MODELIZACIÓN ESTADÍSTICA. EVALUCIÓN CONTINUA

Primera parte de la asignatura

Teoría

Bloque 1

1. Formula el modelo de regresión lineal múltiple y enuncia los resultados que consideres más relevantes.
2. Explica en qué consiste un contraste de hipótesis y define los siguientes conceptos: Región crítica. Error de tipo I. Nivel de significación. p valor del contraste.
3. Define qué es un intervalo de confianza y el significado del coeficiente de confianza. Explica cómo se puede calcular un intervalo de confianza de los parámetros del modelo de regresión.
4. Explica lo que significa el Odds Ratio y sus ventajas e inconvenientes frente al Riesgo Relativo. Explica la relación que exsiste entre los odds ratio y los parámetros de un modelo de regresión logística.
5. Analiza la relación y las diferencias que existen entre los criterios de Akaike y BIC para la selección de variables de un modelo de regresión.

Bloque 2

1. Formula el modelo de regresión lineal múltiple y enuncia los resultados que consideres más relevantes.
2. En un modelo de regresión lineal múltiple con residuales normales, independientes y homocedásticos, demuestra que la varianza del estimador de β_j viene dada por la expresión siguiente y comenta su significado:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{n\, var(x_j)\,(1 - R^2_{j.resto})}$$

donde $var(x_j) = \dfrac{\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2}{n}$ y $R^2_{j.resto}$= coeficiente de determinación de la regresión de x_j con las otras variables del modelo, incluida la constante.

Sugerencia. Dado que $Cov(\hat{\beta}) = \sigma^2\,(X'\,X)^{-1}$, utiliza la inversa de matrices por bloques para calcular la varianza de cada $\hat{\beta}_j$.

`https://es.wikipedia.org/wiki/Matriz_por_bloques`

3. Sea X la matriz de variables independientes de un modelo de regresión. Comprueba que cuando todas esas variables son ortogonales entre si, el valor de las estimaciones de los parámetros coinciden con las de las regresiones simples de la variable dependiente con cada una de las variables predictoras.
4. Define la distancia de Cook e indica para que se utiliza.
5. Explica lo que significa el Odds Ratio y sus ventajas e inconvenientes frente al Riesgo Relativo. ¿En qué condiciones los estimadores de ambos indicadores toman valores similares?

Ejercicios

1. **Ejercicio 1**
   En un estudio sobre el consumo de los automóviles de gasolina, se recogieron las siguientes variables:

   - n.cilindros= número de cilindros
   - cv= potencia del cohe en caballos de vapor
   - relacion.ejes= relación entre los relacion ejes
   - peso= peso en toneladas métricas
   - aceleracion= tiempo que tarda en recorrer 500 m. partiendo de parado
   - ciudad= recorrido por ciudad o por carretera, en horas en las que no hay atascos o retenciones (0= carretera; 1=ciudad)
   - cilindrada= volumen que desplazan los pistones en litros
   - am= Tipo de cambio; automático= 0, manual=1.
   - distancia= distancia recorrido con 50 litos de gasolina

   A partir de los datos que aparecen en el fichero 'Coches.RData' resolver los siguientes apartados:

   a) Determina que variables están relacionadas con la distancia y describe como es esa relación.

   b) Construye un modelo que permita predecir la distancia en función de las variables recogidas. ¿Se puede considerar que $\beta_{peso} = -70$?

   c) ¿Si una de las variables analizadas no entra en el modelo, se puede afirmar que no tiene relación con la distancia recorrida?¿Por qué razón?

   d) Calcula un intervalo de confianza para la media de la distancia que recorre un coche con las siguientes características:
   $(n.cilindros = 6, cv = 102, relacion.ejes = 3{,}89, peso = 1{,}04, aceleracion = 18, ciudad = 1,$
   $cilindrada = 2{,}2, am = 1)$

2. **Ejercicio 2.**
   En una investigación cuyo objetivo es diagnosticar de forma precoz la diabetes en mujeres mayores de 21 años, se hizo un estudio en el que se midieron las siguientes variables:

   nembar : número de embarazos
   glucosa : concentración de glucosa en plasma(mg/dl)
   psdiast : presión sanguínea diastólica (mm Hg)
   triceps : espesor de la piel en el pliegue del triceps
   insulina : niveles de insulina en suero (micro U / ml)
   imc : índice de masa corporal (peso en kg / altura en metros^2)
   tendencia: tendencia familiar a tener diabetes
   edad : edad en años
   diabetes : diagnóstico de diabetes (negativo; positivo)

   Utiliza los datos del fichero D0.RData para resolver los siguientes apartados:

   a) Estudia la relación de la diabetes con el resto de las variables.

   b) Construye un modelo para predecir la diabetes, analiza su capacidad de predicción y explica su significado.

   c) Predice el comportamiento de dos mujeres con las siguientes características:

nembar	glucosa	psdiast	triceps	insulina	imc	tendencia	edad
3	137	84	27	92	27.3	0.231	59
0	102	64	31	78	30.6	0.496	16

NOTAS.

- La fecha límite de entrega es el lunes 9 de diciembre de 2024.
- El trabajo se puede hacer en grupos con un máximo de tres personas.
- Se puede elegir libremente uno de los bloques de teoría.
- El informe completo no debe sobrepasar las doce páginas, con tamaño de letra de 12 puntos, interlineado sencillo y tener formato pdf.
- En un fichero anexo deben estar los comandos de R empleados en los análisis.
- En la evaluación de los ejercicios se tendrán en cuenta tanto la parte técnica en la aplicación de los métodos estadísticos (70 %) como la claridad en la presentación de los resultados y de las conclusiones (30 %).
- El informe final se debe enviar al correo electrónico habitual: *norbert@uniovi.es*