

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

Facultad de Matemáticas, Departamento de Estadística

**DIPLOMADO EN ESTADISTICA – VERSION 2023**

Control Regresión Logística

Nombre: Martín Venegas Márquez

31 de agosto de 2023

**\*NOTA DEL ESTUDIANTE\*: Considerando el formato de la prueba, marqué en azul todo lo que correspondiese a una respuesta mía y así delimitar qué es parte de la prueba y que es una respuesta mía. Espero facilite la revisión.**

**1 - {3 puntos} Cálculo de OR y ajuste de regresión logística. ¿Sexo implica Fumar?**

1.1 Se recoge información respecto al fumar mediante entrevistas a 100 alumnos. Parte de los resultados se presentan en la siguiente tabla (los datos están en **fumaedad.xlsx**):

|  |  |  |
| --- | --- | --- |
| Sexo | SI Fuma | No Fuma |
| Masc | 34 | 16 |
| Femen | 24 | 26 |

P(Fuma | Masc) = \_\_\_\_\_\_\_\_**68%**\_\_\_\_ {0.1 puntos}.

P(Fuma | Fem) = \_\_\_\_\_\_\_\_**48%**\_\_ {0.1 puntos}.

P(No Fuma | Masc) = \_\_\_\_\_\_\_\_**32%**\_\_\_\_ {0.1 puntos}.

P(No Fuma | Fem) = \_\_\_\_\_\_\_\_\_**52%**\_ {0.1 puntos}.

1.2 ¿Cuál es la hipótesis de interés? {0.1 pts}.

Hint: Es la hipótesis de un test Chi-Cuadrado bajo nuestro contexto.

Las hipótesis son las siguientes:

H0*:*Fumar es independiente del sexo

H1:Fumar depende del sexo

El test Chi-Cuadrado resulta ser 2 = \_\_\_\_\_\_4.1051\_\_\_ {0.3 pts} con valor-p = \_\_\_\_.04275\_\_\_\_\_ {0.3 pts}. ¿Test 2 con o sin corrección?: \_\_\_\_\_\_SIN CORRECCIÓN\_\_\_ {0.1 pts}

En conclusión, con un nivel de confianza del 95%, existe evidencia suficiente para rechazar la hipótesis nula de independencia entre sexo y fumar (valor p = .043 < .05). Es decir, existe una relación estadísticamente significativa entre el sexo de los estudiantes de la muestra y el fumar {0.1 puntos}.

1.3 Con el objeto de cuantificar esa asociación se obtiene de la tabla

OR 🡪 \_\_\_\_\_\_\_2.30\_\_\_ {0.3 pts}. Es decir, los hombres tienen 2.3 veces más chances de fumar que las mujeres. Dicho de otra forma, los hombres tienen un 130% más de chances de fumar que las mujeres. {0.3 pts}.

1.4 ¿Qué otras variables son relevantes incluir en el análisis (indique al menos dos)?

1. Un potencial factor relevante podría ser la edad. En la revisión sistemática de Konrad et al (1992) se encontró un apoyo en la literatura a una relación positiva entre la edad y fumar. Una posible explicación es que a mayor edad se tiene más facilidades para adquirir el hábito (e.g. mayores recursos). {0.1 puntos}
2. Estudios más recientes han mostrado que un factor relevante podría ser la cantidad de amigos cercanos que fuman (Cheng et al., 2022), en la medida que las personas (y jóvenes en particular) suelen adquirir y aprender hábitos y prácticas de sus redes más cercanas. {0.1 puntos}

Referencias:

Conrad, K. M., Flay, B. R., & Hill, D. (1992). Why children start smoking cigarettes: predictors of onset. British journal of addiction, 87(12), 1711-1724.

Cheng, X., Guo, X., & Jin, C. (2022). Social determinants of smoking among school adolescents in Beijing, China. Tobacco Induced Diseases, 20.

1.5 Ahora, verifique el OR obtenido a través de una regresión logística con R (o software), revise bien los datos:

Por tanto, la probabilidad estimada:

P(Fumar | Femen) = exp(a) / {1+exp(a)} = .48, siendo a = Intercepto = -.08004271. {0.3 puntos}

P(Fumar | Masc) = exp(a+b) / {1 + exp(a+b)} = .68, siendo b = log odds para masculino = .83381. {0.3 puntos}

Recordemos que de la tabla 2x2 se tiene OR = 2.30 🡨 🡪 exp(beta) = exp(.83381) = 2.302073 {0.3 puntos}

Hint: a y b son valores/coeficientes que deben identificar en su regresión.

2. **{3 puntos}** Como usted determinó, un factor relevante y ausente del análisis podría ser la edad. Así, que los datos anteriores son presentados, incluyendo la edad, en formato “por caso - línea”, y se han recodificado para efecto de ajustar e interpretar la regresión logística. Se ha usado FUMA=1 implica “SI” y FUMA=2 implica “NO”. De igual forma, se recodifico SEX=1 cuando es MASC y SEX=0 en caso contrario (es decir, Femen). Así, se interpreta la presencia (Sex=1) con respecto a la ausencia (sex=0).

Los datos de acuerdo con esta estructura se encuentran en **fumaedad.xlsx** – hoja fumar (ver encabezado)

|  |  |  |  |
| --- | --- | --- | --- |
| caso | FUMA | sex | edad |
| 1 | 1 | 1 | 18 |
| 2 | 1 | 1 | 21 |
| 3 | 1 | 1 | 24 |
| 4 | 1 | 1 | 21 |

La idea es que en esta segunda etapa incluya en los análisis EDAD, dado que se piensa que es un factor relevante.

A realizar. Se propone realizar los siguientes ajustes

1. Fuma 🡨 Sexo
2. Fuma 🡨 Edad ¿edad en forma continua o categórica? \_\_\_\_\_\_\_\_\_continua\_\_\_\_\_\_\_\_.
3. Fuma 🡨 sexo + edad

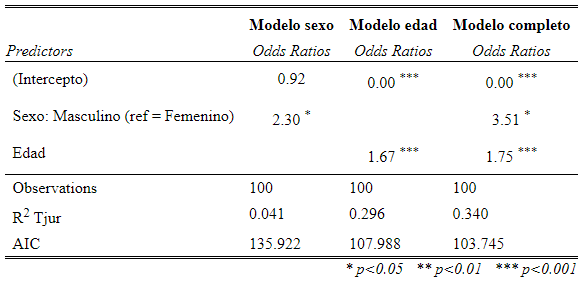
Para a), b) y c) resuma e interprete los resultados obtenidos. Indicando si sexo y/o edad es factor de riesgo o protector {0.3 puntos}, evalué el OR y su interpretación {0.3 puntos} e indique significancia {0.3 puntos}.

Finalmente ¿qué modelo prefiere? {0.3 puntos}

**RESPUESTA:**

Previo a realizar los análisis, es importante señalar que la variable respuesta (FUMA) fue recodificada de tal manera que *0 = No Fuma y 1 = Fuma*. Esta recodificación se realizó con el fin de asegurarse que el éxito correspondiese al hecho de fumar y así interpretar correctamente los coeficientes.

Tabla 1. Modelos de regresión logística binaria para fumar (0 = No Fuma, 1 = Fuma)



En la Tabla 1 se muestran los tres modelos solicitados: un modelo simple que incluye solo la variable sexo, un modelo simple que incluye solo la variable de edad y un modelo completo que incluye ambas variables. A partir de la información de los *Odds Ratios* (OR), se pueden efectuar las siguientes interpretaciones:

* En el modelo simple de sexo, se observa una relación estadísticamente significativa al 95% de confianza entre sexo y fumar, en tanto los hombres tienen 2.3 veces más chances de fumar que las mujeres.
* En el modelo simple de edad, también se observa una relación estadísticamente significativa al 95% de confianza (e incluso al 99.9% de confianza) entre edad y fumar, en donde por cada año las chances de fumar aumentan en 67%.
* En el modelo completo la significancia de ambas variables se mantiene, sin embargo, las magnitudes de los OR aumentan. En el caso de sexo, los hombres tienen 3.51 veces más chances de fumar que las mujeres, asumiendo la edad constante. Asimismo, por cada año de edad las chances de fumar aumentan en 75%, controlando por el sexo de los estudiantes.
* En suma, tanto el sexo (ser hombre) como la edad son factores de riesgo para fumar.

En lo que respecta a las medidas de ajuste, en la Tabla 1 se observan dos estadísticos: AIC y R2 Tjur. Ambos indican que el modelo de edad es mejor que el modelo de sexo y que el modelo completo es mejor que ambos modelos simples. En el caso del AIC, esto se evidencia en que el valor en el modelo completo es menor al de los otros modelos. En el caso del R2 Tjur, se evidencia en la medida que el valor es mayor al de los otros modelos.

Con tal de corroborar el ajuste, también se calculó otro tipo de Pseudo R2 (Nagelkerke), el cual confirma la tendencia: el modelo de edad (R2 Nagelkerke = .369) ajusta mejor que el modelo de sexo (R2 Nagelkerke = .055) y el modelo completo (R2 Nagelkerke = .428) ajusta mejor que los modelos anteriores.

Por último, a modo de corroboración final, se compararon los distintos modelos en términos de su capacidad predictiva a partir de la métrica de área bajo la curva (AUC). Para ello, se dividió la muestra en razón 70-30 para la data de entrenamiento y validación, respectivamente. Luego, se estimaron seis curvas ROC, una para cada modelo por cada data. A partir de la Figura 1, se puede concluir lo siguiente:

* Se mantiene la tendencia, el modelo completo presenta mejor ajuste que los demás. En detalle, la data de entrenamiento muestra que el modelo completo puede ser clasificado de excelente (80% =< AUC < 90%). Sin embargo, el modelo que solo incluye edad sigue siendo aceptable (70% =< AUC < 80%).
* Independiente de que el modelo completo tenga un mejor ajuste general, se observa que el modelo de edad presenta algunos puntos de corte que maximizan la sensibilidad con una menor tasa de falsos positivos (TVP ≈ 85%; TFP ≈ 45%) en comparación al modelo completo (TVP ≈ 90%; TFP ≈ 60%). Por lo que, si se buscara priorizar la sensibilidad, podría ser más recomendable escoger el modelo de edad.
* La data de validación confirma la tendencia, sin embargo, se observa que los AUC de cada modelo son mucho más extremos entre sí, lo cual se puede deber a la baja cantidad de datos (n = 30). Sería recomendable recolectar más datos en aras de obtener un mejor modelo.

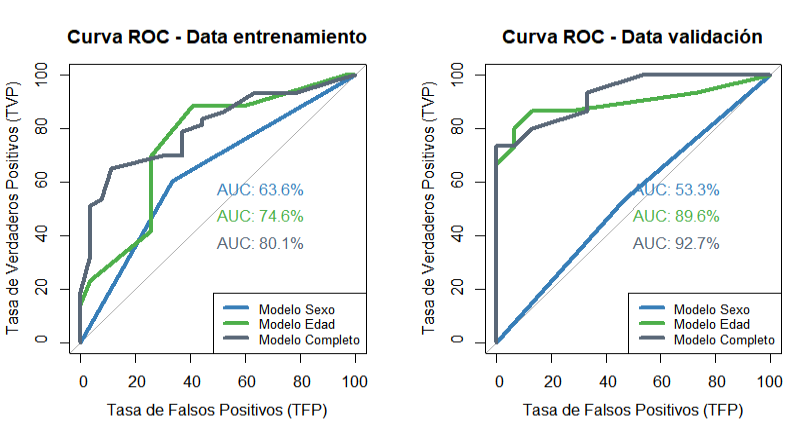


Figura 1. Curvas ROC para modelos logísticos en data de entrenamiento y data de validación

Indicaciones

* Redacte un informe (pregunta/respuesta) que permita responder cada punto indicado.
* NO coloque el script en el informe (solo tablas, gráficos, métricas de interés, valores resultantes que ayuden a respaldar sus respuestas, no relleno).
* Por ningún motivo el último taller es pauta para este control, recaude toda la información pertinente (teórica como práctica).
* El trabajo es individual, pero puede ser grupal también (hasta 3 personas, sin más excepciones).
* El informe completo se entrega en el buzón “Entrega Final” (disponible en la plataforma del curso). No olvide poner el nombre de los miembros del grupo en el encabezado del informe!.
* **Fecha de entrega: 02 de septiembre hasta las 18:00 hrs.**
* Se debe entregar si o si un avance el 31 de agosto hasta las 21:30 hrs.