Metodología. Tarea II

1. La base de datos *sexbias.csv* contiene la información coleccionada de un estudio realizado por la Universidad de California en Berkeley, para evaluar si los hombres estaban recibiendo un trato preferencial sobre las mujeres en la admisión a ciertos programas de posgrado.

   Suponiendo que los hombres y las mujeres que solicitaron la admisión a los programas de posgrado fueron igualmente bien calificados, se esperaría tasas de aceptación iguales o semejantes por género. Esta base contiene la experiencia de 4526 sujetos que deseaban ingresar a los *seis* programas de posgrado más prestigiosos de esta universidad. Las variables coleccionadas son:

   Base: Sexbias

Variable	Descripción
SEX	1: MALE 2: FEMALE (Factor de "exposición")
ACCEPT	Aceptado en el programa: +/− (La respuesta)
MAJOR	Departamentos (A, B, C, D, E, F) (UC Berkeley policy does not allow majors to be identified by name)

   Con esta información se puede afirmar a un nivel de significancia, $\alpha = 0.05$, que existe evidencia que los hombres estaban recibiendo un trato preferencial en detrimento de las mujeres. Argumente cuidadosamente su respuesta.

2. La base de datos *senic.csv* corresponde a *Study on the Efficacy of Nosocomial Infection Control (SENIC Project)* (La descripción de los objetivos y variables de este estudio, se encuentran en el documento: SENIC data set (description)). Con esta base realice lo siguiente

   - Ajuste un modelo de regresión lineal simple entre las variable *INFRISK* (Riesgo de infección. Como variable de respuesta) y *LOS* (Duración del periodo de estancia o tiempo de duración de la estancia. Como variable explicativa).

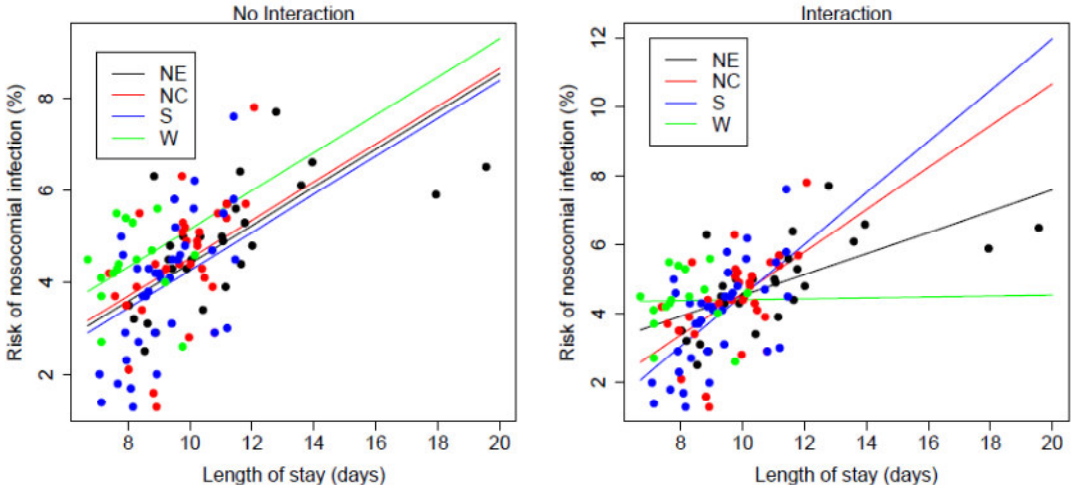
- Interprete el parámetro estimado asociado a su única variable predictora.

- ¿Tiene una interpretación lógica el intercepto del modelo ($\hat{\beta}_0$) en el contexto de los datos que lo generan? Si es así, interprete este parámetro estimado.

- Ajuste el mismo modelo controlando por región (REGION) (recuerde que esta variables es nominal con cuatro categorías). Interprete los parámetros estimados para las categorías de región. ¿Se puede considerar que esta variable es de confusión? Argumente su respuesta.

- Adicione al modelo inicial la interacción (multiplicativa) entre las variable LOS y REGION. ¿Se puede afirmar a un nivel de significancia, $\alpha = 0.05$, que existe interacción entre estas variables? Y si así fuera, ¿qué significa en el contexto de esta investigación?.

- Genere y muestre las gráficas de regresión para el caso del modelo que incluye a la región y el que incluye a la interacción. Las gráficas deben ser similares a las siguientes.


No Interaction
Risk of nosocomial infection (%)
Length of stay (days)
NE
NC
S
W
Interaction
Risk of nosocomial infection (%)
Length of stay (days)
NE
NC
S
W

3. El riesgo de *cáncer de mama* está directamente relacionado con la edad de las mujeres a las que tuvieron su primer hijo. Cuando se evalúa la asociación entre el número total de hijos y el riesgo de cáncer de mama, ¿debe controlarse por edad de la madre a la que tuvieron el primer hijo? Argumente su respuesta.
   En esta cadena causal, ¿qué roles tienen las tres variables involucradas?.

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.

Alergias

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

   - 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 $\alpha = 0.05$).

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

   - Cuál sería la conclusión con base a los resultados del estudio.

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* (Coffe).

CHD, Smoke and Coffe

CHD	Smoke	Coffe	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

Información sobre el estudio.

- Estudio de casos y controles (Enfermedad=CHD)
- Hombres entre 40-50 años de edad, previamente en buen estado de salud
- **Preguntas sobre el estudio**: Utilizando un modelo de regresión logística (logit), responda las siguientes preguntas.
  - ¿Están el tabaquismo y/o el café relacionadas (aquí lo que se pregunta es si cualquiera de las dos o ambas están relacionadas) con el incremento del momio de CHD? (Use $\alpha = 0.05$).
  - ¿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 $\alpha = 0.05$).

6. La base de datos *Melanoma.csv*, contiene información sobre un estudio de cohorte para estudiar muerte por *melanoma* realizado en Dinamarca. La descripción de esta base es la siguiente:

   Description: The Melanoma data frame has data on 205 patients in Denmark with malignant melanoma. This data frame contains the following columns: *time*: survival time in days, possibly censored. *status*: 1 died from melanoma, 2 alive, 3 dead from other causes. *sex*: 1 = male, 0 = female. *age*: age in years. *year*: of operation. *thickness*: tumour thickness in mm. *ulcer*: 1 = presence, 0 = absence. Con estos datos realice

   - Calcule la tasa de incidencia anual por 1000 de muerte por *melanoma*
   - Realice los 4 tipos intervalos de confianza vistos en clase (la programación debe ser de ustedes, no la del script) ($\alpha = 0.05$)
   - ¿Es estadísticamente distinta esta tasa de incidencia entre hombres y mujeres?. Argumente su respuesta. ($\alpha = 0.05$)

- Calcule la tasa de incidencia anual por 1000 de muerte dentro de este estudio.

- ¿Es estadísticamente distinta esta tasa de incidencia de la de muerte sólo por melanoma?. Argumente su respuesta. ($\alpha = 0.05$)

7. La descripción de la base de datos *Montana.csv* es:

   The dataset Montana was extracted from an occupational cohort study conducted to test the association between respiratory deaths and exposure to arsenic in the industry, after adjusting for various other risk factors. The main outcome variable is *respdeath*. This is the count of the number of deaths among *personyrs* or *personyears* of subjects in each category. The other variables are independent covariates including age group *agegr*, period of employment *period*, starting time of employment *start* and the level of exposure to arsenic during the study period *arsenic*. Los niveles de factores de las variables categóricas son:

   $agegr : \mathbf{1} : 40 - 49, \mathbf{2} : 50 - 59, \mathbf{3} : 60 - 69, \mathbf{4} : 70 - 79$
   $period : \mathbf{1} : 1938 - 1949, \mathbf{2} : 1950 - 1959, \mathbf{3} : 1960 - 1969, \mathbf{4} : 1970 - 1977$
   $start : \mathbf{1} : pre - 1925, \mathbf{2} : 1925 \ \& \ after$
   $arsenic : \mathbf{1} :< 1 \ year, \mathbf{2} : 1 - 4 \ years, \mathbf{3} : 5 - 14 \ years, \mathbf{4} : 15 \ + \ years$

   Con esta información, realice lo siguiente (El nivel de significancia siempre será ($\alpha = 0.05$))

   - ¿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 compándolas como en el ejercicio previo.

   - 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.

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

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

Felíz 4 de Marzo