Lista de Ejercicios 1

# 1.

Supón que X e Y tienen la siguiente función de masa de probabilidad conjunta:

| X\Y | 0 | 1 | 2 |
| --- | --- | --- | --- |
| 1 | 10% | 10% | 20% |
| 2 | 20% | 10% | 10% |
| 3 | 10% | 10% | 0% |

**a) Encuentra las probabilidades marginales de X e Y.**

- P(X=1) = 0.10 + 0.10 + 0.20 = 0.40

- P(X=2) = 0.20 + 0.10 + 0.10 = 0.40

- P(X=3) = 0.10 + 0.10 + 0.00 = 0.20

- P(Y=0) = 0.10 + 0.20 + 0.10 = 0.40

- P(Y=1) = 0.10 + 0.10 + 0.10 = 0.30

- P(Y=2) = 0.20 + 0.10 + 0.00 = 0.30

**b) Calcula E[X|Y= y], para y= 0,1 y 2.**

- Para Y=0:

- Para Y=1:

- Para Y=2:

**c) Calcula E[E[X|Y]].**

**d) Calcula E[X] directamente y verifica la Ley de la Esperanza Iterada.**

**e) Calcula E[Y] (cómo prefieres).**

**f) Calcula la Covarianza entre X y Y.**

Solo los términos donde y ≠ 0 aportan:

- (1,1): 1×1×0.1 = 0.1

- (1,2): 1×2×0.2 = 0.4

- (2,1): 2×1×0.1 = 0.2

- (2,2): 2×2×0.1 = 0.4

- (3,1): 3×1×0.1 = 0.3

Total:

Entonces:

# 2.

Sean y .

1. **Calcula E[X|Y = 0.25]**

Si Y = 0.25, entonces , y ambos valores son igualmente probables bajo U[-1,1].

1. **Describa la función E[X|Y=y]**

Dado que , entonces X dado Y=y es simétrico en , por lo tanto:

1. **¿Son X y Y independientes?**

No, porque Y está completamente determinado por X. Hay dependencia funcional.

1. **¿Es X independiente en media de Y?**

Sí, porque

1. **¿Es Y independiente en media de X?**

No, porque , y

Y depende funcionalmente de X.

1. **Calcule la covarianza entre X y Y**

Sabemos , y

(por simetría de la función impar sobre U[-1,1])

Alternativamente, sabemos que independencia en média implica cero covarianza.

# 

# 3.

En R, use los seguintes comandos para carregar una base de datos sobre atletas olímpicos:

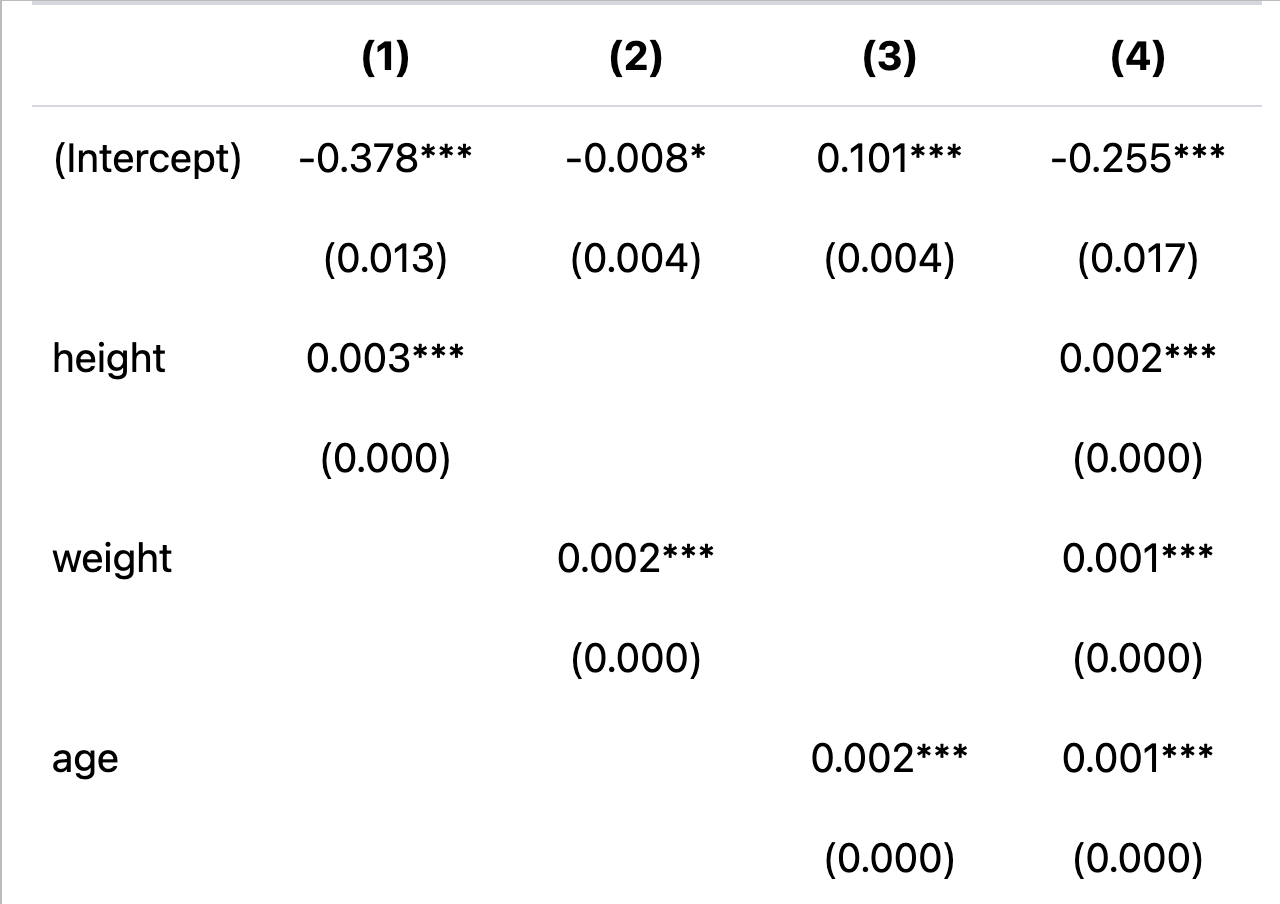
install.package(“tidytuesdayR”) # Solo necesario una vez

tuesdata <- tidytuesdayR::tt\_load(2024, week = 32)

olympics <- tuesdata$olympics

Se puede encontrar más información sobre eses datos en este enlace: <https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-08-06/readme.md>

1. Nos interesa conocer cómo la altura, el peso, y la edad de los atletas afecta su desempeño deportivo. La medida de desempeño va a ser si el atleta ganó una medalla (cualquier tipo). Estima tres regresiones con cada una de esas características. Después estima una regresión con las tres al mismo tiempo. Presenta una tabla de resultados e interprétalos. (Antes de empezar, te recomiendo que mantenga apenas las observaciones sin valores faltantes en altura, peso y edad. Y tenga cuidado con valores NA cuando crias la variable de desempeño.)



El modelo 1 nos dice que 1 cm a más de altura se asocia con 0.3 puntos porcentuales a más en la probabilidad de ganar una medalla. En el modelo 2, un kilograma a más de peso significa 0.2 puntos porcentuales más probabilidad de medalla. En el modelo 3, cada año de edad aumenta la probabilidad de suceso en 0.2 puntos porcentuales. Cuándo utilizamos el modelo conjunto, esos valores son 0.2 p.p., 0.1 p.p. y 0.1 p.p. Todos los resultados son estadísticamente significativos.

1. Estima una regresión de desempeño en el log de la altura, el log del peso, edad, sexo, y temporada. Interpreta cada coeficiente. ¿Cómo se explica el signo del coeficiente en temporada?

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.3239817 0.0961185 -24.178 < 2e-16 \*\*\*

log(height) 0.3569178 0.0226868 15.732 < 2e-16 \*\*\*

log(weight) 0.1552565 0.0071814 21.619 < 2e-16 \*\*\*

age 0.0009912 0.0001452 6.828 8.61e-12 \*\*\*

seasonWinter -0.0302735 0.0019688 -15.376 < 2e-16 \*\*\*

sexM -0.0742183 0.0019824 -37.439 < 2e-16 \*\*\*

* Un aumento de 1% en altura aumenta la chance de medalla en 0.35 puntos porcentuales, manteniendo fijas las otras características.
* Un aumento de 1% en peso aumenta la chance de medalla en 0.15 puntos porcentuales, manteniendo fijas las otras características.
* Un año de edad a más aumenta la chance de medalla en 0.1 punto porcentual, manteniendo fijas las otras características.
* Los atletas en las olimpíadas de invierno tienen 3 puntos porcentuales a menos de chance de ganar medallas, en comparación con atletas con las mismas características en las olimpiadas de verano.
* Hombres tienen 7 puntos porcentuales a menos de chance de ganar medallas en comparación con mujeres con características similares.

El coeficiente negativo para las olimpiadas de invierno significa que las pruebas son más competitivas, es decir, existen más competidores para cada medalla, y por tanto la chance de cada uno es menor.

1. A partir del modelo de la parte b, nos interesa saber si el impacto de edad es mayor o menor para mujeres. Testa la hipótesis y explica su conclusión.

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.3144932 0.0961146 -24.081 < 2e-16 \*\*\*

log(height) 0.3479199 0.0227166 15.316 < 2e-16 \*\*\*

log(weight) 0.1556281 0.0071806 21.673 < 2e-16 \*\*\*

seasonWinter -0.0307126 0.0019695 -15.594 < 2e-16 \*\*\*

age 0.0024701 0.0002473 9.990 < 2e-16 \*\*\*

sexM -0.0192239 0.0077031 -2.496 0.0126 \*

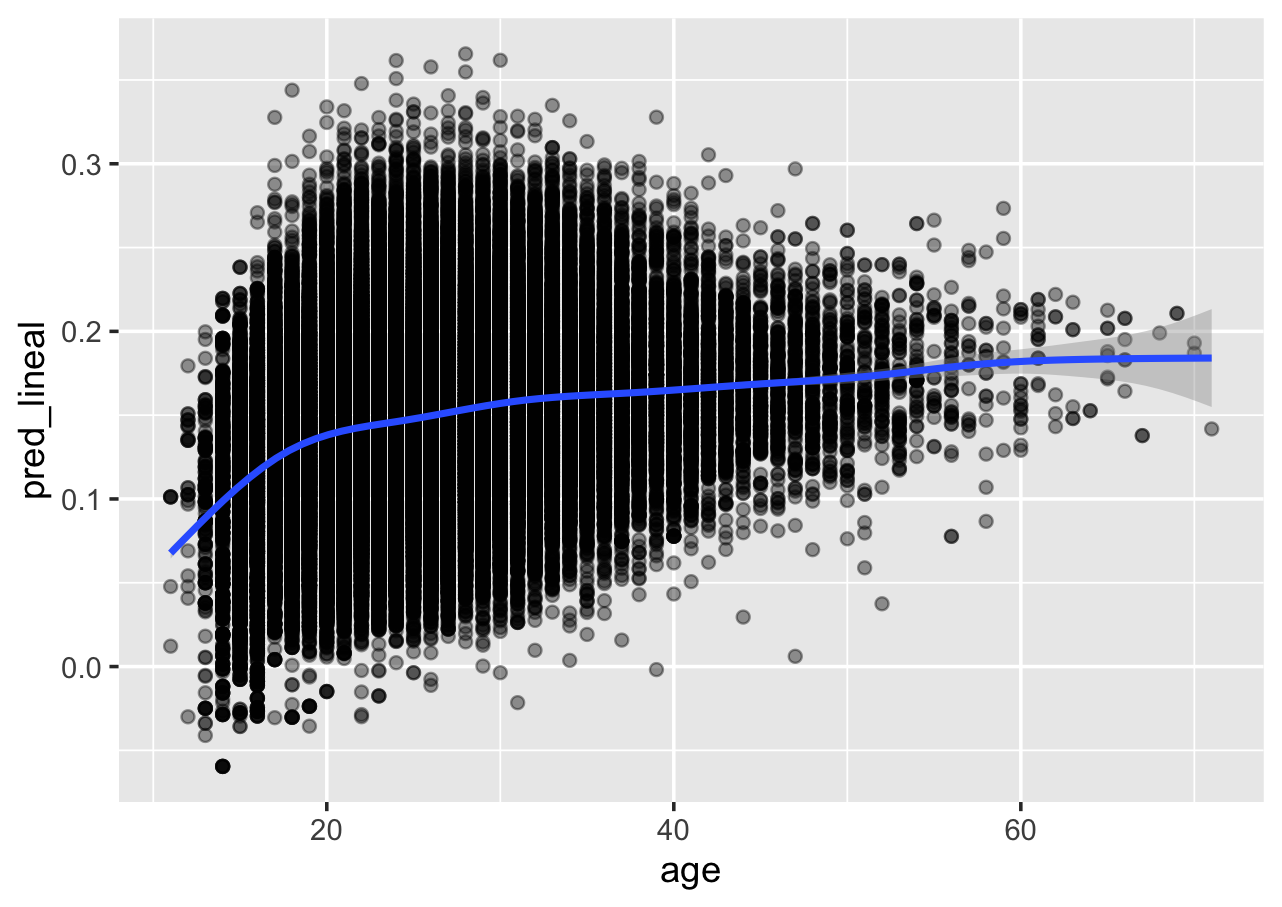
age:sexM -0.0022325 0.0003022 -7.388 1.5e-13 \*\*\*

El coeficiente de age:sexM es negativo y significativo. Eso quiere decir que el efecto de edad para mujeres es mayor que el efecto para hombres. Si probamos el efecto de edad para hombres, la estimación es 0.00024, no estadísticamente diferente de cero.

1. Con el modelo en b, haga predicciones sobre la probabilidad de ganar una medalla para cada atleta. Compara la probabilidad estimada para los atletas que ganaron contra los que no ganaron. Presenta un gráfico de la relación entre edad y la predicción.

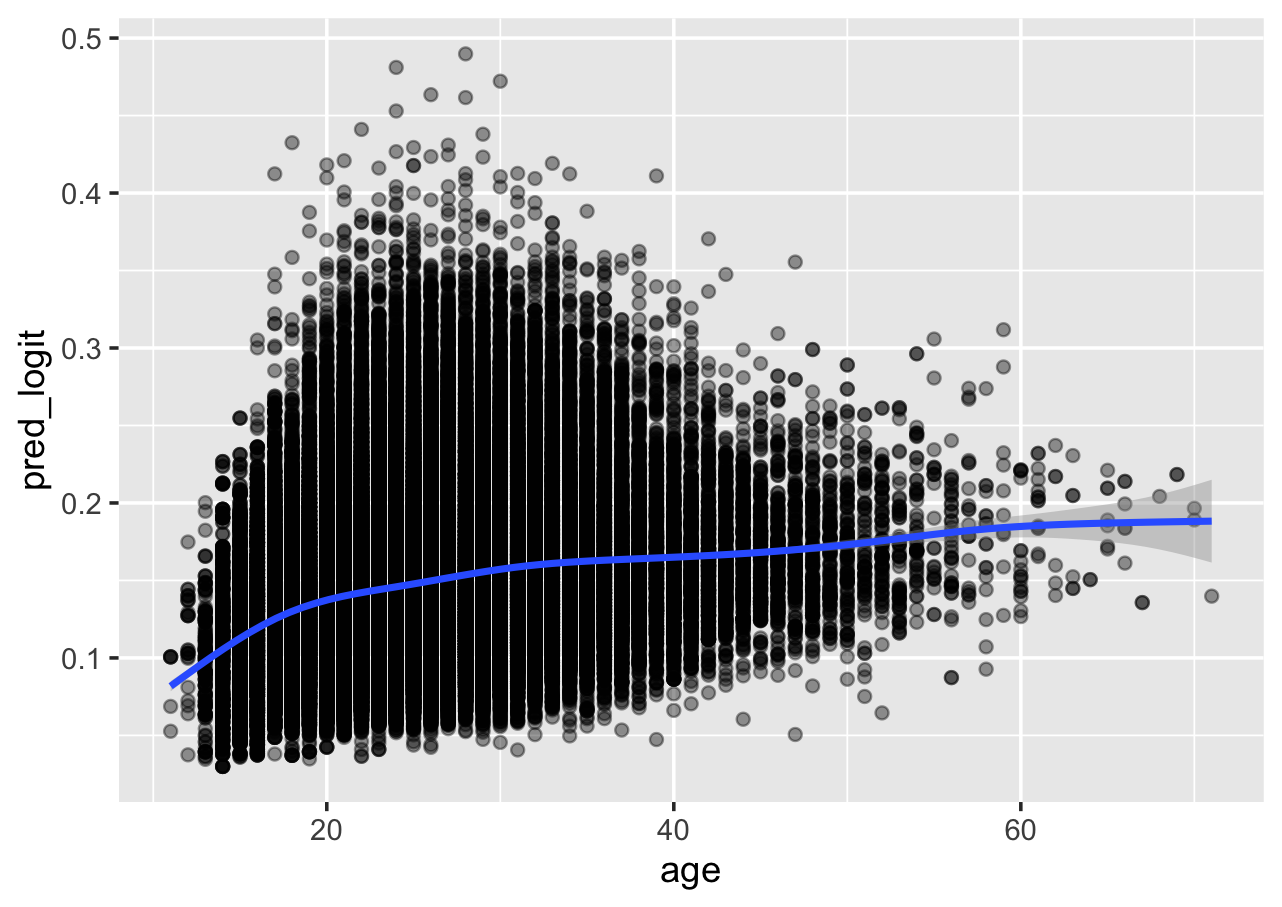
| Ganó Medalla | Prob(medalla) |
| --- | --- |
| 0 | 0.144 |
| 1 | 0.160 |

El modelo parece dar probabilidades muy similares sobre la probabilidad de medalla. Esas informaciones tienen un poco de información sobre el desempeño, pero no mucho.



1. Repita la parte d, utilizando un modelo logit. ¿Te parece que el modelo es mejor que el lineal?

| Ganó Medalla | Prob(medalla) |
| --- | --- |
| 0 | 0.144 |
| 1 | 0.160 |



El modelo logit hace predicciones casi iguales. La única diferencia importante es que las probabilidades son siempre mayores que cero, lo que no es verdad para el modelo lineal.

# 

# 

# library(tidyverse)

# library(modelsummary)

# 

# # Cargando los datos

# tuesdata <- tidytuesdayR::tt\_load(2024, week = 32)

# olympics <- tuesdata$olympics

# 

# #crear la variable de desempeño

# olympics <- olympics %>% mutate(wonmedal = 1-is.na(medal)) %>%

# filter(!is.na(height),!is.na(weight),!is.na(age)) # filtrar datos sin NA

# 

# 

# # Parte A

# model1 <- lm(wonmedal ~ height, data=olympics)

# model2 <- lm(wonmedal ~ weight, data=olympics)

# model3 <- lm(wonmedal ~ age, data=olympics)

# model4 <- lm(wonmedal ~ age + weight + height, data=olympics)

# 

# msummary(list(model1,model2,model3,model4),stars=TRUE)

# 

# # Parte B

# model\_b <- lm(wonmedal ~ log(height) +

# log(weight) +

# age +

# season +

# sex,

# data= olympics )

# 

# summary(model\_b)

# 

# # Parte C

# model\_c <- lm(wonmedal ~ log(height) +

# log(weight) +

# season +

# age\*sex,

# data= olympics )

# 

# summary(model\_c)

# 

# 

# library(car) # para linearHypothesis

# linearHypothesis(model\_c, "age + age:sexM =0")

# 

# library(gmodels) # para estimable

# estimable(model\_c,c("age"=1,

# "age:sexM"=1))

# 

# 

# # Parte D

# olympics$pred\_lineal <- predict(model\_b, newdata = olympics)

# 

# olympics %>% group\_by(wonmedal) %>% summarise(mean(pred\_lineal))

# 

# ggplot(olympics, aes(x = age, y = pred\_lineal)) +

# geom\_point(alpha = 0.4) + geom\_smooth()

# 

# # Parte E

# model\_logit <- glm(wonmedal ~

# log(height) +

# log(weight) +

# age +

# sex +

# season ,

# data= olympics ,

# family = binomial(link = "logit"))

# summary(model\_logit)

# 

# olympics$pred\_logit <- predict(model\_logit,type="response")

# 

# olympics %>% group\_by(wonmedal) %>% summarise(mean(pred\_logit))

# 

# ggplot(olympics, aes(x = age, y = pred\_logit)) +

# geom\_point(alpha = 0.4) + geom\_smooth()

# 

# 4.

### **Pregunta: Exposición a Plaguicidas y Salud Mental**

Estás colaborando con un equipo de investigación que estudia el impacto potencial de la exposición a plaguicidas sobre la salud mental en poblaciones rurales. El equipo ha recolectado datos observacionales y está considerando cómo modelar las relaciones causales entre las variables mediante el uso de diagramas causales.

La pregunta de investigación principal es: **¿Cuál es el efecto de la exposición a plaguicidas sobre la salud mental?**

**Descripción de variables:**

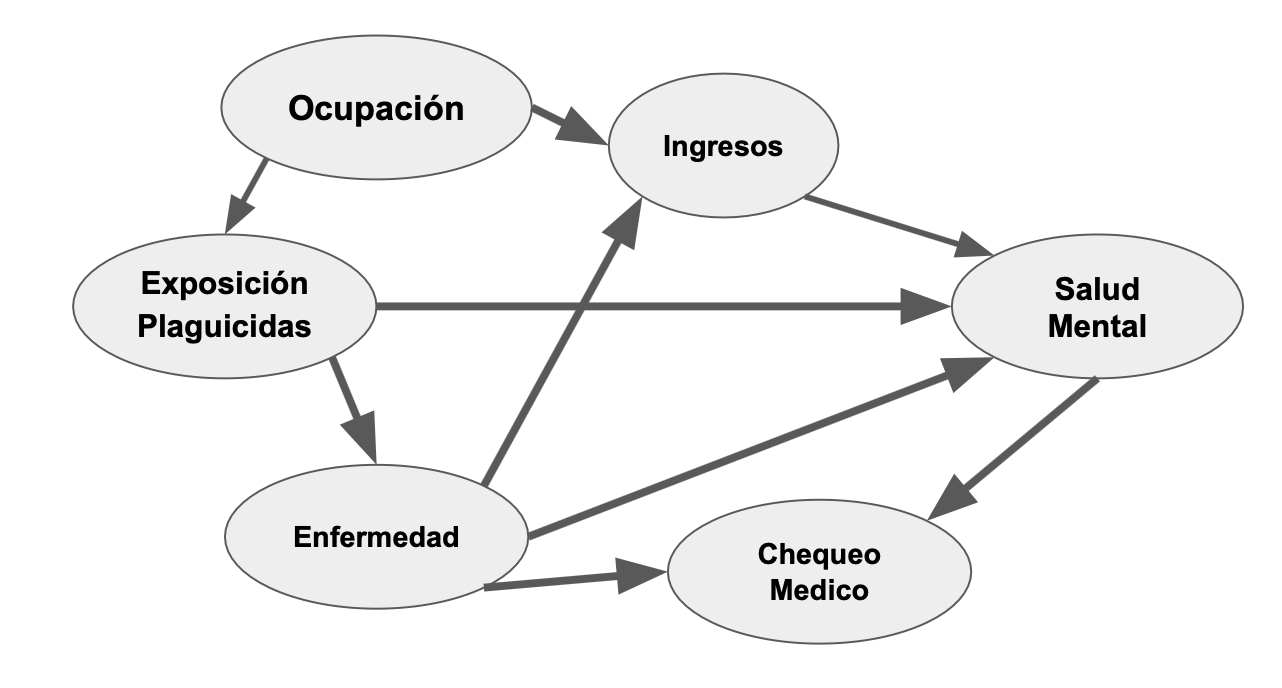
* **ExposiciónPlaguicidas**: Si la persona trabaja en un entorno con alta exposición a plaguicidas.
* **SaludMental**: Diagnóstico de depresión, ansiedad u otros trastornos del estado de ánimo.
* **Enfermedad**: Síntomas físicos o episodios de enfermedad posiblemente relacionados con la exposición.
* **Ocupación**: Tipo de empleo, especialmente si involucra trabajo agrícola.
* **Ingreso**: Nivel de ingresos mensuales.
* **ChequeoMédico**: Si la persona ha visitado a un médico recientemente para una evaluación general.

El equipo propone las siguientes relaciones entre variables:

* **SaludMental** es influida por **ExposiciónPlaguicidas**, **Enfermedad** e **Ingreso**.
* **ExposiciónPlaguicidas** depende de la **Ocupación** (por ejemplo, trabajo agrícola vs. no agrícola).
* **Enfermedad** es causada por la **ExposiciónPlaguicidas**.
* **Ingreso** depende tanto de la **Ocupación** como de la **Enfermedad**.
* **ChequeoMédico** representa si una persona ha tenido un chequeo médico reciente. Depende de la **SaludMental** y de la **Enfermedad**.

### **Preguntas**

**a)** Dibuje el diagrama causal que representa las relaciones causales propuestas.



**b)** Enumere todos los caminos distintos entre **ExposiciónPlaguicidas** y **SaludMental**.  
 (i) Clasifique cada uno como un camino de **front door** o **backdoor**.  
 (ii) Indique si el camino está **abierto** o **cerrado**, asumiendo que no se condiciona en ninguna variable.

1. Exposición -> Salud Mental. Front door, abierto
2. Exposición -> Enfermedad -> Salud Mental. Front door, abierto
3. Exposición -> Enfermedad -> Chequeo Médico <- Salud Mental. Backdoor, cerrado ( Chequeo médico es un collider)
4. Exposición -> Enfermedad -> Ingresos -> Salud Mental. Frontdoor, abierto
5. Exposición <- Ocupación -> Ingresos -> Salud Mental. Backdoor, abierto
6. Exposición <- Ocupación -> Ingresos <- Enfermedad -> Salud Mental. Backdoor, cerrado(Ingresos es un collider)
7. Exposición <- Ocupación -> Ingresos <- Enfermedad -> Chequeo Médico <- Salud Mental. Backdoor, cerrado(Ingresos y Chequeo Médico son colliders)

**c)** ¿Qué variables se deben controlar para estimar el **efecto causal total** de la exposición a plaguicidas sobre la salud mental? Justifique su respuesta.

Es necesario controlar solamente por Ocupación. Eso cierra el camino 5, que es el único backdoor abierto, mantiene cerrados los otros caminos de backdoor, y abiertos los caminos front door.

**d)** ¿Qué variables se deben controlar para estimar el **efecto directo** de la exposición a plaguicidas sobre la salud mental?

En este caso, es suficiente que controlamos Ocupación y Enfermedad.

**e)** Suponga que el conjunto de datos solo incluye personas que se han hecho un **chequeo médico reciente**.  
 (a) ¿Qué tipo de sesgo podría introducirse?  
 (b) ¿Qué caminos se ven afectados por este condicionamiento?

Eso introduce un sesgo de selección. Estamos condicionando en chequeo médico, lo que va abrir el camino 3.

**f)** Una colega sugiere que la **Ocupación** podría tener un efecto **directo** sobre la salud mental (por ejemplo, debido al estrés laboral o las condiciones del trabajo).  
 ¿Cómo podrías evaluar esta hipótesis utilizando las asociaciones observadas en los datos, suponiendo que el resto del diagrama es correcto?

Podemos examinar la relación entre Ocupación y Salud Mental condicionando en Ingresos y Exposición a Plaguicidas. Si existe una relación directa, encontraremos una correlación mismo manteniendo fijas estas variables.