Departamento de Estadística e I.O.

Máster en Estadística Aplicada

UNIVERSIDAD
DE GRANADA

MODELOS DE RESPUESTA DISCRETA APLICACIONES BIOSANITARIAS

Tema 2
Guía de trabajo autónomo

Profesores

Ana María Aguilera del Pino

Manuel Escabias Machuca

Título original: Modelos de Respuesta Discreta. Aplicaciones Biosanitarias. Guía de Trabajo Autónomo del Tema 2

Índice general

Capítulo 1

Guía de Trabajo Autónomo del Tema 2

Para asimilar los contenidos del Tema 2 el alumno debe estudiar los apuntes de teoría del tema y realizar las actividades de trabajo autónomo de esta guía. Estas actividades tienen como objetivo facilitar el autoaprendizaje del alumno y la consecución de los objetivos del tema. Es conveniente que el alumno las trabaje individualmente y plantee en el foro sus posibles dudas e inquietudes relacionadas con las mismas.

1.1. Justificación del Tema 2

Este capítulo está dedicado al estudio detallado de los modelos de regresión logística en el caso de variables explicativas cuantitativas observadas sin error. Como es usual en modelización estadística, se comenzará por la formulación del modelo e interpretación de sus parámetros, después se abordará el problema de su estimación y contrastes de bondad de ajuste, y finalmente se llevará a cabo la validación y selección del modelo más apropiado. También se realizarán ajustesde modelos logit con datos reales mediante el uso del software libre R.

1.2. Objetivos del Tema 2

- Estudiar la formulación, interpretación, estimación y validación del modelo de regresión logística a partir de una o varias variables cuantitativas relacionadas.
- Interpretar la relación entre la variable de respuesta y las explicativas

en términos de cocientes de ventajas.

- Manejar los métodos stepwise para seleccionar el modelo logit más adecuado (con o sin interacción) para explicar una variable cualitativa a partir de varias variables relacionadas con ella.

- Aprender a manejar un el software estadístico R para aplicar a datos biomédicos o en cualquier otro campo los modelos estadísticos estudiados.

1.3. Actividades del Tema 2

1. Haga una lectura comprensiva de los apuntes del Capítulo 2 para obtener una visión global del tema.

2. Explique detalladamente la relación entre la ventaja y la probabilidad de un suceso. ¿Cuánto vale la probabilidad de un suceso cuya ventaja es 50? ¿conoce algún ejemplo de acontecimiento de la vida real en el que la incertidumbre se evalúe en términos de ventajas en lugar de probabilidades?

3. En el modelo logit simple, ¿qué significa que el cociente de ventajas de respuesta Y=1 cuando se incrementa en una unidad la variable explicativa es igual a la unidad? ¿cuánto vale el parámetro beta del modelo en este caso? ¿cómo se interpreta la exponencial del parámetro constante?

4. En un modelo logit múltiple, ¿qué representa la exponencial de cada uno de los parámetros asociados a cada una de las variables explicativas del modelo? ¿qué significa que dicha exponencial valga 1 para una de las variables? 5. Consideremos un estudio epidemiológico en el que se quiere estudiar la relación entre el infarto de miocardio (enfermedad) y el tabaquismo (factor de riesgo) controlando a su vez el peso (p). Para ello se han estimado dos modelos logit uno con sólo la variable número de cigarrillos fumados (x) y otro con las dos variables (x,p) Modelo 1: ln[p(x,p)/[1-p(x,p)]] = 1+ x Modelo 2: ln[p(x,p)/[1-p(x,p)]] = 1+2x+p ¿Es la variable peso un factor de confusión para el estudio de la relación entre el infarto y el tabaco? ¿es el peso un modificador del efecto del tabaco sobre el infarto? ¿en base al modelo ajustado, qué relación hay entre el infarto y el tabaco? ¿y entre el infarto y el peso? Justifique las respuestas.

5. Estudie el los párrafos 4.1 y 4.2 relativos a la estimación por máxima verosimilitud de los parámetros de un modelo logit. ¿Qué analogía hay entre las ecuaciones de verosimilitud del modelo logit y las del modelo lineal general? ¿bajo que condiciones no existen los estimadores de máxima verosimilitud? Inventa un ejemplo sencillo de una muestra de una variable de respuesta binaria y una explicativa cuantitativa para la que no existan los estimadores MV del modelo logit. ¿Por qué es necesario utilizar el método de Newton-Raphson para resolver las ecuaciones de verosimilitud?

6. Lea detenidamente el párrafo 4.3 en el que se explica el procedimiento de estimación por mínimos cuadrados ponderados del modelo logit. ¿Cree que es posible usar este método de estimación con cualquier muestra? ¿qué ocurre cuando hay una única observación en cada combinación diferente de valores observados de las variables explicativas?

7. ¿Es conveniente usar el test de Hosmer y Lemeshow para contrastar la bondad del ajuste de un modelo logit? ¿qué diferencias y similitudes hay entre el test Chi-Cuadrado de bondad de ajuste y el de Razón de Verosimilitudes?

8. ¿Qué relación hay entre el contraste de Wald con nivel de significación de igualdad a cero de un parámetro del modelo logit (basado en la distribución normal asintótica del parámetro) y el intervalo de confianza de nivel $(1-\alpha)$ para dicho parámetro?

9. ¿Cómo se identifican datos con falta de ajuste? ¿qué se hace para evaluar la estabilidad y robustez de los parámetros estimados frente a datos con falta de ajuste?

10. ¿Cuál es el cutpoint más adecuado para discriminar entre respuestas Y=1 e Y=0 en el modelo logit? ¿qué considera más adecuado para estimar la capacidad del modelo para discriminar, la tasa de clasificaciones correctas o la currva ROC? Razone su respuesta.

11. En los métodos de selección paso a paso (stepwise) de modelos logit, ¿por qué se toma el nivel de significación del contraste de entrada de variables algo menor que el de salida?

12. Lea y estudie en profundidad, ejecutando las sentencias con R, la guía de prácticas de este tema en la que se desarrollan varias aplicaciones del modelo logit simple y múltiple con datos agrupados y no agrupados.

Para afrontar con éxito las actividades de evaluación, es muy importante que domine tanto el manejo de R para obtener los resultados numéricos de estimación de los modelos como la interpretación de los mismos.

13. Elabore sus propios apuntes sobre como se realiza la estimación, inferencia y selección paso a paso de modelos logit con variables explicativas cuantitativas con R, y utilícelos para resolver los ejercicios que se proponen a continuación. Plantee cualquier tipo de duda o comentario sobre su resolución en el foro de este tema.

14. Se cree que la causa de la explosión de la lanzadera espacial americana Challenger pudo ser la temperatura del momento del despegue que fué de 31 grados Fahrenheit. Para contrastar estadísticamente esta hipótesis se dispone de los datos de la siguiente tabla que corresponden a 23 simulacros de lanzamientos realizados previamente al lanzamiento real:

Temperatura	Simulacro
53	1
57	1
58	1
63	1
66	0
67	0
67	0
67	0
68	0
69	0
70	0
70	0
70	1
70	1
72	0
73	0
75	0
75	1
76	0
76	0
78	0
79	0
81	0

Teniendo en cuenta que, el resultado de cada simulacro de lanzamiento es 1 si se produce algún fallo y 0 si no se produce ninguno, se pide

a) Ajustar un modelo de regresión logística para explicar la probabilidad de fallo en función de la temperatura del despegue. Interpretar sus parámetros en términos de cocientes de ventajas y estudiar su significación estadística mediante test de Wald e intervalos de confianza del 95 %.

b) Contrastar la bondad del ajuste del modelo ajustado mediante el test más adecuado y validarlo mediante un análisis de residuos y medidas de influencia.

c) Construir, para el modelo ajustado, un test de diagnóstico óptimo para decidir si el lanzamiento de un cohete será fallido en función de la temperatura en el momento de lanzamiento. Calcular la tasa de verdaderos positivos y falsos positivos asociada.

d) Obtener e interpretar la curva ROC.

15. Se desea estudiar si el nivel de ácido fosfórico (AF) en sangre en enfermos de cáncer de próstata determina o no el desarrollo de nuevos nódulos. Para ello se dispone de los datos de la siguiente tabla, donde la variable Y toma el valor 1 para aquellos individuos con nuevos nódulos de cáncer y el valor 0 para los que no los tienen, se pide:

Y	AF	Edad	Y	AF	Edad	Y	AF	Edad	Y	AF	Edad
0	48	66	0	47	67	0	50	61	1	81	50
0	56	68	0	49	50	0	40	64	1	76	60
0	50	66	0	50	56	0	55	52	1	70	45
0	52	56	0	78	60	0	59	66	1	78	56
0	50	58	0	83	52	1	48	58	1	70	45
0	49	60	0	98	56	1	51	57	1	67	67
0	46	65	0	52	57	1	49	65	1	82	63
0	62	61	0	75	63	0	48	65	1	67	57
1	56	50	1	99	59	0	63	59	1	72	51
0	55	49	0	102	61	1	89	64	0	62	61
1	136	61	0	76	53	1	126	68	0	71	58
1	82	63	0	95	67	0	65	51	0	40	63
0	66	53	1	67	67	0	50	64	1	84	65

a) Ajustar un modelo de regresión logística simple para explicar la probabilidad de un nuevo brote de cáncer en función del nivel de ácido fosfórico.

b) Estudiar la bondad del ajuste del modelo ajustado y realizar una validación completa de modelo mediante un análisis de residuos y medidas de influencia.

c) Interpretar los parámetros del modelo en términos de cocientes de ventajas, construyendo para estos últimos intervalos de confianza de nivel 99 %. Estimar el cambio que se produce en la ventaja de tener un brote por cada aumento de diez unidades en el nivel de ácido fosfórico, proporcionando también un intervalo de confianza aproximado para el valor poblacional asociado.

d) Estimar la tasa de verdaderos positivos y de falsos positivos del test de diagnóstico asociado al modelo ajustado, así como la tasa de clasificaciones correctas con cutpoint 0.5. Razonar si este es el cutpoint más adecuado para discriminar entre individuos a los que se estima que desarrollarán de nuevo la enfermedad y aquellos que no lo harán.

e) Estimar el nivel de ácido fosfórico para el que la probabilidad estimada de desarrollar nuevos nódulos de cáncer es el cutpoint óptimo (aquel que proporciona la máxima tasa de clasificaciones correctas). Obtener además la curva ROC e interpretarla.

f) Realizar un contraste condicional de razón de verosimilitudes para decidir si el modelo que contiene como variables explicativas la edad y el nivel de ácido fosfórico es más adecuado que el contiene sólo el nivel de ácido fosfórico? ¿puede considerarse la edad un factor de confusión en la relación entre la aparición de nuevos nódulos de cáncer y el nivel de ácido fosfórico?

16. Analizar mediante un modelo logit los datos de la siguiente tabla sobre ocurrencia de vasoconstricción (1 indica constricción y 0 lo contrario) en la piel de los dedos en función de la tasa y el volumen de aire respirado:

Constricción	Volumen	Tasa	Constricción	Volumen	Tasa
1	0.825	3.7	0	2.0	0.4
1	1.09	3.5	0	1.36	0.95
1	2.5	1.25	0	1.35	1.35
1	1.5	0.75	0	1.36	1.5
1	3.2	0.8	1	1.78	1.6
1	3.5	0.7	0	1.5	0.6
0	0.75	0.6	1	1.5	1.8
0	1.7	1.1	0	1.9	0.95
0	0.75	0.9	1	0.95	1.9
0	0.45	0.9	0	0.4	1.6
0	0.57	0.8	1	0.75	2.7
0	2.75	0.55	0	0.03	2.35
0	3.0	0.6	0	1.83	1.1
1	2.33	1.4	1	2.2	1.1
1	3.75	0.75	1	2.0	1.2
1	1.64	2.3	1	3.33	0.8
1	1.6	3.2	0	1.9	0.95
1	1.415	0.85	0	1.9	0.75
0	1.06	1.7	1	1.625	1.3
1	1.8	1.8			

a) Seleccionar paso a paso el modelo de regresión logística más adecuado para explicar la probabilidad de padecer vasoconstricción en función de la tasa y el volumen de aire respirados. Estudiar la bondad del ajuste del modelo ajustado y realizar una validación completa de modelo mediante un análisis de residuos y medidas de influencia.

b) Interpretar los parámetros del modelo en términos de cocientes de ventajas, construyendo para estos últimos intervalos de confianza de nivel 99 %. Estimar el cambio que se produce en la ventaja de padecer vasoconstricción en función de la tasa y del volumen de aire respirados.

c) Estimar la tasa de verdaderos positivos y de falsos positivos del test de diagnóstico asociado al modelo ajustado, así como la tasa de clasificaciones correctas con cutpoint 0.5. Razonar si este es el cutpoint más adecuado para discriminar entre individuos a los que se estima que desarrollarán de nuevo la enfermedad y aquellos que no lo harán. Obtener la curva ROC e interpretarla.