

Metodología de la Investigación: Tarea 2

Badillo Hernandez Christian Francisco Lopez Hernandez Lizbeth
Nuñez Rodas Luis Antonio Pichardo Jiménez Sealtiel
Santana Jaimes Luz María Concepción

Tabla de contenidos

Ejercicio 4	2
<i>OR</i> para asma y alergia	2
Variable de interacción	3
Variable de confusión	5
Conclusión	7
Ejercicio 5	8
Ejercicio 7	11

Ejercicio 4

En la siguiente tabla se muestran los resultados de un estudio de casos y controles para estudiar el efecto del asma y el sexo sobre la alergia primaveral a las gramíneas. Se definen como casos a los individuos alérgicos y como controles a los no alérgicos.

Tabla 1: Alergias

	Hombre		Mujer	
	Asmáticos	No Asmáticos	Asmáticos	No Asmáticos
Casos	19	8	10	24
Controles	51	84	31	89

OR para asma y alergia

- ¿Existe relación entre el hecho de ser asmático y la definición de caso y control (sin considerar la variable sexo). Utilizar la medida de riesgo adecuada, e interpretar los resultados obtenidos (Usar $\alpha = 0.05$).

Como no se tomará en cuenta el sexo, se deben sumar hombres y mujeres pero haciendo la distinción entre casos y controles; asmáticos y no asmáticos:

Tabla 2: Tabla de Asma y Alergia

Asma	Alergia (Casos)	No Alergia (Controles)	Total
Sí	29	82	111
No	32	173	205
Total	61	255	316

Como medida de riesgo, se utilizará la razón de momios (*OR*), para ello se necesita primero el cálculo de los momios en cada caso, estos se definen como:

$$\Omega_1 = \frac{\mathbb{P}[Y = 1|X = 1]}{\mathbb{P}[Y = 2|X = 1]} = \frac{\pi_1}{1 - \pi_1} \quad \Omega_2 = \frac{\mathbb{P}[Y = 1|X = 2]}{\mathbb{P}[Y = 2|X = 2]} = \frac{\pi_2}{1 - \pi_2}$$

Donde Y es el desarrollo o no de la alergia; y X es el factor de exposición: tener o no asma. Entonces, el *OR* se calcula como el cociente del momio de tener alergia entre los que tienen asma y el momio de tener alergia entre los que no tienen asma:

$$\theta = \frac{\Omega_1}{\Omega_2}$$

Con intervalo de confianza igual a:

$$OR \in \left(\exp \left\{ \log(\hat{OR}) \pm Z_{(1-\alpha/2)} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right\} \right)$$

Se genera una función para calcular el cociente de momios y su intervalo de confianza:

```
IC_OR <- function(X, alpha) {  
  a <- X[1]  
  b <- X[3]  
  c <- X[2]  
  d <- X[4]  
  
  OR <- (a*d) / (b*c)  
  
  sd<-sqrt(1/a+1/b+1/c+1/d)  
  IC<-exp(log(OR)+c(-1,1)*qnorm(1-alpha/2)*sd)  
  
  return(list(OR = OR, IC = IC))  
}
```

Se calcula el cociente de momios y su intervalo a un 95% de confianza:

```
resultado <- IC_OR(matrix_asma_alergia, alpha = 0.05)  
OR_alergia <- resultado$OR  
IC_alergia <- resultado$IC
```

El *OR* de tener alergia en los que tienen asma es 1.912 veces el momio de tener alergia en los que no tienen asma. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.084, 3.371]. Dado que el intervalo no contiene al valor nulo, se puede decir que existe una relación positiva débil entre tener asma y desarrollar alergia primaveral. O bien, que las personas que tienen asma desarrollan más la alergia primaveral que las personas que no tienen asma.

Variable de interacción

- ¿Es la variable sexo modificadora del efecto? Justifique su respuesta.

Para poder detectar si la variable sexo es modificadora del efecto, también llamada una variable de interacción, se puede realizar análisis estratificado calculando el *OR* en los hombres y en las mujeres por separado y, al promediarlos, se debería tener un valor próximo al del *OR* sin estratificar.

Se genera la tabla de 2x2 para el estrato de los hombres:

Tabla 3: Tabla de Asma y Alergia en Hombres

Asma	Alergia (Casos)	No Alergia (Controles)	Total
Sí	19	51	70
No	8	84	92
Total	27	135	162

Se calcula el *OR* del estrato de los hombres:

```
resultado <- IC_OR(matrix_asma_alergia_h, alpha = 0.05)  
OR_alergia_h <- resultado$OR  
IC_alergia_h <- resultado$IC
```

El *OR* de tener alergia en los que tienen asma es 3.912 veces el momio de tener alergia en los que no tienen asma; en los hombres. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.596, 9.586]. Dado que el intervalo no contiene al valor nulo, se podría decir que existe una relación positiva moderada, en los hombres, entre tener asma y desarrollar alergia primaveral.

Ahora se calculará el *OR* para el estrato de las mujeres. Primero se genera la tabla de 2x2 para el estrato de las mujeres:

Tabla 4: Tabla de Asma y Alergia en Mujeres

Asma	Alergia (Casos)	No Alergia (Controles)	Total
Sí	10	31	41
No	24	89	113
Total	34	120	154

Se calcula el *OR* del estrato de las mujeres:

```
resultado <- IC_OR(matrix_asma_alergia_m, alpha = 0.05)
OR_alergia_m <- resultado$OR
IC_alergia_m <- resultado$IC
```

En las mujeres, el *OR* de tener alergia en los que tienen asma es 1.196 veces el momio de tener alergia en los que no tienen asma. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [0.515, 2.78]. Dado que el intervalo sí contiene al valor nulo, se podría decir que no existe una relación, en las mujeres, entre tener asma y desarrollar alergia primaveral.

Tomando en cuenta que en un caso sí hay relación y en otro no, ahora se realiza el promedio de los *OR* estratificados:

```
promedio_OR <- (OR_alergia_h + OR_alergia_m) / 2
```

El promedio de los *OR* es de 2.554 y el *OR* sin estratificar era de 1.912, hay una pequeña diferencia entre ambos valores. Aunque la diferencia entre el *OR* de los hombres es 3.27 veces el de las mujeres.

Para confirmar si esta diferencia es lo suficientemente significativa, se recurre a la prueba de homogeneidad de Woolf (1955), que permite comparar el *OR* de los dos estratos (hombres y mujeres). En el caso en que el sexo sea una variable de interacción, los *OR* deberían diferir de manera significativa entre sí. Se toma como hipótesis nula que los grupos son homogéneos y como altera que no son homogéneos, tomando de referencia $\alpha = 0.05$.

Se aplica el estadístico de prueba:

```
alergias <- xtabs(freq ~ .,
                 cbind(expand.grid(risk=c("Asma", "No Asma"),
                                   response=c("Alergia", "No Alergia"),
                                   gender=c("Hombre", "Mujer")),
                 freq=c(19, 8, 51, 84, 10, 24, 31, 89))
)

WoolfTest(alergias)
```

```
Woolf Test on Homogeneity of Odds Ratios (no 3-Way assoc.)
```

```
data: alergias
X-squared = 3.5601, df = 1, p-value = 0.05918
```

Dado que el *p-value* resultó mayor a 0.05, se rechaza la hipótesis alterna de que los grupos no son homogéneos. Por lo tanto, la variable sexo no es una de interacción o que modifique el efecto de la relación entre el asma y el desarrollo de alergia. Es decir, el efecto del asma sobre la alergia en realidad es consistente para hombres y mujeres.

Variable de confusión

Se consideran los siguientes requisitos para poder considerar a una variable como de confusión. Se define al factor de exposición como la presencia de asma, la presencia del factor de riesgo o enfermedad es poseer alergia de primavera y al factor de confusión como el sexo de la persona.

1. **Asociación entre el factor de confusión y la exposición en los no enfermos.** Se debe encontrar una asociación entre el sexo y el asma en el grupo de personas no alérgicas (controles). Es decir, un *OR* diferente a 1 en las personas que no tienen alergia.

Se genera la tabla con los datos:

Tabla 5: Tabla de Sexo y Asma para Controles (No Alergia)

Sexo	Asma (Sí)	Asma (No)	Total
Hombre	51	84	135
Mujer	31	89	120
Total	82	173	255

Se calcula el *OR* para los controles:

```
resultado <- IC_OR(matrix_sexo_asma_controles, 0.05)
OR_sex_asm_control <- resultado$OR
IC_sex_asm_control <- resultado$IC
```

En los controles, el *OR* de tener asma en los hombres es 1.743 veces el momio de tener asma en las mujeres. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.019, 2.982]. Como el intervalo no contiene al uno, se podría decir que existe una asociación positiva entre el tener asma y ser hombre. O bien, que dentro de las personas que no tienen alergias, los hombres desarrollan más el asma que las mujeres. Para comprobar la fuerza de esta asociación, se realizará una prueba χ^2 para independencia.

Se calcula el estadístico de prueba utilizando la función que toma en cuenta la corrección por continuidad de Yates:

```
chisq.test(matrix_sexo_asma_controles)

Pearson's Chi-squared test with Yates' continuity correction

data:  matrix_sexo_asma_controles
X-squared = 3.6251, df = 1, p-value = 0.05691
```

El *p-value* obtenido con la función devuelve un valor de 0.05691, el cual es mayor al 0.05. Entonces se rechaza la hipótesis nula de que las variables no son independientes, es decir que no hay suficiente evidencia como para rechazar la independencia de las variables. Al igual que el resultado obtenido en el análisis de variables de interacción, se puede decir que el asma influye de la misma manera a hombres y mujeres que no tienen alergias. Por lo que, al no haber relación, el primer requisito no se cumple.

2. **Asociación entre el factor de confusión y la enfermedad en los grupos no expuestos.** Se debe evaluar si existe una relación entre el sexo y la presencia de alergia en el grupo de personas que no tienen asma (no expuestos).

Se genera la tabla con los datos:

Tabla 6: Tabla de Sexo y Alergia para No Expuestos (No Asma)

Sexo	Asma (Sí)	Asma (No)	Total
Hombre	8	84	92
Mujer	24	89	113
Total	32	173	205

Se calcula el *OR* para los no expuestos:

```
resultado <- IC_OR(matrix_sexo_asma_no_exp, 0.05)
OR_no_exp <- resultado$OR
IC_no_exp <- resultado$IC
```

En las personas que no tienen asma, el *OR* de tener alergia en las mujeres es 2.831 veces el momio de tener alergia en los hombres. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.206, 6.65]. Como el intervalo no contiene al uno, se podría decir que existe una asociación positiva entre el tener alergia y ser mujer. O bien, que dentro de las personas que no tienen asma, las mujeres desarrollan más alergias que los hombres. Para comprobar la fuerza de esta asociación, se realizará una prueba χ^2 para independencia.

Pearson's Chi-squared test with Yates' continuity correction

```
data: matrix_sexo_asma_no_exp
X-squared = 5.1421, df = 1, p-value = 0.02335
```

El *p-value* obtenido con la función devuelve un valor de 0.02335, el cual es menor al 0.05. Por lo que se rechaza la hipótesis nula de que las variables son independientes. En este caso, se puede decir que existe una relación positiva, en las personas que no tienen asma, de desarrollar alergia y ser mujer. Por lo que, al haber asociación, el segundo requisito sí se cumple.

- 3. El factor de confusión no debe ser un paso intermedio en la secuencia causal entre el tipo de tratamiento y su éxito.** En este caso, como el sexo es una característica definida o intrínseca de los individuos, no se puede considerar como un paso intermedio entre el asma y la alergia. Es decir, no es posible que el asma provoque el sexo de las personas y con ello un mayor o menor aumento del desarrollo de alergias. Por lo que este criterio se cumple.

Tomando de referencia que el valor del *OR* bruto es 1.912 con un intervalo de confianza de [1.084 , 3.371] se puede apreciar que el intervalo está muy cerca del valor nulo, donde se demostraría que no hay asociación entre el desarrollo de alergia primaveral y tener o no asma, independientemente del sexo. Al estratificar por sexo, el *OR* de los hombres es 3.912 con un intervalo de confianza de [1.596 , 9.586], el cual revelaría una posible asociación positiva moderada entre tener alergia primaveral y asma para los hombres. Por último, el *OR* de las mujeres es 1.196 con un intervalo de confianza de [0.515 , 2.78], el cual demuestra que en el caso de las mujeres no se encuentra asociación entre tener alergia primaveral y asma. Dentro de los criterios, se encontró que dos de los tres se cumplen para ser una variable de confusión y uno no lo cumple por tener un *p-value* de 0.05691, es decir un valor muy cercano a la zona de rechazo. Tomando toda esta información, se concluye que la variable sexo sí es una variable de confusión, la cual genera confusión al enmascarar el efecto, donde, de tener más información, se podría notar que el *OR* global sea cercano a 1 (sin asociación) pero que el *OR* de un estrato muestra asociación (el de los hombres) y el otro no.

Conclusión

Para toda la información obtenida, se puede decir que, al no estratificar por sexo, no hay asociación entre el desarrollo de alergia primaveral y tener asma. Al estratificar por sexo, solo los hombres muestran una asociación moderada entre el desarrollo de alergia y tener asma; esta relación no se encuentra en las mujeres. Por ello, la variable sexo se considera como una de confusión, pero no una de interacción; esto último debido a no tener suficiente evidencia estadística para indicar que los *OR* de cada estrato difieran de manera significativa entre ellos.

Ejercicio 5

La siguiente tabla corresponde a la clasificación cruzada de las variables: Enfermedad coronaria (CHD, por sus siglas en inglés), Fumar (Smoke) y bebedor de café (Coffee).

Tabla 7: CHD, Smoke and Coffee

CHD	Smoke	Coffee	n
Sí	No	No	15
No	No	No	42
Sí	Sí	No	11
No	Sí	No	8
Sí	No	Sí	15
No	No	Sí	21
Sí	Sí	Sí	25
No	Sí	Sí	14

- Utilizando un modelo de regresión logística (*logit*) responda: ¿Están el tabaquismo y/o el café relacionados con el incremento del momio de CHD? (Use $\alpha = 0.05$).

Para aplicar el modelo de regresión logística, primero modificamos la tabla para que represente a cada observación en las filas y no como al conjunto de las frecuencias en cada grupo:

```
chd_expanded <- chd_smoke_coffee %>%  
  uncount(n)  
  
head(chd_expanded)
```

```
CHD Smoke Coffee  
1  Sí   No    No  
2  Sí   No    No  
3  Sí   No    No  
4  Sí   No    No  
5  Sí   No    No  
6  Sí   No    No
```

Se convierte la variable dependiente (CHD) a una variable *dummy* de 0 y 1 y se aplica el modelo logístico con *Smoke* y *Coffee* como variables independientes; este se definiría como:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot \text{Smoke} + \beta_2 \cdot \text{Coffee}$$

Donde:

π = probabilidad de que ocurra CHD

β_0 = Intercepto. El $\log(OR)$ de tener CHD en los no fumadores y no bebedores de café

β_1 = Coeficiente asociado a fumar (*Smoke*)

β_2 = Coeficiente asociado al consumo de café (*Coffee*)

```
chd_expanded$CHD_bin <- ifelse(chd_expanded$CHD == "Sí", 1, 0)  
  
modelo <- glm(CHD_bin ~ Smoke + Coffee, family = binomial, data = chd_expanded)
```



```
summary(modelo)
```

Call:

```
glm(formula = CHD_bin ~ Smoke + Coffee, family = binomial, data = chd_expanded)
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9572     0.2703   -3.541 0.000398 ***
SmokeSí       1.1020     0.3610    3.053 0.002269 **
CoffeeSí      0.5270     0.3542    1.488 0.136798
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 206.93  on 150  degrees of freedom
Residual deviance: 191.74  on 148  degrees of freedom
AIC: 197.74
```

Number of Fisher Scoring iterations: 4

Se calcula e^{β_i} con $i = 1, 2$ para obtener cada *OR*

```
exp(cbind(OR = coef(modelo), confint(modelo)))
```

```
              OR      2.5 %      97.5 %
(Intercept) 0.3839539 0.2215322 0.6424622
SmokeSí      3.0101153 1.4935502 6.1770819
CoffeeSí     1.6938032 0.8447583 3.4022183
```

El *OR* de tener CHD en los fumadores es 3.01 veces el *OR* de los no fumadores que tienen CHD; y con una confianza al 95% su intervalo no contiene al valor nulo, por lo que se dice que hay una asociación positiva entre tener CHD y ser fumador. O bien, al examinar el $\ln(OR)$ se encuentra que el *p-value* es menor a 0.05, confirmando la asociación entre ser fumador y tener CHD.

El momio de tener CHD en los que sí beben café es 1.69 veces el momio de los que tienen CHD y no beben café; y su intervalo contiene al valor nulo, por lo que se dice que no hay una asociación entre tener CHD y ser bebedor de café, con una confianza de 95%. O bien, al examinar el $\ln(OR)$ se encuentra que el *p-value* es mayor a $\alpha = 0.05$, confirmando que no hay asociación entre ser bebedor de café y tener CHD.

En conclusión, solo el ser fumador se encuentra asociado de manera positiva con tener enfermedad coronaria

- ¿Es la asociación de café y CHD mayor entre los fumadores? Es decir, ¿fumar es un efecto modificador de las asociaciones de café-CHD? (Use $\alpha = 0.05$)

Se genera el modelo que incluye a la interacción entre fumar y ser bebedor de café y se examina el coeficiente de la variable de interacción; este sería:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot \text{Smoke} + \beta_2 \cdot \text{Coffee} + \beta_3 \cdot (\text{Smoke} \cdot \text{Coffee})$$

Donde:

β_3 = Coeficiente asociado a la interacción (*Smoke* · *Coffee*)

```
# Modelo con interacción
modelo_con_interaccion <- glm(CHD_bin ~ Smoke * Coffee,
                             family = binomial,
                             data = chd_expanded)

summary(modelo_con_interaccion)
```

Call:

```
glm(formula = CHD_bin ~ Smoke * Coffee, family = binomial, data = chd_expanded)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.0296	0.3008	-3.423	0.000619	***
SmokeSí	1.3481	0.5535	2.435	0.014873	*
CoffeeSí	0.6931	0.4525	1.532	0.125573	
SmokeSí:CoffeeSí	-0.4318	0.7295	-0.592	0.553899	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 206.93 on 150 degrees of freedom
 Residual deviance: 191.39 on 147 degrees of freedom
 AIC: 199.39

Number of Fisher Scoring iterations: 4

Se calcula e^{β_i} con $i = 1, 2, 3$ para obtener cada OR

```
exp(cbind(OR = coef(modelo_con_interaccion), confint(modelo_con_interaccion)))
```

	OR	2.5 %	97.5 %
(Intercept)	0.3571429	0.1917723	0.6292934
SmokeSí	3.8500000	1.3186041	11.7863193
CoffeeSí	2.0000000	0.8242030	4.9042592
SmokeSí:CoffeeSí	0.6493506	0.1534047	2.7107380

Como se puede observar, el *p-value* del coeficiente asociado a la interacción entre las variables de ser bebedor de café y fumador es mayor al $\alpha = 0.05$, por lo que se puede concluir que no hay interacción significativa de los fumadores en la relación de ser bebedor de café y tener enfermedad coronaria. Si se aplica el exponencial y calculando el intervalo a un 95% de confianza, se observa que el intervalo $[0.1534, 2.7101]$ contiene al valor nulo, por lo que se confirma que, de manera significativa, no hay interacción de los fumadores en la relación de tomar café y tener enfermedad coronaria

Ejercicio 7

- ¿Existe diferencia significativa en las tasas de muerte al tiempo de inicio en el empleo (`start`)?. Responda esta pregunta construyendo las tasas por nivel de la covariable y después comparándolas como en el ejercicio previo.

Tabla 8: Tabla de tasas de incidencia

	pre1925	post1925	Total
Casos	161.0	115.0	276.0
Tiempo-Persona	96819.9	28170.8	124990.7

Suponiendo que nuestro nivel de exposición aumento con el tiempo, se tomará como exposición al periodo posterior a 1925 y como no expuesto al periodo anterior a 1925.

Definimos una función para estimar la razón de tasas de incidencia (IRR).

```
## Función para calcular el RR y su intervalo de confianza
calcular_IC_RR <- function(E1, T1, E2, T2, nivel_confianza = 0.95) {
  # Calcular el RR
  RR <- (E1 / T1) / (E2 / T2)

  # Logaritmo natural del RR
  ln_RR <- log(RR)

  # Error estándar del log(RR)
  SE_ln_RR <- sqrt(1 / E1 + 1 / E2)

  # Valor crítico de Z para el nivel de confianza especificado
  Z <- qnorm(1 - (1 - nivel_confianza) / 2)

  # Intervalo de confianza en escala log
  lower_ln <- ln_RR - Z * SE_ln_RR
  upper_ln <- ln_RR + Z * SE_ln_RR

  # Convertir a escala original
  IC_lower <- exp(lower_ln)
  IC_upper <- exp(upper_ln)

  # Resultado
  list(
    RR = RR,
    IC = c(IC_lower, IC_upper),
    SE = SE_ln_RR,
    nivel_sig = nivel_confianza)
}
```

Estimamos el IRR y hacemos la prueba de hipótesis.

```
rr.stats <- calcular_IC_RR(tasas.start.year$cases[2],
                           tasas.start.year$personyear[2],
                           tasas.start.year$cases[1],
                           tasas.start.year$personyear[1])

Z.test <- log(rr.stats$RR) / rr.stats$SE

p.value.rr.rate <- 2*(1-pnorm(Z.test))
```

Las tasas de incidencia 1000 persona-año es de 1.6629 para el periodo del inicio de trabajo antes de 1925 y 4.0822 para el periodo posterior a 1925. La razón de las tasas de incidencia es de 2.4549, con un intervalo de confianza al 95% de [1.9325, 3.1186].

Usando un nivel de significancia del 0.05, la prueba de hipótesis ($H_0 : \log RR = 0$) resulta ser significativa con un $p\text{-value} = 1.898481\text{e-}13$ y un estadístico Z de 7.3558. Por lo cual podemos rechazar la hipótesis nula de que ambas tasas de incidencia son iguales, al ser de dos colas solo se puede afirmar que son distintas sin indicar su dirección.

2. Utilizando el modelo de regresión Poisson, utilizando como *offset* la variable `personyrs`, responda la misma pregunta del inciso anterior. Este ejercicio es para que comprueben que sale lo mismo con los dos procedimientos y aprendan a usar este modelo.

El modelo de regresión Poisson es:

$$\begin{aligned}\log \text{rate} &= X_i^T \beta \\ &= \log \text{personyrs} + \beta_0 + \beta_1 \cdot \text{start}\end{aligned}$$

Donde:

$\text{rate} = \frac{\text{respdeath}}{\text{personyrs}}$ = es la tasa de incidencia de muerte respiratoria.

$\log \text{personyrs}$ = Variable *offset* que ajusta por el tiempo de exposición.

β_0 = Corresponde al log de la tasa esperada en el nivel de referencia ($\text{start} = 1$).

β_1 = Coeficiente asociado a la razón de la tasa de incidencia IRR para el tiempo de inicio en el empleo (start).

Se escribe el modelo en R.

```
poisson.rate <- glm(
  formula = respdeath ~ start,
  family = "poisson",
  offset = log(personyrs),
  data = df.resp.deaths)
```

Tabla 9: Estimadores de Regresión Poisson (start).

Variable	IRR ¹	95% CI ¹	p-value
(Intercept)	0.00	0.00, 0.00	<0.001

Tabla 9: Estimadores de Regresión Poisson (start).

Variable	IRR ¹	95% CI ¹	p-value
start	2.45	1.93, 3.11	<0.001

¹ IRR = Razón de tasas de incidencia, CI = Intervalo de confianza

Como se observa en la tabla 9, el coeficiente estimado por el modelo de regresión corresponde al riesgo relativo estimado anteriormente, además de que su intervalos de confianza son idénticos. Además su estimación es significativa estadísticamente a un nivel de $\alpha = 0.05$ ($Z = 7.356$, $p\text{-value} = 1.9e - 13$).

Se puede probar la suposición de que la varianza y la media de los datos son iguales, que subyace al modelo de regresión Poisson, con la siguiente instrucción de R.

```
epiDisplay::poisgof(poisson.rate)

$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 326.0066

$df
[1] 112

$p.value
[1] 8.963121e-23
```

Dado que nuestra hipótesis nula es que este supuesto se cumple, podemos ver que el modelo de Poisson no es el más indicado, por lo cual se realizará con el modelo quasipoisson.

```
qpoisson.rate <- glm(
  formula = respdeath ~ start,
  family = "quasipoisson",
  offset = log(personyrs),
  data = df.resp.deaths)
```

Tabla 10: Estimadores de Regresión QuasiPoisson (start).

Variable	IRR ¹	95% CI ¹	p-value
(Intercept)	0.00068	0.00033, 0.00134	<0.001
start	2.45492	1.52958, 3.89271	<0.001

¹ IRR = Incidence Rate Ratio, CI = Confidence Interval

Se observa en la tabla 10 que la IRR no cambia a diferencia de la amplitud del intervalo de confianza que se amplía tomando en consideración que existe sobre dispersión o subdispersión, para verificar este hecho

se utiliza el mismo modelo estimado.

```
summary(qpoisson.rate)$dispersion
```

```
[1] 3.77482
```

Lo que nos indica una sobre dispersión de 3.77, por ende es más adecuado el uso de un modelo de quasipoisson o uno binomial negativo.

- Utilizando regresión Poisson pruebe si hay diferencias significativas en las tasas muerte por nivel de arsénico.

El modelo de regresión Poisson es:

$$\begin{aligned}\log \mu_i &= X_i^T \beta \\ &= \log \text{personyrs} + \beta_0 + \beta_{1l} \cdot \text{arsenic}\end{aligned}$$

Donde:

$\text{rate} = \frac{\text{respdeath}}{\text{personyrs}}$ = es la tasa de incidencia de muerte respiratoria.

$\log \text{personyrs}$ = Variable *offset* que ajusta por el tiempo de exposición.

β_0 = Corresponde al log de la tasa esperada en el nivel de referencia ($\text{arsenic} = 1$)

β_{1l} = Coeficiente asociado a la razón de la tasa de incidencia IRR para los distintos niveles de exposición ($l = 2, 3, 4$) al arsénico (arsenic).

Se escribe el modelo en R.

```
poisson.rate.arsenic <- glm(
  formula = respdeath ~ arsenic,
  family = "poisson",
  offset = log(personyrs),
  data = df.resp.deaths)
```

Tabla 11: Estimadores de Regresión Poisson (arsenic).

Variable	IRR ¹	95% CI ¹	p-value
(Intercept)	0.00161	0.00137, 0.00188	<0.001
arsenic			
1	—	—	
2	2.08395	1.51816, 2.81843	<0.001
3	2.03434	1.33190, 2.99317	<0.001
4	4.08421	2.86695, 5.69117	<0.001

¹ IRR = Incidence Rate Ratio, CI = Confidence Interval

Los coeficientes estimados son significativos (véase tabla 11), indicando que el riesgo de muerte por enfermedad respiratoria es dos veces mayor para un nivel de exposición 2 (1-4 años) y 3 (5-14 años) con respecto al nivel de exposición 1 (<1 año) y es cuatro veces mayor cuando el tiempo de exposición al arsénico es mayor a 15 años (nivel 4). El nivel de riesgo esperado cuando el nivel de exposición es el más bajo es prácticamente nulo, por tanto el modelo indica que este nivel de exposición no afecta la tasa de muerte por enfermedad respiratoria.

Igualmente se comprueba la suposición del modelo de regresión Poisson.

```
epiDisplay::poisgof(poisson.rate.arsenic)

$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 311.1098

$df
[1] 110

$p.value
[1] 4.214491e-21
```

Se vuelve a obtener que el modelo no es el más apropiado. Al ajustar un modelo quasipoisson, se obtiene que el parámetro de dispersión es de 3.31293 (menor que la del modelo para `start`), indicando una sobre dispersión. Igualmente se obtienen los mismos estimadores de los coeficientes del modelo con un intervalo de confianza al 95% más amplio (véase tabla 12) como consecuencia, el coeficiente estimado para el nivel de exposición 3 al arsénico resulta no ser significativo.

Tabla 12: Estimadores de Regresión QuasiPoisson (arsenic).

Variable	IRR ¹	95% CI ¹	p-value
(Intercept)	0.00161	0.00119, 0.00213	<0.001
arsenic			
1	—	—	
2	2.08395	1.15455, 3.58003	0.012
3	2.03434	0.90845, 4.01967	0.061
4	4.08421	2.10103, 7.37576	<0.001

¹ IRR = Incidence Rate Ratio, CI = Confidence Interval

- Utilizando regresión Poisson verifique si existe interacción entre el tiempo de inicio en el empleo y los niveles de exposición de arsénico.

El modelo de regresión Poisson es:

$$\begin{aligned}\log \mu_i &= X_i^T \beta \\ &= \log \text{personyrs}_i + \beta_0 + \beta_1 \cdot \text{start} + \beta_{2l} \cdot \text{arsenic} + \beta_{3l} \cdot \text{start} * \text{arsenic}\end{aligned}$$

Donde:

$\text{rate} = \frac{\text{respdeath}}{\text{personyrs}}$ = es la tasa de incidencia de muerte respiratoria.

$\log \text{personyrs}_i$ = Variable *offset* que ajusta por el tiempo de exposición.

β_0 = Corresponde al log de la tasa esperada en el nivel de referencia ($\text{arsenic} = 1$).

β_1 = Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para el tiempo de inicio en el empleo (start).

β_{2l} = Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para los distintos niveles de exposición ($l = 2, 3, 4$) al arsénico (arsenic).

β_{3l} = Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para los distintos niveles de interacción entre el tiempo de inicio en el empleo (start) y el nivel de exposición al arsénico ($l = 2, 3, 4$).

Estimamos el modelo de interacción con R.

```
poisson.rate.inter <- glm(
  formula = respdeath ~ start*arsenic,
  family = "poisson",
  offset = log(personyrs),
  data = df.resp.deaths)
```

Tabla 13: Estimadores de Regresión Poisson (Interacción: $\text{arsenic} * \text{start}$).

Variable	IRR ¹	95% CI ¹	p-value
(Intercept)	0.00067	0.00041, 0.00108	<0.001
start	2.00270	1.41941, 2.79100	<0.001
arsenic			
1	—	—	
2	1.81760	0.71373, 4.58525	0.2
3	1.90768	0.51954, 6.37911	0.3
4	1.72964	0.32878, 6.92450	0.5
start * arsenic			
start * 2	1.14016	0.57694, 2.19058	0.7
start * 3	0.98853	0.43239, 2.23949	>0.9
start * 4	1.37609	0.62033, 3.39683	0.5

¹ IRR = Incidence Rate Ratio, CI = Confidence Interval

Al ajustar el modelo se obtiene que los estimadores de la interacción no son significativos a un nivel de $\alpha = 0.05$, por lo cual no hay evidencia estadística para hablar de una posible interacción entre el tiempo de inicio del trabajo y el nivel de exposición al arsénico. Solo un efecto principal es significativo (*start*) indicando que la razón de la tasa de incidencia es dos veces mayor cuando se empieza el empleo en años posteriores a 1925 controlando por el nivel de exposición al arsénico. Esta exposición resulta no ser significativa cuando se controla por el periodo de inicio del trabajo.

Se procede a corroborar la aplicabilidad del modelo.

```
epiDisplay::poisgof(poisson.rate.inter)

$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 277.981

$df
[1] 106

$p.value
[1] 2.32494e-17
```

Tabla 14: Estimadores de Regresión QuasiPoisson (Interacción: arsenic * start).

Variable	IRR ¹	95% CI ¹	p-value
(Intercept)	0.00067	0.00029, 0.00151	<0.001
start	2.00270	1.09554, 3.52872	0.021
arsenic			
1	—	—	
2	1.81760	0.35778, 8.98415	0.5
3	1.90768	0.18526, 14.8249	0.6
4	1.72964	0.07751, 16.8197	0.7
start * arsenic			
start * 2	1.14016	0.34261, 3.48143	0.8
start * 3	0.98853	0.23114, 4.10759	>0.9
start * 4	1.37609	0.36327, 7.23870	0.7

¹ IRR = Incidence Rate Ratio, CI = Confidence Interval

El modelo no es el más adecuado. La dispersión estimada es de 2.9637428 indicando la presencia de sobre dispersión (se estimo usando un modelo quasipoisson con la interacción incluida). Los parámetros estimados se muestran en la tabla 14. A excepción de la amplitud de los intervalos de confianza, las estimaciones e interpretaciones de este modelo no son distintas al anterior.