Actividad 3: Modelización predictiva

Enunciado

25 de abril 2018

Contents

En esta actividad se usará el fichero del World Hapiness Report ya preparado, es decir, después del preproceso que se ha realizado en la primera actividad. Recordad que con la instrucción: `read.csv("2016_clean.csv")` podéis leer el fichero en R.

Después de preparar el fichero y realizar los análisis propios de la estadística descriptiva e inferencial, se pasará a estudiar la causalidad.

Esta base de datos contiene 157 registros y 13 variables. Las variables son Country, Region, HR, HS, LCI, UCI, GpC, Family, LE, Freedom, GC, Generosity, DR. Son las mismas variables de la actividad 1 y de la actividad 2.

---

1 Modelo de regresión lineal

Primeramente, estudiaremos como cambia el nivel de felicidad en función de algunas características de cada país.

1.1 Modelo de regresión lineal múltiple (regresores cuantitativos)

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la puntuación de felicidad (HS) de un país en función de tres factores cuantitativos: el indicador de renta por cápita (GpC), la esperanza de vida en salud (LE) y la corrupción (GC).

Evaluar la bondad de ajuste a través del coeficiente de determinación (R2). Podéis usar la instrucción de R **lm**.

1.2 Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la puntuación de felicidad (HS) de un país en función de cuatro factores. Además de los tres anteriores (renta, esperanza de vida y corrupción) ahora se añade la región del mundo (región). Usar como categoría de referencia la región "Western Europe" (para ello usar el factor combinado con relevel(region, ref = "Western Europe")).

Evaluar la bondad del ajuste a través del coeficiente de determinación (R^2) y comparar el resultado de este model con el obtenido en el apartado 1.1. Podéis usar la instrucción de R **lm** y usar el coeficiente R-cuadrado ajustado en la comparación. Interpretar también el significado de los coeficientes obtenidos y su significación estadística.

1.3 Efectuar una predicción del índice de felicidad en los dos modelos

Suponer un país de la región de Europa Occidental (Western Europe), con una renta de 1.5, una esperanza de vida en salud del 69% y un índice de corrupción de 0.35. Realizar la predicción con los dos modelos. Interpretar los resultados.

---

2 Modelo de regresión logística

Se desea evaluar el fenómeno de la felicidad desde un punto de vista de qué países son los 32 más felices. Por tanto, se evaluará la probabilidad de que un país del mundo esté entre este grupo de los 32 más felices. Para evaluar esta probabilidad se aplicará un modelo de regresión logística, donde la variable depediente será una variable binaria que indicará si el país tiene un lugar en el ránquing de los mejores 32 (donde la posición 32 está incluida). Se usará la muestra disponible para estimar el modelo con las mismas variables que en el modelo 1.1.

2.1 Estimación de un modelo de regresión logística

Estimar el modelo de regresión logística donde la variable dependiente es "best" y las explicativas son el indicador de renta por cápita (GpC) y la corrupción (GC). No incluimos la esperanza de vida puesto que pensamos que queda ya representada con la riqueza en la renta por cápita.

Evaluar si alguno de los regresores tiene influencia significativa (p-valor del contraste individual inferior al 5%).

2.2 Predicción en el modelo lineal generalizado (modelo de regresión logística)

Usando el modelo anterior, calculad la probabilidad de ser uno dels 32 países más felices del mundo para un país que tiene una renta de 1.5, y un índice de corrupción de 0.35.

2.3 Mejora del modelo

Buscar un modelo mejor al anterior añadiendo más variables explicativas. Se realizarán las siguientes pruebas:

Modelo regresor que añade al anterior la variable libertad (Freedom).

Modelo regresor que añade la región.

Modelo regresor que añade libertad y región.

Decidir si se prefiere el modelo inicial o bien uno de los modelos con freedom, con región, o con ambas. El criterio para decidir el mejor modelo es AIC. Cuanto más pequeño es AIC mejor es el modelo.

2.4 Calidad del ajuste

Calcular la matriz de confusión del mejor modelo del apartado 2.3 suponiendo un umbral de discriminación del 80%. Observad cuantos falsos negativos hay e interpretar qué es un falso negativo en este contexto.

2.5 La selección de los mejores países

Establecer un nivel de probabilidad (umbral de discriminación a partir del cual pensáis que el país tiene muchas posibilidades de estar entre los mejores, por ejemplo podéis escoger el 80%). Comparar el nivel de probabilidad que da el modelo con el ránquing del país e identificar los países que no se comportan según lo esperado. Podéis realizar este estudio gráficamente.

2.6 Curva ROC

Realizar el dibujo de la curva ROC (usando la librería pROC y la instrucción **roc** y el plot del objeto resultante). Calcular AUROC usando también este paquete, auc(...) donde debéis pasar el nombre del objeto roc. Interpretar el resultado.

2.7 Puntuaciones de los apartados

. Apartado 1 (30%)

. Apartados 2.1 hasta 2.3 (30%)

. Apartados 2.4 hasta 2.6 (30%)

. Calidad del informe dinámico y del código R (10%)