a los log-odds base para cada una de las categorías de alimentos cuando las variables predictoras (Lago y tamaño) estan al nivel de la referencia.

Para la categoría *"Nivel de respuesta = 1"* el estimado es -0.0908 y el OR es 0.9132 y esto nos indica que en el grupo de referencia (Lago y tamaño), la razón de oportunidad de elegir ese alimento en vez de peces es menor a 1.

Efecto del lago en la elección:

Los coeficientes que se muestran para "*LagoHancock*, *LagoOcklawaha* y el *LagoTrafford*" nos indican como varía la probabilidad relativa de elegir cada categoría de alimento (a comparación de los peces) en el respectivo lago respecto al lago de referencia (el nivel omitido del factor lago).

Por ejemplo, para "*LagoHancock = 1*" el estimado es -1.6584 y el *OR* es 0.1905, lo que sugiere que, para la categoría de alimento correspondiente al "*NivelRespuesta 1*", los caimanes en Hancock tienen odds de elegir ese alimento (en lugar de Peces) aproximadamente 81% menores que en el lago de referencia.

En contraste, "*LagoHancock = 2*" tiene un estimado de 1.2428 y un OR de 3.4652, lo que implica que para esa categoría el "*NivelRespuesta 2*" la probabilidad relativa aumenta más de tres veces en *Hancock*, en comparación con el lago de referencia.

Efecto del Tamaño:

La variable "*Tamaño > 2.3*" compara el grupo de caimanes de mayor tamaño con el grupo de referencia (que, tras la reconfiguración, es "≤ *2.3*").

Por ejemplo, para "*Tamaño > 2.3:1*" el estimado es -1.4582 y el OR es 0.2327. Esto indica que, para la categoría "*NivelRespuesta 1*", los caimánes de mayor tamaño se asocia con odds aproximadamente un 77% menores de elegir ese tipo de alimento (en comparación con Peces) respecto a los caimanes pequeños.

Para las demás categorías (*Niveles 2, 3 y 4*), los efectos varían y, en algunos casos, el OR es mayor que 1 o menor que 1, lo que indica un efecto de aumentar o disminuir las probabilidades relativas en comparación con el grupo de menor tamaño.

hagamos la interpretación específica para alimentos invertebrados:

Ya como comprobamos y sabemos que el "*NivelRespuesta = 1*" corresponde a Invertebrados, el hecho de que "*Tamaño > 2.3:1*" tenga un OR de *0.2327* sugiere que, en los caimanes de mayor tamaño, **la probabilidad de que el alimento primario sea Invertebrados (en lugar de Peces) es mucho menor que en caimanes pequeños**.

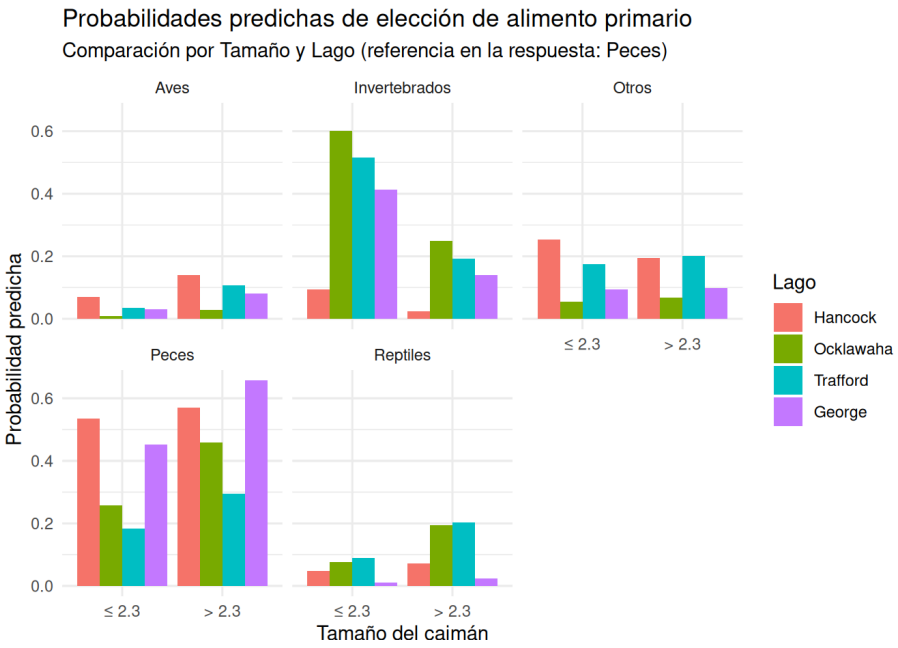
Además, los efectos de Lago varían: por ejemplo, "*LagoOcklawaha = 1*" tiene un OR de *2.5529*, lo que sugiere que, en ese lago, la probabilidad relativa de elegir Invertebrados es mayor, mientras que "*LagoHancock = 1*" tiene un OR menor (0.1905), lo que indica un efecto contrario.

Miremos gráficamente:

```
# Creamos manualmente el data frame de nuevas combinaciones
newdata <- expand.grid(
  Tamaño = c("≤ 2.3", "> 2.3"),
  Lago = c("Hancock", "Ocklawaha", "Trafford", "George")
)

# Convertimos las variables a factor con el orden deseado
newdata$Tamaño <- factor(newdata$Tamaño, levels = c("≤ 2.3", "> 2.3"))
newdata$Lago   <- factor(newdata$Lago, levels = c("Hancock", "Ocklawaha", "Trafford", "George"))

# Comprobamos newdata
print(newdata)
```

```
##   Tamaño      Lago
## 1  ≤ 2.3   Hancock
## 2  > 2.3   Hancock
## 3  ≤ 2.3 Ocklawaha
## 4  > 2.3 Ocklawaha
## 5  ≤ 2.3  Trafford
## 6  > 2.3  Trafford
## 7  ≤ 2.3    George
## 8  > 2.3    George
```

```
# Vamos a usar newdata para hacer unas predecciones
pred <- predict(modelo, newdata = newdata, type = "response")

# La matriz 'pred' tiene 5 columnas correspondientes a las categorías de alimento
newdata$Peces <- pred[,1]
newdata$Invertebrados <- pred[,2]
newdata$Reptiles <- pred[,3]
newdata$Aves <- pred[,4]
newdata$Otros <- pred[,5]

# Convertimos a formato largo para graficar
library(tidyr)
newdata_long <- pivot_longer(newdata,
                             cols = c("Peces", "Invertebrados", "Reptiles", "Aves", "Otros"),
                             names_to = "Alimento",
                             values_to = "Probabilidad")

# Ahora el Gráfico: Probabilidades predichas para cada categoría, comparando Tamaño y diferenciando por Lago
library(ggplot2)
ggplot(newdata_long, aes(x = Tamaño, y = Probabilidad, fill = Lago)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Alimento) +
  labs(title = "Probabilidades predichas de elección de alimento primario",
       subtitle = "Comparación por Tamaño y Lago (referencia en la respuesta: Peces)",
       x = "Tamaño del caimán",
       y = "Probabilidad predicha",
       fill = "Lago") +
  theme_minimal()
```



Podemos observar que para varias categorias y en especial para Invertebrados y Peces se ve que el grupo de *caimanes ≤ 2.3m* tiende a mostrar probabilidades diferentes a su contraparte los *caimanes > 2.3m*, por ejemplo en la categoría de Invertebrados, los caimanes de menor tamaño tienen altas probabilidades en os lagos de *Trafford y George* de consumirlos en comparación de caimanes de mayor tamaño.

Tambien observamos que dentro de la categoría de alimentos, los colores nos indican que entre lagos la probabilidad predicha varia bastante, en los peces por ejemplo en los lagos "*Hancock (en azul) y Ocklawaha (en naranja)*" podemos tener mas probabilidades ó en comparacion d elos lagos "*Trafford (en verde) y George (en violeta)*" mas bajas ó mas altas según el tamaño del caiman.

Sobre los invertebrados, se aprecia que "*Trafford (en verde) y George (en violeta)*" se presentan probabilidades mucho mas altas en el grupo de caimanes con tamaño ≤ 2.3m que en otros lagos.

Al respecto de las **Aves**, los **Reptiles** y **Otros** en general tienen probabilidades bajas de ser la dieta primaria, mientras los **Peces** y los **Invertebrados** tienen las probabilidades mas altas aunque tambien depende del tamaño.

Ahora si se puede observar que el **tamaño del caimán** influye en la elección e ciertos tipos de alimentos, por ejemplo en la categoría de *invertebrados* donde es más porbable en caimanes pequeños (≤ 2.3m) exceptuando el lago *Hancock*.

El **Lago** tambien juega un papel importante ya que hay lagos donde los **peces** siguen siendo el alimento más probable sin importar el tamaño, mientras que se observa que en otros hay un aumento notable en invertebrados e incluso otros reptiles.

Conclusiones generales

Siguiendo la lógica los caimanes de menor tamaño tienden a tener mayor probabilidad de incluir a los invertebrados como alimento principal en algunos lagos como Trafford y George.

La eleccion de peces como alimento varía segun el lago y el tamaño aunque en algunos lagos como Hancock y Ocklawaha se mantiene una alta probabilidad para peces y en especial para caimanes de mayor tamaño.

Las categorías Aves, Reptiles y Otros muestran probabilidades menores o estables, con unas pequeñas diferencias dependiendo del lago y el tamaño

Si tomamos en cuenta que la dieta principal de todos los caimanes son los peces podremos entender el estado de salud del lago respecto a la poblacion de caimanes, ya que la lucha por recursos entre estos causará que los caimanes más débiles ó mas pequeños se vean obligados a incluir en su dieta especies que no estan habitualmente en su dieta tal como aves, invertebrados, reptiles u otras especies, esto peude causarse por una población mayor de caimanes y/o una población menor de peces.

Situación 5

Sea el modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \ldots, n),$$

Dónde $E[\epsilon] = 0$ *y* $Var[\epsilon] = \sigma^2 I_n$.

Encuentre los estimadores por mínimos cuadrados de β_0 y β_1. Demuestre que estos estimadores son no correlacionados si y solo si $\overline{x} = 0$.

Solución

- Definir los estimadores de minimos cuadrados para este modelo:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \ldots, n),$$

con $E[\epsilon] = 0$ y $Var[\epsilon] = \sigma^2 I_n$, los estimadores por mínimos cuadrados o MC son:

1. Estimador de β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\therefore \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad y \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

2. Estimador de β_0:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

- Demostrar la No Correlación entre $\hat{\beta}_0$ y $\hat{\beta}_1$

Para poder entender la correlación enre $\hat{\beta}_0$ y $\hat{\beta}_1$ devemos obtener la covarianza $Cov(\hat{\beta}_0, \hat{\beta}_1)$:

1. podemos expresar $\hat{\beta}_1$ en funcion del error, recordemos que $Y_i = \beta_0 + \beta_1 x_i + \epsilon - i$, asi que:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

2. Ahora igual ara $\hat{\beta}_0$:

$$dado\ que: \qquad \hat{\beta}_0 = \overline{\overline{Y}} - \hat{\beta}_1 \overline{\overline{x}}, \quad y \quad \overline{\overline{Y}} = \beta_0 + \beta_1 \overline{\overline{x}} + \bar{\epsilon} \quad (con\ \bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i)$$

tenemos entonces que:

$$\hat{\beta}_0 = \beta_0 + \beta_1 \overline{\overline{x}} + \bar{\epsilon} - \hat{\beta}_1 \overline{\overline{x}} = \beta_0 + \bar{\epsilon} - \overline{\overline{x}}(\hat{\beta}_1 + \beta_1)$$

Siendo $\hat{\beta}_0$ es entonces que:

$$\hat{\beta}_0 - \beta_0 = \bar{\epsilon} - \overline{\overline{x}}(\hat{\beta}_1 - \beta_1)$$

3. Utilizamos la lienalidad de la covarianza calculemos entonces $Cov(\hat{\beta}_0, \hat{\beta}_1)$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov(\bar{\epsilon} - \overline{\overline{x}}(\hat{\beta}_1 - \beta_1), \hat{\beta}_1 - \beta_1)$$

$$Cov(\bar{\epsilon}, \hat{\beta}_1 - \beta_1) - \overline{\overline{x}} Var(\hat{\beta}_1 - \beta_1)$$

Sabiendo que $E[\epsilon_i] = 0$, por lo que $E[\bar{\epsilon}] = 0$ y que $\bar{\epsilon}$ es la media de los errores y es independiente o no correlacionada con la combinación $\sum(x_i - \overline{x})\epsilon_i$ y ya que x_i no es aleatoria entonces en consecuencia tenemos que:

$$Cov(\bar{\epsilon}, \hat{\beta}_1 - \beta_1) = 0 \Rightarrow Cov(\hat{\beta}_0, \beta_1) = -\overline{\overline{x}} Var(\hat{\beta}_1 - \beta_1)$$

Además, tambien comocemos que:

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_1 - \beta_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \Rightarrow Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{\overline{x}} \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Dónde la covarianza es cero si y solo si $\overline{\overline{x}} = 0$

Resultados finales

- Estimadores:

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(Y_i - \overline{Y})}{\sum(x_i - \overline{x})^2} \quad y \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

* No correlación si y solo si $\overline{\overline{x}} = 0$:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{\overline{x}} \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

```r
# Semilla para que sea reproducible
set.seed(123)

# Número de observaciones simuladas
n <- 100

# Simular x distribucion normal o cualquier otra
x <- rnorm(n, mean = 0, sd = 1)  # Si mean = 0, se cumple \bar{x}=0

# Parámetros verdaderos
beta0 <- 2
beta1 <- 3
sigma <- 1

# Simulamos los errores
epsilon <- rnorm(n, mean = 0, sd = sigma)

# Generarmos y
Y <- beta0 + beta1 * x + epsilon

# Calculamos las medias
x_bar <- mean(x)
Y_bar <- mean(Y)

# los estimadores de mínimos cuadrados
beta1_hat <- sum((x - x_bar) * (Y - Y_bar)) / sum((x - x_bar)^2)
beta0_hat <- Y_bar - beta1_hat * x_bar

cat("Estimación de beta0:", beta0_hat, "\n")
```

```
## Estimación de beta0: 1.897197
```

```r
cat("Estimación de beta1:", beta1_hat, "\n")
```

```
## Estimación de beta1: 2.947528
```

```r
# Calculamos la varianza de beta1_hat (la fórmula teórica)
var_beta1_hat <- sigma^2 / sum((x - x_bar)^2)
cat("Var(beta1_hat):", var_beta1_hat, "\n")
```

```
## Var(beta1_hat): 0.01212267
```

```r
# Calculamos la covarianza entre beta0_hat y beta1_hat teóricamente
cov_beta0_beta1 <- -x_bar * var_beta1_hat
cat("Cov(beta0_hat, beta1_hat):", cov_beta0_beta1, "\n")
```

```
## Cov(beta0_hat, beta1_hat): -0.001095961
```

Análisis:

- Estimación de β_0 ó Intercepto: 1.897197 indica que en promedio cuando $x = 0$ se espera que la funcion corte el eje Y en un valor de aproximadamente de 1.89797
- Estimación de β_1 ó Pendiente: 2.947528 indica que por cada incremento en una unidad en el eje x, esperamos que Y incremente en promedio 2.9475 unidades.
- Estimación de la varianza de $\hat{\beta}_1$: 0.01212267 esta es la varianza teórica del estimador de pendiente, es un valor cercano a cero y sugiere que el estimador $\hat{\beta}_1$ tiene muy poca variabilidad y se estima con bastante precision.
- Estimación de la covarianza $Cov(\hat{\beta}_0, \hat{\beta}_1)$ = -0.001095961 Este valor es la covarianza entre el estiamdor del intercepto y el de la pendiente, ya que es este ejemplo x se generó con la média cercana a cero, esperamos que esta covarianza sea tambien igua a cero. En este ejemplo el valor es muy cercano a cero y esto indica que los estimadores prácticamente son no correlacionados como teóricamente se demuestra cuando $\overline{x} = 0$

Situación 6

Considere la siguiente tabla de $3 \times 4 \times 4$ que resume datos de severidad de los efectos colaterales de una operación, S (1=ninguno, 2=leve, y 3=moderada), tipo de operación, O, y hospital, H . La cuatro operaciones, que corresponden a tratamientos para úlcera duodenal, son numeradas de acuerdo a su grado de invasividad.

Nota: esta tabla la generé usando [https://www.tablesgenerator.com/html_tables# (https://www.tablesgenerator.com/html_tables#)]

	Severidad											
	Hospital 1			Hospital 2			Hospital 3			Hospital 4		
Operación	1	2	3	1	2	3	1	2	3	1	2	3
1	23	7	2	18	6	1	8	6	3	12	9	1
2	23	10	5	18	6	2	12	4	4	15	3	2
3	20	13	5	13	13	2	11	6	2	14	8	3
4	24	10	6	9	15	2	7	7	4	13	6	4

a. Proponga un modelo apropiado para relacionar la severidad de los efectos colaterales y el tipo de operación.
b. De acuerdo al modelo previamente ajustado, encuentre una expresión para $P(S = 1|O), P(S = 2|O)$ y $P(S = 3|O)$ para el tipo de operación 1.
c. Con una significancia del 5%, ¿tiene algunas de las operaciones 1, 2, o 3 un efecto significativo en comparación con la operación más invasiva, 4? Explique.

Entendiendo el ejercicio

En la tabla podemos observar que el estudio cubre 4 hospitales, 4 tipos de operaciones ordenadas por nivel de invasividad y el número de casos observados de efecos colaterales organizados por 3 niveles de severidad donde:

- Severidad 1: No presentan efectos colaterales.
- Severidad 2: Presentan efectos colaterales leves.
- Severidad 3: Presentan efectos colaterales moderados.

Podemos observar segun la tabla:

Operación tipo 1. Es ménos invsiva y los conteos generalmente son altos en la categoría de la severidad 1 y relativamente bajos con severidades 2 y 3. Esto concuerda con la lógica ya que una cirugía menos invasiva tendrá una curva mas rápida de recuperacion y los pacientes presentarán pocos o ningún efecto colateral.

Operación tipo 3 y 4. Observamos que en algunas combinaciones entre hospital y severidad se disminuyen los conteos en severidad 1 y aumentan en los niveles 2 y 3, esto nos sugiere que las operaciones mas invasivas se tienede a estar asociadas a una probabilidad mayor de presentar efectos colaterales de mayor severidad.

Se puede observar que los conteos de casos tambien varías según el hospital, por ejemplo en los hospitales 1 y 2 presentan alta frecuencia en severidad 1 relacionada con la operacion 1 (23,7,2) y (18,6,1) casos, en el hospital 3 se observan es esta misma operacion (8,6 y 3) casos lo que indica que la proporcion es distinta a los otros dos hospitales y en el hospital 4 se observan (12,9,1) casos lo que indica que aunque la mayoría no teien efectos colaterales si hay una proporción relativamente mayor de casos leves.

La tabla nos sugiere que el grado de invasibidad de la operación de 1 a 4 está relacionado con la distribución de la severidad de los efectos colaterales. De forma general, las operaciones las invasivas pueden estar asociadas a una mayor proporción de incidencia de casos con efecto colateral leves o moderados.

Tambien observamos diferencias entre los hospitales con lo cual podemos concluir que existen factores propios de cada centro (protocolos, experiencia e l personal, caracteristicas de la población, nivel de esterilizacion, etc.) que influyen en la severidad de los efectos observados despues de una operacion.

Según se observó la tabla, este modelo puede usar una regresión logística multinomial o similar porque la variable de respuesta es la severidad (nivel 1,2,3) y esta se puede considerar nominal (podria ser ordinal pera para este ejercicio se asume como nominal) y las variables predictoras son el tipo de operación y el hospital.

Podriamos fijar por ejemplo la categoría 1 (ninguna) como referencia y así podriamos obtener las razones de cambio que cuantifiquen como aumenta o disminuye la probabilidad de presentar efectos colaterales leves o moderados en funcion de la operación ó grado de invasividad y el hospital.

```
# Cargar librerías necesarias
library(tidyr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

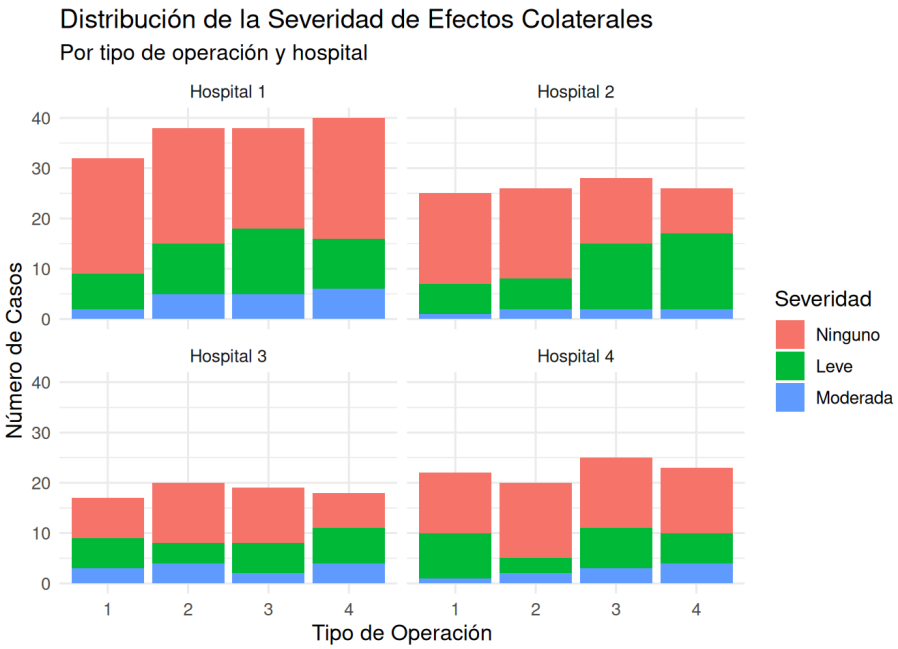
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Crear un data frame con los datos (cada fila corresponde a una combinación de Operación y Hospital)
# Los datos se han extraído de la tabla proporcionada:
datos <- data.frame(
  Operacion = rep(1:4, each = 4),
  Hospital = rep(c("Hospital 1", "Hospital 2", "Hospital 3", "Hospital 4"), times = 4),
  S1 = c(23, 18, 8, 12,   23, 18, 12, 15,    20, 13, 11, 14,    24, 9, 7, 13),
  S2 = c(7, 6, 6, 9,       10, 6, 4, 3,   13, 13, 6, 8,    10, 15, 7, 6),
  S3 = c(2, 1, 3, 1,       5, 2, 4, 2,      5, 2, 2, 3,     6, 2, 4, 4)
)

# Visualizar los datos originales
print(datos)
```

```
##    Operacion   Hospital S1 S2 S3
## 1          1 Hospital 1 23  7  2
## 2          1 Hospital 2 18  6  1
## 3          1 Hospital 3  8  6  3
## 4          1 Hospital 4 12  9  1
## 5          2 Hospital 1 23 10  5
## 6          2 Hospital 2 18  6  2
## 7          2 Hospital 3 12  4  4
## 8          2 Hospital 4 15  3  2
## 9          3 Hospital 1 20 13  5
## 10         3 Hospital 2 13 13  2
## 11         3 Hospital 3 11  6  2
## 12         3 Hospital 4 14  8  3
## 13         4 Hospital 1 24 10  6
## 14         4 Hospital 2  9 15  2
## 15         4 Hospital 3  7  7  4
## 16         4 Hospital 4 13  6  4
```

```
# Convertir el data frame a formato "largo" para poder graficar con ggplot2.
# Esto crea columnas "Severidad" y "Count"
datos_long <- pivot_longer(datos,
                           cols = c("S1", "S2", "S3"),
                           names_to = "Severidad",
                           values_to = "Count")

# Convertir la variable Severidad a factor con etiquetas significativas
datos_long$Severidad <- factor(datos_long$Severidad,
                               levels = c("S1", "S2", "S3"),
                               labels = c("Ninguno", "Leve", "Moderada"))
datos_long$Operacion <- factor(datos_long$Operacion)

# Crear un gráfico de barras apiladas:
ggplot(datos_long, aes(x = Operacion, y = Count, fill = Severidad)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~ Hospital) +
  labs(title = "Distribución de la Severidad de Efectos Colaterales",
       subtitle = "Por tipo de operación y hospital",
       x = "Tipo de Operación",
       y = "Número de Casos",
       fill = "Severidad") +
  theme_minimal() +
  theme(text = element_text(size = 12))
```



Solucion:

DEspues de analizar en profundidad el ejercicio propone relacionar la severidad de los efectos colaterales y el tipo de operación, ahora si deseamos analizar este caso en esta perspectiva seguramente la diferencia entre hospitales no son el foco principal y por tanto es factible oder promediarlas, para ello podemos obtener la media agregada por cada uno para representar el comportamiento global de la severidad por operacion.

Asi simplificamos el analisis al reducir las dimensiones (3x4x4) a una table de severidad por Operación (3x4).

Además que nos ayudará a ajustar un modelo multinomial que relaciones directamente el tipo de operación con la distribucion de severidad sin tener que modelar explicitamente la variable hospital.

entonces tenemos que:

Severidad		1				2				3			
Hospital		1	2	3	4	1	2	3	4	1	2	3	4
Operación	1	23	18	8	12	7	6	6	9	2	1	3	1
	2	23	18	12	15	10	6	4	3	5	2	4	2
	3	20	13	11	14	13	13	6	8	5	2	2	3
	4	24	9	7	13	10	15	7	6	6	2	4	4

- Para la Operación 1:

$$S_1 = 23 + 18 + 8 + 12 = 61, \quad S_2 = 7 + 6 + 6 + 9 = 28, \quad S_3 = 2 + 1 + 3 + 1 = 7$$

* Para la Operación 2:

$$S_1 = 23 + 18 + 12 + 15 = 68, \quad S_2 = 10 + 6 + 4 + 3 = 23, \quad S_3 = 5 + 2 + 4 + 2 = 13$$

* Para la Operación 3:

$$S_1 = 20 + 13 + 11 + 14 = 58, \quad S_2 = 13 + 13 + 6 + 8 = 40, \quad S_3 = 5 + 2 + 2 + 3 = 12$$

* Para la Operación 4:

$$S_1 = 24 + 9 + 7 + 13 = 53, \quad S_2 = 10 + 15 + 7 + 6 = 38, \quad S_3 = 6 + 2 + 4 + 4 = 16$$

* Entonces el nivel de severidad vs el tipo de operación es:

```
# Datos agregados por operación
datos_agg <- data.frame(
  Operacion = factor(1:4),
  S1 = c(61, 68, 58, 53),
  S2 = c(28, 23, 40, 38),
  S3 = c(7, 13, 12, 16)
)
datos_agg
```

```
##   Operacion S1 S2 S3
## 1         1 61 28  7
## 2         2 68 23 13
## 3         3 58 40 12
## 4         4 53 38 16
```

Propuesta de modelo y ajuste multinomial

Propongo para este ejercicio ajustar un modelo de regresión logística multinomial en el que la respuesta es el vector de casosd (conteos)

$$(Y_1, Y_2, Y_3) = (S1, S2, S3)$$

en este caso la única variable predictora es el tipo de operacion, O - En este modelo vamos a fijar $S = 1$ (“ninguno”) como la referencia de la respuesta. - Vamos a reconfigurar la variable Operación para que la operacion 4 que es la mas invasiva sea la referencia.

```
library(nnet)

# La respuesta se construye con cbind() a partir de los conteos de severidad.
Y <- as.matrix(datos_agg[, c("S1", "S2", "S3")])

# Fijamos la operación 4 como referencia
datos_agg$Operacion <- relevel(datos_agg$Operacion, ref = "4")

# Ajustamos el modelo multinomial
modelo_mult <- multinom(Y ~ Operacion, data = datos_agg)
```

```
## # weights:  15 (8 variable)
## initial  value 458.121324
## iter  10 value 382.446270
## final  value 382.271052
## converged
```

```
summary(modelo_mult)
```

```
## Call:
## multinom(formula = Y ~ Operacion, data = datos_agg)
##
## Coefficients:
##    (Intercept) Operacion1 Operacion2  Operacion3
## S2  -0.3327091 -0.4459688 -0.7513067 -0.03885621
## S3  -1.1977026 -0.9672472 -0.4568509 -0.37784409
##
## Std. Errors:
##    (Intercept) Operacion1 Operacion2 Operacion3
## S2   0.2125647  0.3119162  0.3215089  0.2956774
## S3   0.2852504  0.4905264  0.4159283  0.4265480
##
## Residual Deviance: 764.5421
## AIC: 780.5421
```

Probabilidad para la Operacion 1

En un modelo multinomial donde se fija $S = 1$, “*ninguno*” como la categoría de referencia, se modelan las razones logaritmicas:

$$log(\frac{P(S = 2|O = k)}{P(S = 1|O = k)}) = \alpha_2 + \beta_{2,k}$$

$$log(\frac{P(S = 3|O = k)}{P(S = 1|O = k)}) = \alpha_3 + \beta_{3,k}$$

Entonces tenemos que para cualquier operacion $O = k$ el modelo estima:

$$log(\frac{P(S=j|O=k)}{P(S=1|O=k)}) = \alpha_j + \beta_{j,k}, \quad j = 2,3$$

Para la Operacion 1, donde $k = O = 1$:

$$\eta_2 = \alpha_2 + \beta_{2,1} \quad y \quad \eta_3 = \alpha_3 + \beta_{3,1}$$

Ahora usando la regla del softmax para la funcion de enlace, las probabilidades condicionales quedan expresadas como:

$$P(S=1|O=1) = \frac{1}{1+e^{\eta_2}+e^{\eta_3}},$$

$$P(S=2|O=1) = \frac{e^{\eta_2}}{1+e^{\eta_2}+e^{\eta_3}},$$

$$P(S=3|O=1) = \frac{e^{\eta_3}}{1+e^{\eta_2}+e^{\eta_3}},$$

En nuestro caso para obtener η_2 y η_3 en nuestro script de R, debemos sumar los coeficientes correspondientes al intercepto y a "*Operacion_1*" ya que la operación 1 y el predictor dummy es 1.

```
# Extraer y mostrar la matriz de coeficientes del modelo
coef_modelo <- summary(modelo_mult)$coefficients
print(coef_modelo)
```

```
##    (Intercept) Operacion1 Operacion2  Operacion3
## S2  -0.3327091 -0.4459688 -0.7513067 -0.03885621
## S3  -1.1977026 -0.9672472 -0.4568509 -0.37784409
```

```
# Para la operación 1, sumamos:
eta2_op1 <- coef_modelo["S2", "(Intercept)"] + coef_modelo["S2", "Operacion1"]
eta3_op1 <- coef_modelo["S3", "(Intercept)"] + coef_modelo["S3", "Operacion1"]

cat("Para la operación 1:\n")
```

```
## Para la operación 1:
```

```
cat("P(S=1|O=1) = 1 / (1 + exp(", round(eta2_op1, 4), ") + exp(", round(eta3_op1, 4), "))\n")
```

```
## P(S=1|O=1) = 1 / (1 + exp( -0.7787 ) + exp( -2.1649 ))
```

```
cat("P(S=2|O=1) = exp(", round(eta2_op1, 4), ") / (1 + exp(", round(eta2_op1, 4), ") + exp(", round(eta3_op1, 4),
"))\n")
```

```
## P(S=2|O=1) = exp( -0.7787 ) / (1 + exp( -0.7787 ) + exp( -2.1649 ))
```

```
cat("P(S=3|O=1) = exp(", round(eta3_op1, 4), ") / (1 + exp(", round(eta2_op1, 4), ") + exp(", round(eta3_op1, 4),
"))\n")
```

```
## P(S=3|O=1) = exp( -2.1649 ) / (1 + exp( -0.7787 ) + exp( -2.1649 ))
```

Evaluar la significancia

Como deseamos conocer si algunas de las operaciones 1,2 o 3 tiene un impacto significativo (a nivel 5%) en comparación con la operación 4, examinamos los coeficiientes del modelo junto a los errores estándares y si podemos calcular los p-valores.

```
modelo_sum <- summary(modelo_mult)
print(modelo_sum)
```

```
## Call:
## multinom(formula = Y ~ Operacion, data = datos_agg)
##
## Coefficients:
##    (Intercept) Operacion1 Operacion2  Operacion3
## S2  -0.3327091 -0.4459688 -0.7513067 -0.03885621
## S3  -1.1977026 -0.9672472 -0.4568509 -0.37784409
##
## Std. Errors:
##    (Intercept) Operacion1 Operacion2 Operacion3
## S2   0.2125647  0.3119162  0.3215089  0.2956774
## S3   0.2852504  0.4905264  0.4159283  0.4265480
##
## Residual Deviance: 764.5421
## AIC: 780.5421
```

En nuestro caso y segun lo definimos en el enunciado vamos a analizar la significancia usando cada coeficiente (S2 vs S1 y S3 vs S1), utilizaremos el cociente estimado / error estandard (estadístico z) para deducir la significancia que usualmene comparado con un valor crítico de 1.96 en valor absoluto para $\alpha = 0.05$

- Severidad Leve vs ninguno: Para el Intercepto S2:

$$z = -0.3327091/0.2125647 \approx -1.565 \Rightarrow |z| = 1.565 < 1.96 \rightarrow p > 0.05$$

  El intercepto para S2 no es significativamente distinto de cero, lo que indica que en la operación de referencia (Operación 4) la relación log-odds de S2 vs S1 no difiere significativamente con el valor base.

Para la Operacion 1, S2:

$$z = -0.44597/0.31192 \approx -1.431 \Rightarrow |z| = 1.431 < 1.96 \rightarrow p > 0.05$$

Para S2, la diferencia en log-odds entre la Operación 1 y la Operación 4 que es la referencia, no se encuentra que sea estadísticamente significativa.

Para la Operación 2, S2:

$$z = -0.75131/0.32151 \approx -2.337 \Rightarrow |z| = 2.337 > 1.96 \rightarrow p < 0.05$$

Este coeficiente si es estadísticamente significativo al 5% e indica que para S2 (leve) en comparación con S1, la operacion 2 tiene log-odds significativamente más bajos que la operacion 4. La razón de cambio es $OR = exp(-0.75131) \approx 0.472$ y esto nos sugiere que la Operacion 2 reduce la probabilidad relativa de presentar efectos leves (mas que ningun efecto) en comparacion con la Operacion 4.

Para la Operación 3, S2:

$$z = -0.03886/0.29568 \approx -0.131 \Rightarrow |z| = 0.131 < 1.96 \rightarrow p >> 0.05$$

No se detecta ningun efecto significativo para la Operación 3 (en S2) frente a la referencia Operación 4.

Para S3, comparacion de severidad moderada vs ninguno:

Para el Intercepto S3:

$$z = -1.19770/0.28525 \approx -4.198 \Rightarrow |z| = 4.198 > 1.96 \rightarrow p < 0.0001$$

Este intercepto es altamente significativo, para la operación 4 la relacion log-odds para S3 vs S1 es significativamente menor.

Operación 1, S3:

$$z = -0.96725/0.49053 \approx -1.972 \Rightarrow |z| = 1.972 \approx 1.96 \rightarrow papprox 0.048$$

Podemos decir que ese coeficiente es marginalmente significativo (cercano al 5%) y nos indica que para S3 (moderada) la operación 1 presenta log-odds significativamnete más bajos en comparación con la Operación 4. La OR es $exp(-0.96725) \approx 0.38$, lo que implica que la Operacion 1 reduce la probabilidad relativa de efectos moderados frente a ningun efecto y en comparación con la operacion 4.

Operacion 2, S3:

$$z = -0.45685/0.41593 \approx -1.099 \Rightarrow |z| = 1.099 < 1.96 \rightarrow p < 0.05$$

No es siginicativo, la operacion 2 no difiere significativamente de la operacion 4 para la severidad 3.

Operacion 3, S3:

$$z = -0.37784/0.42655 \approx -0.885 \Rightarrow |z| = 0.885 < 1.96 \rightarrow p < 0.05$$

Tampoco es siginicativo, la operacion 3 no difiere significativamente de la operacion 4 para la severidad 3.

Resumen de significancia:

- Para S2 (leve), solo la diferencia entre la operacion 2 y la operacion 4 es significativa (p<0.05)
- Para S3 (moderada), la diferencia entre las operaciones 1 y4 es marginalmente significativa $(p \approx 0.048)$, mientras que las operaciones 2 y 3 no muestran diferencias significativas a la operacion 4.

- Entonces, para responder el ejercicio:

Para la categoría S2: Solo el coeficiente asociado a Operacion 2 es significativamente distinto de cero (p < 0.05). Esto indica que, para efectos colaterales leves (S2), la operación 2 difiere significativamente de la operación 4. En términos de razón de chances, OR = exp(–0.75131) ≈ 0.472, lo que sugiere que la operación 2 reduce los odds de presentar efectos leves (frente a ningún efecto) en comparación con la operación 4.

Para la categoría S3: El coeficiente para Operacion1 es marginalmente significativo (p ≈ 0.048), indicando que para efectos moderados (S3), la operación 1 tiene un efecto significativo en comparación con la operación 4. La razón de chances es OR = exp(–0.96725) ≈ 0.38, lo que indica que la operación 1 reduce los odds de efectos moderados (frente a ningún efecto) respecto a la operación 4.

Con una significancia del 5% - La operacion 2 tiene un efecto significativo para la severidad leve (S2) comparada con la operacion 4 - La operacion 1 presenta un efecto marginalmente sigificativo para la severidad moderada (S3) comparada con la operacion 4. - La operacion 3 no demuestra diferencias significativas.

Calcular las Razones de Chances (Odds ratios)

Podemos calcular las razones de chances (OR) al exponente de cada coeficiente y preentarlas en una tabla, para facilitar la interprestacion del modelo:

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
coef_mult_long <- melt(modelo_sum$coefficients)
names(coef_mult_long) <- c("Severidad", "Operacion", "Estimate")
coef_mult_long$OR <- exp(coef_mult_long$Estimate)
library(knitr)
kable(coef_mult_long, digits = 4,
      caption = "Coeficientes y Razones de Chances (OR) del Modelo Multinomial")
```

Coeficientes y Razones de Chances (OR) del Modelo Multinomial

Severidad	Operacion	Estimate	OR
S2	(Intercept)	-0.3327	0.7170
S3	(Intercept)	-1.1977	0.3019
S2	Operacion1	-0.4460	0.6402
S3	Operacion1	-0.9672	0.3801
S2	Operacion2	-0.7513	0.4717
S3	Operacion2	-0.4569	0.6333
S2	Operacion3	-0.0389	0.9619
S3	Operacion3	-0.3778	0.6853

Conclusiones finales

EL modelo multinomial relaciona la distribucion de severidad de los efectos colaterales con el tipo de operacion, usando $S = 1$, "*ninguno*" como la referencia en la respuesta y la operacion 4 o sea la mas invasiva como la referencia de los predictores.

Para el calculo de probabilidades para la operacion 1, Las expresiones obtenidas permiten calcular $P(S = 1|O = 1), P(S = 2|O = 1)$ y $P(S = 3|O = 1)$ en función de los coeficientes del modelo, en el caso de la operacion 1 tenemos que $\eta_2 = -0.4460$ y $\eta_3 = -0.9672$

$$P(S = 1|O = 1) = \frac{1}{1 + e^{-0.4460} + e^{-0.9672}} = \frac{1}{2.0203} = 0.495$$

$$P(S = 2|O = 1) = \frac{e^{-0.4460}}{1 + e^{-0.4460} + e^{-0.9672}} = \frac{e^{-0.4460}}{2.0203} = \frac{0.6402}{2.0203} = 0.317$$

$$P(S = 3|O = 1) = \frac{e^{-0.9672}}{1 + e^{-0.4460} + e^{-0.9672}} = \frac{e^{-0.9672}}{2.0203} = \frac{0.3801}{2.0203} = 0.188$$

donde $e = euler \approx 2.71828$

```
```

Notas:

AIC

El Criterio de Información de Akaike (AIC) es una medida estadística que se usa para evaluar la calidad relativa de los modelos estadisticos que se aplican a un conjunto de datos, el AIC ayuda a seleccionar entre un conjunto de modelos candidatos cual es el mas aporpiado considerando la bondad del ajuste con la complejidad el modelo.

El AIC se calcula con la fórmula

$$AIC = 2k - 2ln(L)$$

Dónde: - k es el número de parámetros estimados en el modelo. - L es el valor máximo de la función de verosimilitud para el modelo.

El AIC no nos proporciona una prueba absoluta de la calidad del modelo, compara la calidad relativa entre varios modelos, lo podemos interpretar asi:

- **Menor AIC**: indica que el modelo es mejor porque el modelo pierde menos información y tiene un mejor equilibrio entre el ajuste y la simplicidad.
- **Mayor AIC**: Indica que el modelo es menos adecuado porque puede estar sobreajustado o por el contrario no se ajusta a los datos suministrados.

Log-Odds

La escala de log-odds es una forma de expresar las probabilidades en términos logarítmicos, esta es utilizada generalmenteen regresión logística y otros modelos estadísticos, son útiles para predecir la probabilidad de un evento en función de una o más variables independientes.

Los log-odds se refieren al logaritmo natural de la razon de probabilidades (odds). las odds se calculan como la razón entre la probabilidad de que un evento ocurra o nó, ahora si asumimos que la probabilidad de éxito es p:

$$odds = \frac{p}{1-p} \;\Rightarrow\; log-odds = ln(\frac{p}{1-p})$$

* **Simetría**: Los log-odd se comportan de forma simétrica alrededor de cero y esto facilita el análisis y la interpretación. * **Linealidad**: En la regresión logística, estos permiten modelar la relacion que hay entre las variables independientes y la variable dependendiente de manera lineal. * **Rango**: Ya que las probabilidades están limitadas ente 0 y 1, los log-odds se mueven en cualquier valor real y esto puede ser util en algunos análisis.

Estimadores por mínimos cuadrados

Estos estimadores son una técnica estadística que es usada para estimar los paráameros de un modelo de regresión lineal. Estos nos sirven para encontrar los valores de los parámetros que minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por un modelo en estudio.

Estos estimadores se utilizan ampliamente en econometría, ingeniería, ciencias sociales y otras donde se construyan modelos predictivos y analicen relaciones entre variables.

En el modelo de regresión lineal simple la relación entre una variable dependiente y una independiente x se expresa de la siguiente manera:

$$y = \beta_0 + \beta_1 x + \epsilon \quad \therefore \quad \beta_0 = intercepto, \beta_1 = pendiente, \epsilon = error$$

* Como se calculan los estimadores? Pára un modelo de regresión lineal múltiple, los estimadores por mínimos cuadrados se calculan con la siguiente ecuación matricial:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y \quad \therefore$$

Símbolo	Descripción
X	Matriz de variables independientes
y	Vector de valores observados o variable dependiente
$\hat{\beta}$	Vector de estimadores de parámetros

Los estimadores de regresión lineal múltiple tienen las siguientes propiedades:

- Son Insesgados: significa que, en promedio los etimadores son iguales a los verdaderos valores de os parámetros.
- La varianza es mínima: entre todos los estiamdores este tipo de estimadores de regresión lineal múltiple tienen el valor de varianza mas bajo