REGRESIÓN LOGÍSTICA Y CURVAS ROC

MARCELA LILIANA NATAL

Resumen de la Tesis de Magister Scientiae en Biometría defendida el 22 de diciembre de 2000

Introducción: El valor umbral de los tests de laboratorio para discriminar entre sujetos sanos y enfermos es la decisión más importante en el área del diagnóstico clínico. Las curvas ROC (receiver operating characteristic) son un instrumento idóneo para facilitar esta toma de decisión. Un enfoque estadístico para el test de diagnóstico consiste en el empleo de técnicas predictivas de regresión no lineal basadas en la distribución logística. Hui y Miller (1991) y Lemeshow (1996), entre otros, utilizaron curvas ROC para refinar la inferencia estadística de los modelos de regresión logística y aplicaron la metodología ROC a la variable respuesta. Sintetizando las curvas ROC sólo son empleadas para la variable dependiente. Cuando las covariables intervinientes en un modelo de en regresión logística son continuas, en ocasiones, por razones de interés clínico, se las suele categorizar para el análisis. En este caso, la inferencia estadística no es invariante a la categorización utilizada en el análisis. El objetivo principal de la tesis es: proponer una metodología de trabajo no convencional para la aplicación de la **curva ROC** en la determinación del **punto de corte** sobre **variables explicativas continuas,** en situaciones en que la variable respuesta pueda ser analizada mediante **regresión logística.**

Metodología: Se utiliza la metodología ROC tanto para definir el punto de corte en una variable continua como para evaluar la capacidad del modelo logístico para discriminar entre los éxitos o fracasos en función de los diferentes valores de probabilidad. Los índices de precisión derivados a partir del análisis ROC que se usan son el área bajo la curva ROC, el largo proyectado de la curva (PLC) y el área "barrida" por la curva ROC (ASC). Para el cáculo de los índices ASC y PLC, se desarrolló un programa en lenguaje Pascal. El Modelo de Regresión Logística se usa para describir la relación de una o más variables explicativas, con una variable dependiente dicotómica. Se construyó el modelo logístico. Mediante el programa: ROCKIT (1998) se construye la curva ROC para variables continuas usando estimadores de máxima verosimilitud, suponiendo modelos bi-normales; se usó el paquete BMDP, versión 7.0 para DOS, para la construcción de los modelos logísticos y una planilla de cálculos para graficar. Se utiliza una base de datos correspondientes a pacientes infectados con el virus de HIV atendidos por la Unidad de Infectología del Hospital Interzonal de la ciudad de Mar del Plata desde 1986 hasta julio de 1997.

Resultados: Las variables categóricas: SEXO, RIESGO1, OMSA1, SIDA (1: enfermo, 0: no enfermo) y las variables continuas son EDAD y CD4A. Se pretende predecir el riesgo de enfermedad de SIDA ajustando un modelo de regresión logística considerando la variable dependiente SIDA y como variables independientes a: SEXO, EDAD, CD4A, RIESGO1. Se construyó el modelo considerando a CD4A como variable continua y luego utilizando las curvas ROC se determinó un punto de corte en la variable continua CD4A. Se determinó un punto de corte en la variable CD4 que resultó ser 200. Se ajustaron distintos modelos de regresión logística por el método paso a paso discretizando la variable CD4A según los distintos puntos de corte propuestos. Para la selección del modelo se consideró el valor p del ajuste, el test de Hosmer, el área bajo la curva ROC (en vble. Predicha), el porcentaje de clasificaciones correctas, el índice PLC y ASC. El modelo seleccionado fue el que consideró la variable CD4A discretizada con el punto de corte 200.

Conclusiones: El punto de corte obtenido utilizando la curva ROC para discretizar la variable CD4A coincide con el utilizado por criterio médico para determinar si el paciente pasa de la situación de portador a enfermo. La metodología propuesta fue aplicada en un estudio sobre el Mal de Chagas en Argentina obteniéndose similares conclusiones. Los pasos de la metodología propuesta son: a)Ajuste del modelo de regresión logística con variable continua; b) determinación del punto de corte en la variable explicativa continua a través de la curva ROC; c) ajuste del modelo logístico empleando la variable explicativa discretizada; d) comparación de los modelos obtenidos utilizando tanto metodología ROC como análisis clásico de regresión. Basados en los funda-mentos teóricos de las curvas ROC, sus bases estadísticas y en nuestra investigación queda establecido que la curva ROC es un medio adecuado para determinar un punto de corte en una variable explicativa de la Regresión Logística, proponiéndose esta metodología de investigación para resolver problemas análogos en otras ciencias.