16.43 Bioestadística
Instituto Tecnológico de Buenos Aires

Examen Parcial

Profesores:

- Santa María, Victoria
- Berrino, Eugenia
- Cirignoli, Camila

Alumno:

- Romano, Alejo Nicolás

Fecha de entrega: 1/6/2022

Índice

1. Ejercicio 1

El trabajo planteado corresponde al diseño de un estudio de cohortes, dado que el estudio es de tipo observacional y consiste en seguir un grupo de muestras en un intervalo de tiempo. Esto tiene sentido dado que este diseño de estudio se utiliza para evaluar la relación entre una exposición y la ocurrencia de un evento de interés, lo que permite identificar factores de riesgo.

La pregunta PICO para esta investigación es la siguiente:

- **Población (P).** Pacientes que consultaron por primera vez en el año 2010 y 2011 al servicio de Cardiología del Hospital de Clínicas.
- **Intervención (I).** Las covariables del estudio, es decir, la exposición que tienen los sujetos (sexo, colesterol, cigarrillos, presiones, etc.).
- **Comparación (C).** No hay comparación, dado que no se evalúan poblaciones diferentes entre sí.
- **Resultado (O).** Enfermedad coronaria (EC).

2. Ejercicio 2

El diseño experimental puede presentar los siguientes sesgos:

- **De selección.** Dado que los pacientes asisten al servicio de cardiología de un hospital, estos presentan mayor prevalencia del evento de interés. Además, puede ser que los sujetos presenten otros factores de riesgo o enfermedades graves, los cuales pueden interferir en el estudio o incluso hacer que el paciente abandone el estudio (p. ej. sujeto fallece por otra enfermedad cardíaca).
- **De clasificación.** Es probable que los sujetos ya presenten el evento de interés al inicio del estudio, dado que asisten al servicio de cardiología, pero que aún no se haya detectado.
- **De información.** Es posible que haya errores en la medición o tomas de datos en la etapa inicial del estudio, lo cual afecta la exposición de los sujetos y, por ende, afecta la relación con el outcome.

3. Ejercicio 3

Para analizar la distribución de las variables numéricas continuas, se utilizan métodos gráficos (Q-Q plot e histograma) y pruebas estadísticas (Test de Shapiro-Wilk y Test de Kolmogorov-Smirnov). Estos test toman como hipótesis nula que la variable analizada presenta una distribución normal, mientras que toman como hipótesis alternativa que la variable no cumple con la normalidad.

Con respecto al tipo de variable, el rol y la unidad, estos se obtuvieron de acuerdo a la consigna y a la base de datos provista.

Por último, para la elección de los test, que permiten realizar un análisis univariado, se tuvieron en cuenta la distribución, el tipo de variable y la independencia de las muestras. Dado el escenario y el diseño de la investigación, se puede inferir la independencia de las observaciones.

Variable	Tipo de variable	Rol	Unidad	Distribución	Test que utilizaría
Edad	Numérica	Explicatoria	años	No normal	Wilcoxon runk-sum
PAS	Numérica	Explicatoria	mmHg	No normal	Wilcoxon runk-sum
PAD	Numérica	Explicatoria	mmHg	No normal	Wilcoxon runk-sum
Colesterolemia	Numérica	Explicatoria	mg %	No normal	Wilcoxon runk-sum
Hombre	Dicotómica	Explicatoria	-	Binomial	Chi Cuadrado
Cig	Numérica	Explicatoria	$\frac{cigarrillos}{día}$	No normal	Wilcoxon runk-sum
Packs	Categórica ordinal	Explicatoria	-	-	Chi Cuadrado
EC	Dicotómica	Respuesta	-	Binomial	-

4. Ejercicio 4

Para el análisis univariado, se utilizan los Test de Wilcoxon rank-sum y Test de Chi Cuadrado, tal como se mencionó en el ejercicio anterior. El **Test de Wilcoxon rank-sum** tiene como hipótesis nula que las mediana de los dos grupos analizados son iguales, mientras que tiene como hipótesis alternativa que las medianas son significativamente diferentes. Por otro lado, el **Test de Chi Cuadrado** toma como hipótesis nula que todas las probabilidades de ocurrencia de las celdas de la tabla de contingencia son iguales entre sí, es decir, que las variables son independientes. La hipótesis alternativa establece que al menos una de esas probabilidades es distinta al resto, es decir, que las variables no son independientes.

Variable		No EC (0)	SI EC (1)	p-value
Edad [mediana (iqr)]		52.0 (8.00)	54.0 (8.00)	4.046e-06
PAS [mediana (iqr)]		140.0 (30.00)	154.5 (36.00)	2.475e-12
PAD [mediana (iqr)]		88.0 (16.00)	92.0 (20.00)	1.46e-08
Colesterolemia [mediana (iqr)]		230.0 (63.00)	232.0 (70.25)	0.04481
Hombre [% (n)]	0	56.22 % (615)	38.81 % (104)	4.455e-07
	1	43.78 % (479)	61.19 % (164)	
Cig [mediana (iqr)]		0.0 (15.00)	0.5 (20.00)	0.01382
Packs = 0 [% (n)]		56.12 % (614)	50.00 % (134)	0.03701
Packs = 1 [% (n)]		35.47 % (388)	36.94 % (99)	
Packs = 2 [% (n)]		8.41 % (92)	13.06 % (35)	

Para la variable **Edad**, el Test de Wilcoxon rank-sum da un p-value ($4{,}046e-06$) menor a 0.05, por lo que se rechaza la hipótesis nula y se acepta la alternativa. Esto quiere decir que hay diferencias significativa entre las edades de los grupos con y sin EC, con una diferencia en la mediana de 2 años (grupo sin EC es más joven).

Para la variable **PAS**, El Test de Wilcoxon rank-sum da un p-value ($2{,}175e-12$) menor a 0.05, por lo que se acepta la hipótesis alternativa y, por ende, hay diferencias significativas entre las presiones sistólicas de los grupos con y sin EC. El grupo con EC presenta una mediana con 14.5 mmHg más de presión.

Para la variable **PAD**, El Test de Wilcoxon rank-sum da un p-value ($1{,}46e-08$) menor a 0.05, por lo que se acepta la hipótesis alternativa y, por ende, hay diferencias significativas entre las presiones diastólicas de los grupos con y sin EC. El grupo con EC presenta una mediana con 4 mmHg más de presión.

Para la variable **colesterolemia**, El Test de Wilcoxon rank-sum da un p-value (0,04481)

levemente menor a 0.05, por lo que se acepta la hipótesis alternativa y, por ende, hay diferencias significativas entre las colesterolemias de los grupos con y sin EC. El grupo con EC presenta una mediana con 2 mg % más de colesterolemia.

Para la variable **cig**, El Test de Wilcoxon rank-sum da un p-value (0,01382) menor a 0.05, por lo que se acepta la hipótesis alternativa y, por ende, hay diferencias significativas entre la cantidad de cigarrillos fumados por día de los grupos con y sin EC. El grupo con EC presenta una mediana con 0.5 cigarrillos por día más.

Para la variable **hombre**, El Test de Chi Cuadrado da un p-value ($4{,}455e-07$) menor a 0.05, por lo que se acepta la hipótesis alternativa y, por ende, la variable en cuestión presenta una relación (dependencia) con la variable resultado. Las proporciones se puede observar en la tabla.

Para la variable **packs/y**, El Test de Chi Cuadrado da un p-value (0,03701) menor a 0.05, por lo que se acepta la hipótesis alternativa y, por ende, la variable en cuestión presenta una relación (dependencia) con la variable resultado. Las proporciones se puede observar en la tabla.

5. Ejercicio 5

Dado que la variable resultado es categórica y dicotómica, se debe utilizar un modelo de regresión logística para poder obtener la probabilidad de ocurrencia de dicha variable a partir de las variables explicatorias. De esta manera, se puede realizar un análisis multivariado de la base de datos que se está estudiando.

Antes de realizar el modelo saturado, se generan las variables dummy para la variable **packs/y**. Esto permite asignar un coeficiente para cada clase de la variable y poder comparar de manera individual el efecto que tienen estas clases en el modelo y en la variable respuesta.

Luego, se realiza el modelo lineal generalizado saturado, es decir, utilizando todas las variables del dataframe, con excepción de la variable **ID** y la variable **packs** sin modificar. Este modelo de regresión logística queda definida entonces por la siguiente ecuación:

$$logit(p) = \beta_0 + \beta_1 Edad + \beta_2 PAS + \beta_3 PAD + \beta_4 Col + \beta_5 Homb + \beta_6 cig + \beta_7 packs0 + \beta_8 packs1 + \beta_9 packs2 \quad (1)$$

donde p es la probabilidad de ocurrencia de la variable respuesta (EC).

El modelo generado contiene los valores de los coeficientes asociados a las variables y sus respectivos p-values, los cuales surgen del Test de hipótesis beta, el cual tiene como hipótesis nula que el coeficiente β_i es igual a cero, y tiene como hipótesis alternativa que el coeficiente es distinto de cero.

Después, se realiza el Test de Omnibus para ver la significancia o contribución que tienen las variables, con sus respectivos coeficientes, al modelo generado. Este Test tiene como hipótesis nula que todos los coeficientes son nulos, mientras que tiene como alternativa que al menos uno de los coeficientes no es nulo.

Por último, se calculan los Odds Ratio para las variables, con sus respectivos intervalos de confianza. Los OR proveen una medida de la asociación que hay entre una variable explicativa y el outcome.

En la siguiente tabla, se puede observar todos los resultados obtenidos para la serie de pasos mencionada.

Variable	Coef. Beta	p-value (regresión)	p-value (omnibus)	OR	IC del OR
Intercept	-8.768825	2.17e-14	-	-	-
Edad	0.057720	0.000234	2.02e-06	1.05942	1.02746 - 1.09269
PAS	0.015013	0.000187	3.41e-11	1.01513	1.00718 - 1.02321
PAD	0.004891	0.533586	0.2970	1.00490	0.98953 - 1.02053
Colest.	0.004614	0.003495	0.1169	1.00462	1.00151 - 1.00774
Hombre	0.906132	3.57e-08	3.82e-11	2.47473	1.79685 - 3.42536
Cig	0.009888	0.495009	0.0767	1.00994	0.98171 - 1.03919
Packs0	-0.070526	0.897272	0.7974	0.93190	0.32282 - 2.75700
Packs1	-0.079069	0.835520	0.8205	0.92398	0.43945 - 1.96057

En primer lugar, se puede observar que las variables PAD, cig, packs0 y packs1 presentan un p-value (regresión) mayor a 0.05, por lo que se acepta la hipótesis nula y, por ende, estos coeficientes son iguales a 0 (no significativos para el modelo). Por otro lado, el resto de las variables, incluida la ordenada al origen (β_0) presentan p-values menores a 0.05, por lo que se rechaza la hipótesis nula y se acepta la alternativa. Esto quiere decir que los coeficientes asociados a estas variables son significativos y no nulos: la ordenada al origen es -8.768825, el coeficiente para la edad es 0.057720, el coeficiente para la PAS es 0.015013, el coeficiente para la colesterolemia es 0.004614 y el coeficiente para hombre es 0.906132. Estos coeficientes indican la relación que existe entre sus respectivas variables y la probabilidad de ocurrencia de EC.

Con respecto a los p-values del omnibus, las variables que no son significativas para el modelo global son PAD, Colesterolemia, Cig, Packs0 y Packs1. El resto de las variables son significativas dado que el p-value es menor a 0.05.

Por último, se puede observar en los resultados de OR que las variables PAD, cig, packs0 y packs1 tienen intervalos de confianza que incluyen al 1. Esto quiere decir que dichas variables son independientes del outcome (EC), es decir, no significativas. Por otro lado, el resto de las variables no son independientes del outcome y son significativas. Para la edad, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que por cada año que aumenta la edad, el riesgo de sufrir una EC aumenta en un 6 % (o aumenta 1.06 veces). Para la PAS, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que por cada aumento unitario de PAS, el riesgo de sufrir una EC aumenta un 1.5 % (o aumenta 1.015 veces). Para la variable colesterolemia, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que por cada aumento unitario de colesterolemia, el riesgo de sufrir una EC aumenta en un 0.46 % (o 1.0046 veces). Para la variable Hombre, el OR es mayor a 1 (factor de riesgo) y su intervalo no lo incluye, por lo que se puede decir que ser hombre aumenta el riesgo de sufrir una EC en un 147 % (o aumenta 2.4747 veces).

Para saber que tan bien ajusta el modelo generado al conjunto de datos, se realiza la prueba de bondad de ajuste de Hosmer-Lemeshow. En función de que tan bien esta ajustado el modelo a los datos, se puede inferir que tan certero o acertado va a ser el modelo. Este test compara las frecuencias esperadas con las frecuencias observadas del modelo. Si las frecuencias observadas son similares a las frecuencias esperadas, entonces el modelo presenta un buen ajuste. La hipótesis nula que se plantea para esta prueba es que la diferencia entre frecuencias observadas y las esperadas es igual a cero (observado igual a esperado), mientras que la hipótesis alternativa indica que esta diferencia es distinta de cero.

Como el Test de bondad de ajuste de Hosmer-Lemeshow da un p-value (0.396) mayor a 0.05,

entonces la diferencia entre lo observado y lo esperado es igual a cero y, por ende, el modelo ajusta bien al conjunto de datos que se tiene.

6. Ejercicio 6

Se decide utilizar el método backward elimination para ir sacando aquellas variables que no son significativas para el modelo, es decir, que no tengan una contribución parcial. Para esto, se elimina la variable que tiene menor contribución, es decir, la variable que obtenga el p-value más grande. Estos p-values se obtienen de los test de hipótesis aplicados a los coeficientes, donde las hipótesis son análogas a las mencionadas en el modelo saturado.

Las variables que se eliminaron fueron (ver código para la secuencia de modelos) **packs0**, **packs1**, **packs2**, **PAD** y **cig**. De esta manera, se llegó a un modelo reducido que presenta las variables **Edad**, **PAS**, **Colesterolemia** y **Hombre**, y queda definido por la siguiente ecuación:

$$logit(p) = \beta_0 + \beta_1 Edad + \beta_2 PAS + \beta_4 Colesterolemia + \beta_5 Hombre \tag{2}$$

Una vez obtenido el modelo reducido, se evaluaron las mismas cosas que en el ejercicio anterior (coeficientes beta, p-value de regresión, p-value de omnibus y los Odds Ratio). Todos estos resultados se pueden observar en la siguiente tabla.

Variable	Coef. Beta	p-value (regresión)	p-value (omnibus)	OR	IC del OR
Intercept	-8.347166	<2e-16	-	-	-
Edad	0.051574	0.000689	2.01e-06	1.05293	1.02213 - 1.08491
PAS	0.016868	6.93e-12	3.38e-11	1.01701	1.01216- 1.02198
Colest.	0.004762	0.002461	0.1060	1.00477	1.00168 - 1.00788
Hombre	1.010859	3.78e-11	2.66e-11	2.74796	2.04267 - 3.72134

En primer lugar, se puede observar que todos los coeficientes asociados a las variables Edad, PAS, Colesterolemia y Hombre dan significativos (p-value menor a 0.05), con coeficientes iguales a 0.051574, 0.016868, 0004762 y 1.010859, respectivamente. Además, se obtiene una ordenada al origen significativa con un valor de -8.347166.

Con respecto a los p-values del omnibus, s epuede ver que todas las variables son significativas dado que el p-value es menor a 0.05, salvo la variable colesterolemia, la cual presenta un p-value mayor a 0.05 y, por ende, no es significativa para el modelo global.

Por último, se puede observar en los OR que ninguna de las variables tienen OR iguales a 1 o intervalos que incluyan al 1, por lo que todas las variables no son independientes del outcome y son significativas. Para la edad, se puede decir que por cada año que aumenta la edad, el riesgo de sufrir una EC aumenta en un 5.29 % (o aumenta 1.0529 veces). Para la PAS, se puede decir que por cada aumento unitario de PAS, el riesgo de sufrir una EC aumenta un 1.7 % (o aumenta 1.017 veces). Para la variable colesterolemia, se puede decir que por cada aumento unitario de colesterolemia, el riesgo de sufrir una EC aumenta en un 0.477 % (o 1.00477 veces). Para la variable Hombre, se puede decir que ser hombre aumenta el riesgo de sufrir una EC en un 174796 % (o aumenta 2.74796 veces).

Como el Test de bondad de ajuste de Hosmer-Lemeshow para este modelo da un p-value (0.2635) mayor a 0.05, entonces la diferencia entre lo observado y lo esperado es igual a cero y, por

ende, el modelo ajusta bien al conjunto de datos que se tiene.

Una vez obtenidos los dos modelos, para compararlos se utiliza el Test Likelihood ratio. Este test compara los valores de LogLik y analiza si la diferencia en estos valores de ajuste es significativa o no. La hipótesis nula indica que la diferencia es nula (no significativa), mientras que la alternativa indica que la diferencia no es nula y que es significativa.

Como el Test Likelihood ratio da un p-value (0.4763) mayor a 0.05, se acepta la hipótesis nula. Por lo tanto, la diferencia entre los dos valores LogLik es nula y no significativa, lo cual indica que los dos modelos ajusten de igual manera (o de manera similar).

Sin embargo, se decide elegir el segundo modelo dado que todas las variables son significativas, tiene menor cantidad de variables y presenta un LogLik más chico, por lo que este modelo ajusta mejor.

7. Ejercicio 7

En el análisis univariado se observó que la edad, la PAS, la PAD la colesterolemia y la cantidad de cigarrillos fumados por día presentan diferencias significativas al comparar los sujetos con y sin EC. Por otro lado, se observó que el sexo (ser hombre) y la categoría packs influyen en las ECs.

Luego, en el análisis multivariado se realizó un modelo de regresión logística saturado y se observó que las variables PAD, cig, packs0 y packs1 (variables dummy de packs) no son significativas para dicho modelo (p-values ¿0.05), mientras que el resto de las variables sí lo es (p-values ¡0.05). Además, se investigó las magnitudes de las relaciones existentes entre las variables significativas y el outcome mediante el cálculo de los Odds Ratio. Se observó que para la edad, por cada año que aumenta, el riesgo de sufrir una EC aumenta en un 6 % (o aumenta 1.06 veces). Para la PAS, por cada aumento unitario de PAS, el riesgo de sufrir una EC aumenta un 1.5 % (o aumenta 1.015 veces). Para la variable colesterolemia, por cada aumento unitario de colesterolemia, el riesgo de sufrir una EC aumenta en un 0.46 % (o 1.0046 veces). Para la variable Hombre, el hecho de ser hombre aumenta el riesgo de sufrir una EC en un 147 % (o aumenta 2.4747 veces). Además, gracias al Test de bondad de ajuste de Hosmer-Lemeshow, se probó que el modelo generado presenta un buen ajuste al conjunto de datos (p-value = 0.396).

Como segundo modelo, se propuso uno de regresión logística que contenga únicamente variables significativas. Para obtener este modelo, se realizó el backward elimination, lo que terminó en un modelo con las variables significativas edad, PAS, colesterolemia y Hombre (p-values ¡0.05). Además, mediante el cálculo de los OR, se observó que las cuatro variables presentan una relación con el outcome (EC). Para la edad, por cada año que aumenta, el riesgo de sufrir una EC aumenta en un 5.29 % (o aumenta 1.0529 veces). Para la PAS, por cada aumento unitario, el riesgo de sufrir una EC aumenta un 1.7 % (o aumenta 1.017 veces). Para la variable colesterolemia, por cada aumento unitario, el riesgo de sufrir una EC aumenta en un 0.477 % (o 1.00477 veces). Para la variable Hombre, el hecho de ser hombre aumenta el riesgo de sufrir una EC en un 175 % (o aumenta 2.74796 veces). También se probó, mediante el Test de Hosmer-Lemeshow que el modelo generado presenta un buen ajuste a los datos (p-value = 0.2635).

Por último, mediante el Test Likelihood Ratio, se comprobó que no hay diferencia significativa entre los dos modelos en cuanto al ajuste a los datos (p-value = 0.4763). Sin embargo, se decidió elegir el segundo modelo ya que todas las variables son significativas, tiene menor cantidad de variables y presenta un LogLik más chico (-621.43 contra -619.67), es decir, un mejor ajuste.