## 1.3. Actividades del Tema 2

1. Haga una lectura comprensiva de los apuntes del Cap´ıtulo 2 para obtener una visio´n global del tema.
2. Explique detalladamente la relacio´n entre la ventaja y la probabilidad de un suceso. ¿Cua´nto vale la probabilidad de un suceso cuya ventaja es 50? ¿conoce algu´n ejemplo de acontecimiento de la vida real en el que la incertidumbre se evalu´e en t´erminos de ventajas en lugar de probabilidades?

Probabilidad valores de cero a uno, ventaja de cero a infinito

La ventaja puede ser expresada como el cociente de las probabilidades a favor y en contra de un suceso.

Si la ventaja es 50, el valor de prob. Del suceso debe ser p/(1-p) = 50, al resolver obtenemos que p=0.983

Si en las acciones de la bolsa, que suba o baje la acción, o en el estudio del efecto de dos métodos médicos.

1. En el modelo logit simple, ¿qu´e significa que el cociente de ventajas de respuesta Y=1 cuando se incrementa en una unidad la variable explicativa es igual a la unidad? ¿cua´nto vale el par´ametro beta del modelo en este caso? ¿c´omo se interpreta la exponencial del par´ametro constante?

La respuesta es independiente de la variable explicativa.

Beta vale cero

La exponencial del parámetro B es el cociente de ventajas de respuesta

1 para dos valores de X que se diferencien en una unidad.

1. En un modelo logit mu´ltiple, ¿qu´e representa la exponencial de cada uno de los para´metros asociados a cada una de las variables explicativas del modelo? ¿qu´e significa que dicha exponencial valga 1 para una de las variables?
2. 5. Consideremos un estudio epidemiol´ogico en el que se quiere estudiar la relacio´n 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 s´olo la variable nu´mero 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 confusio´n para el estudio de la relacio´n entre el infarto y el tabaco? ¿es el peso un modificador del efecto del tabaco sobre el infarto? ¿en base al modelo ajustado, qu´e relacio´n hay entre el infarto y el tabaco? ¿y entre el infarto y el peso? Justifique las respuestas.
3. Estudie el los pa´rrafos 4.1 y 4.2 relativos a la estimacio´n por ma´ximaverosimilitud de los par´ametros de un modelo logit. ¿Qu´e analog´ıa hay entre las ecuaciones de verosimilitud del modelo logit y las del modelo lineal general? ¿bajo que condiciones no existen los estimadores de ma´xima 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 qu´e es necesario utilizar el m´etodo de Newton-Raphson para resolver las ecuaciones de verosimilitud?
4. Lea detenidamente el pa´rrafo 4.3 en el que se explica el procedimiento de estimaci´on por m´ınimos cuadrados ponderados del modelo logit. ¿Cree que es posible usar este m´etodo de estimacio´n con cualquier muestra? ¿qu´e ocurre cuando hay una u´nica observaci´on en cada combinacio´n diferente de valores observados de las variables explicativas?
5. ¿Es conveniente usar el test de Hosmer y Lemeshow para contrastar labondad del ajuste de un modelo logit? ¿qu´e diferencias y similitudes hay entre el test Chi-Cuadrado de bondad de ajuste y el de Razo´n de

Verosimilitudes?

1. ¿Qu´e relacio´n hay entre el contraste de Wald con nivel de significacio´nde igualdad a cero de un par´ametro del modelo logit (basado en la distribucio´n normal asinto´tica del par´ametro) y el intervalo de confianza de nivel (1 − *α*) para dicho para´metro?
2. ¿C´omo se identifican datos con falta de ajuste? ¿qu´e se hace para evaluar la estabilidad y robustez de los para´metros estimados frente a datos con falta de ajuste?
3. ¿Cu´al es el cutpoint ma´s adecuado para discriminar entre respuestasY=1 e Y=0 en el modelo logit? ¿qu´e considera ma´s adecuado para estimar la capacidad del modelo para discriminar, la tasa de clasificaciones correctas o la currva ROC? Razone su respuesta.
4. En los m´etodos de seleccio´n paso a paso (stepwise) de modelos logit,¿por qu´e se toma el nivel de significacio´n del contraste de entrada de variables algo menor que el de salida?
5. Lea y estudie en profundidad, ejecutando las sentencias con R, la gu´ıade pra´cticas de este tema en la que se desarrollan varias aplicaciones del modelo logit simple y mu´ltiple con datos agrupados y no agrupados.

Para afrontar con ´exito las actividades de evaluacio´n, es muy importante que domine tanto el manejo de R para obtener los resultados num´ericos de estimacio´n de los modelos como la interpretacio´n de los mismos.

1. Elabore sus propios apuntes sobre como se realiza la estimaci´on, inferencia y selecci´on paso a paso de modelos logit con variables explicativas cuantitativas con R, y util´ıcelos para resolver los ejercicios que se proponen a continuaci´on. Plantee cualquier tipo de duda o comentario sobre su resolucio´n en el foro de este tema.
2. Se cree que la causa de la explosi´on de la lanzadera espacial americanaChallenger pudo ser la temperatura del momento del despegue que fu´e de 31 grados Fahrenheit. Para contrastar estad´ısticamente esta hip´otesis 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 algu´n fallo y 0 si no se produce ninguno, se pide

* 1. Ajustar un modelo de regresi´on log´ıstica para explicar la probabilidad de fallo en funci´on de la temperatura del despegue. Interpretar sus par´ametros en t´erminos de cocientes de ventajas y estudiar su significacio´n estad´ıstica mediante test de Wald e intervalos de confianza del 95%*.*
  2. Contrastar la bondad del ajuste del modelo ajustado mediante el test ma´s adecuado y validarlo mediante un ana´lisis de residuos y medidas de influencia.
  3. Construir, para el modelo ajustado, un test de diagn´ostico ´optimo para decidir si el lanzamiento de un cohete sera´ fallido en funci´on de la temperatura en el momento de lanzamiento. Calcular la tasa de verdaderos positivos y falsos positivos asociada.
  4. Obtener e interpretar la curva ROC.

1. Se desea estudiar si el nivel de a´cido fosf´orico (AF) en sangre en enfermos de c´ancer de pro´stata determina o no el desarrollo de nuevos no´dulos. Para ello se dispone de los datos de la siguiente tabla, donde la variable Y toma el valor 1 para aquellos individuos con nuevos no´dulos de ca´ncer 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 |

* 1. Ajustar un modelo de regresio´n log´ıstica simple para explicar la probabilidad de un nuevo brote de c´ancer en funci´on del nivel de a´cido fosfo´rico.
  2. Estudiar la bondad del ajuste del modelo ajustado y realizar una validaci´on completa de modelo mediante un ana´lisis de residuos y medidas de influencia.
  3. Interpretar los para´metros del modelo en t´erminos de cocientes de ventajas, construyendo para estos u´ltimos 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 a´cido fosfo´rico, proporcionando tambi´en un intervalo de confianza aproximado para el valor poblacional asociado.
  4. Estimar la tasa de verdaderos positivos y de falsos positivos del test de diagno´stico asociado al modelo ajustado, as´ı como la tasa de clasificaciones correctas con cutpoint 0.5. Razonar si este es el cutpoint m´as adecuado para discriminar entre individuos a los que se estima que desarrollara´n de nuevo la enfermedad y aquellos que no lo hara´n.
  5. Estimar el nivel de a´cido fosf´orico para el que la probabilidad estimada de desarrollar nuevos n´odulos de ca´ncer es el cutpoint o´ptimo (aquel que proporciona la ma´xima tasa de clasificaciones correctas). Obtener adem´as la curva ROC e interpretarla.
  6. ) Realizar un contraste condicional de razo´n de verosimilitudes para decidir si el modelo que contiene como variables explicativas la edad y el nivel de a´cido fosfo´rico es ma´s adecuado que el contiene so´lo el nivel de a´cido fosf´orico? ¿puede considerarse la edad un factor de confusi´on en la relacio´n entre la aparicio´n de nuevos no´dulos de ca´ncer y el nivel de a´cido fosf´orico?

1. Analizar mediante un modelo logit los datos de la siguiente tabla sobreocurrencia de vasoconstriccio´n (1 indica constricci´on y 0 lo contrario) en la piel de los dedos en funcio´n de la tasa y el volumen de aire respirado:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Constriccio´n | Volumen | Tasa | Constriccio´n | 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 |  |  |  |

1. Seleccionar paso a paso el modelo de regresi´on log´ıstica ma´s adecuado para explicar la probabilidad de padecer vasoconstriccio´n en funci´on de la tasa y el volumen de aire respirados. Estudiar la bondad del ajuste del modelo ajustado y realizar una validacio´n completa de modelo mediante un ana´lisis de residuos y medidas de influencia.
2. Interpretar los para´metros del modelo en t´erminos de cocientes de ventajas, construyendo para estos u´ltimos intervalos de confianza de nivel 99%*.* Estimar el cambio que se produce en la ventaja de padecer vasoconstriccio´n en funcio´n de la tasa y del volumen de aire respirados.
3. Estimar la tasa de verdaderos positivos y de falsos positivos del test de diagno´stico asociado al modelo ajustado, as´ı como la tasa de clasificaciones correctas con cutpoint 0.5. Razonar si este es el cutpoint m´as adecuado para discriminar entre individuos a los que se estima que desarrollara´n de nuevo la enfermedad y aquellos que no lo hara´n. Obtener la curva ROC e interpretarla.