**XS-2130: MODELOS DE REGRESIÓN APLICADOS**

**II SEMESTRE 2015**

**TAREA No.3**

**SENIC: Study on the Efficacy of Nosocomial Infection Control.**

**El objetivo del estudio es determinar si los programas de control han reducido las tasas de infección nosocomial (adquirida en el hospital) en hospitales de E.U. Se tienen 113 hospitales seleccionados aleatoriamente de los 338 hospitales registrados.**

**Cada línea de datos tiene un número de identificación y provee información de 11 variables de cada hospital. Los datos son del período de estudio de 1975-76.**

1. **ID - Número de identificación.**
2. **LE - Largo de la estadía: largo promedio de la estadía de todos los pacientes en el hospital (en días).**
3. **EDAD: edad promedio de los pacientes (en años).**
4. **RI - Riesgo de infección: probabilidad estimada promedio de adquirir una infección en el hospital (en porcentaje).**
5. **RCR - Razón de cultivos rutinaria: razón del número de cultivos realizados al número de pacientes sin señales o síntomas de infección adquirida en el hospital (multiplicada por 100).**
6. **RRX - Razón de rayos-X de rutina: razón del número de rayo-X realizados al número de pacientes sin señales o síntomas de neumonía (multiplicada por 100).**
7. **CAMAS - Número de camas: número promedio de camas en el hospital durante el período de estudio.**
8. **AFIL - Afiliación al colegio de médicos: 1=sí, 2=no.**
9. **REGION: región geográfica, donde: 1=NE, 2=NC, 3=S, 4=W.**
10. **CENSO - Censo diario promedio: número promedio de pacientes en el hospital por día durante el período de estudio.**
11. **NURSE - Número de enfermeras: número promedio de tiempos completos (equivalente) de enfermeras registradas durante el período.**
12. **FSD - Facilidades y servicios disponibles: porcentaje de facilidades y servicios provistos por el hospital.**

**Use el archivo “senictot.Rdata”**

**Se va a usar el logaritmo natural del largo de estadía promedio de un paciente en un hospital como respuesta (Y) y los siguientes predictores numéricos: la edad, el riesgo de infección, el número de camas, el número de enfermeras, el número de pacientes y las facilidades y servicios disponibles. Además debido a las posibles diferencias entre regiones se va a incluir la región y las interacciones de este factor con todos los predictores numéricos.**

1. **Haga una presentación descriptiva donde se pueda observar cuáles variables podrían ser las más apropiadas para incluir en el modelo. Use siempre el logaritmo natural de la estadía como respuesta. Familiarícese con los datos y las variables para que tengan algún sentido para usted.**
2. **Escriba el modelo más grande que pueda tener con todas estas variables (incluyendo interacciones). A este modelo le llamaremos el modelo completo.**
3. **¿Cuánto es el máximo de la variabilidad que cualquier modelo con estas variables puede llegar a explicar en términos absolutos?**
4. **¿Cuánto es el máximo de la variabilidad que cualquier modelo con estas variables puede llegar a explicar en términos relativos?**
5. **Realice paso a paso el proceso de selección hacia adelante usando el criterio de información de Akaike (AIC).**
6. **¿Cuánta de la variabilidad total (en términos absolutos) logra explicar el modelo seleccionado en el punto anterior? ¿En términos relativos? Compare este resultado con el del modelo completo.**
7. **Ajuste el porcentaje de variabilidad explicada de ambos modelos (el escogido y el completo) y compare los resultados.**
8. **Compare la deviancia de ambos modelos. Compare el AIC de ambos modelos. Hay diferencias en los resultados al comparar deviancia y al comparar AIC?**
9. **Realice paso a paso el proceso de selección hacia atrás usando el criterio de información de Akaike. Escriba el modelo resultante y la ecuación de regresión asociada.**
10. **Escriba en cada paso cuánta variabilidad absoluta queda sin explicar por el modelo que va seleccionando. Observe qué sucede con esta cantidad al ir quitando variables.**
11. **Realice el proceso de selección hacia atrás usando el criterio de información de Bayes.**
12. **Utilice el criterio de R2 ajustado para escoger un conjunto de modelos sin interacciones. Permita un margen de 2% con el máximo R2 ajustado. Escoja uno de los modelos. Escriba el modelo recomendado y la ecuación de regresión asociada.**
13. **Haga el contraste del modelo completo contra cada uno de los modelos propuestos con los diferentes métodos usando la F (se le están pidiendo pruebas de hipótesis formales).**
14. **Calcule paso a paso el PRESS del modelo sugerido por AIC.**
15. **Calcule paso a paso el estadístico de Mallow del modelo sugerido por AIC.**
16. **Cuánto debe dar el estadístico de Mallow para el modelo completo.**
17. **Escriba una función que tenga como argumento un modelo lineal y el modelo completo y que devuelva lo siguiente:** 
    1. **La suma de cuadrados residual.**
    2. **El coeficiente de determinación.**
    3. **El coeficiente de determinación ajustado.**
    4. **El AIC.**
    5. **El BIC.**
    6. **El PRESS**
    7. **El estadístico de Mallow.**
    8. **El valor p (número de coeficientes estimados).**

**Aplique la función al modelo nulo, al modelo con todas las interacciones, al modelo propuesto con R2, con Akaike y con Bayes. Haga una tabla resumen con toda la información que se obtiene.**

* 1. **Utilice el estadístico de Mallow para evaluar el sesgo de los modelos propuestos.**
  2. **Compare los modelos según su capacidad predictiva.**