Modelos de Regresión Aplicados

Tarea 3.

**Ejercicio 1.** Se le da el archivo antropometria.Rdata en formato de R que contiene una muestra de personas a las cuales se les midió su Volumen Espiratorio Forzado VEF, el cual es la fuerza para soplar. Este indicador es utilizado por neumólogos (médicos especializados en los pulmones) para analizar la capacidad respiratoria. El objetivo principal del análisis es estudiar qué variables antropométricas logran predecir el VEF, controlando además por edad, sexo y condición de fumador. Se cuentan con varias variables. Las variables son las siguientes:

|  |  |
| --- | --- |
| Variable | Etiqueta |
| ---------------- | ---------------------------------------------------------------- |
| age | edad correcta al momento de entrevista |
| cinturar | Circunferencia de la cintura (en cm) |
| caderar | Circunferencia de la cadera (en cm) |
| grip | Fuerza de la mano (en Kg) |
| espirom | VEF (en Litros por minuto) |
| pantorrilla | Circunferencia de la pantorrilla (en cm) |
| brazo | Circunferencia del brazo (altura del bíceps) (en cm) |
| tricipital | Pliegue tricipital (en mm) |
| subscap | Pliegue subescapular (en mm) |
| pesokg | Peso (en Kg) |
| tallacm | Altura (en cm) |
| smoker | 1=Fuma, 0=No fuma |
| hombre | 1=Hombre, 0=Mujer |

Además, en la base de datos se encuentran las siguientes variables:

|  |  |
| --- | --- |
| Variable | Etiqueta |
| ---------------- | ---------------------------------------------------------------- |
| id0a | Fecha de la entrevista en aaaammdd |
| id0b | Hora de la entrevista (en segundos desde la medianoche |
| anyomes | Año y mes de la entrevista |
| secuencia | Orden en que fueron entrevistados los individuos |

Conteste con R las siguientes preguntas.

1. Seleccione un mes y año específico. Cree una subbase para Ud. o su grupo únicamente con el mes y año que le corresponde (puede ser con subset o con corchetes cuadrados). (2 ptos).
2. Haga una matriz de correlaciones y señale cuáles pares de variables violan la “regla de ojo del profesor” de que su correlación sea mayor que la correlación entre cada una de ellas y la variable dependiente (3 ptos.)
3. Estime un modelo de regresión en el que prediga el VEF (espirom) por las variables independientes señaladas más arriba (OBVIAMENTE, NO INCLUYA LAS VARIABLES id0a, id0b, anyomes y secuencia). Analice la multicolinealidad en el modelo observando los VIFs. Diga cuál o cuáles variables dan problema y por qué cree que eso pasa (Básese en la respuesta de la pregunta 2). (10 ptos.)
4. Con una prueba de Shapiro y una prueba de Jarque-Bera, analice el supuesto de normalidad de los residuos. Diga si estas dos pruebas coinciden o no. (5 ptos.)
5. Calcule “a pie” el estadístico de prueba de Breusch-Pagan, usando el procedimiento descrito en clases con los residuos ajustados, para analizar el supuesto de homoscedasticidad, y verifíquelo con la función existente en R (5 ptos.)
6. Grafique los siguientes gráficos resaltando los valores extremos o valores influenciales (o sea, tienen que “etiquetar” los valores extremos o residuales. Si no hay, entonces no se etiqueta nada):
   * Valores predichos vs. residuos estandarizados (5 ptos.)
   * Valores predichos vs. residuos estandarizados, con burbujas de tamaño proporcional a los leverage. (5 ptos.)
   * Residuos estandarizados vs. DFFITs. (5 ptos.)
7. Con un procedimiento de selección de variables tipo stepwise, seleccione el mejor modelo usando el BIC en lugar del AIC (investigue cómo se selecciona con BIC en lugar de AIC) (10 ptos.)
8. Con un procedimiento de selección de variables basado en p-values y tipo stepwise, seleccione el mejor modelo (10 ptos.)
9. Al modelo resultante, cálculele los VIFs y la prueba de Breusch-Pagan usando las funciones ya existentes en R. Se mejora de alguna forma el modelo en términos del cumplimiento del modelo cuando se escoge un modelo más parsimonioso? (5 ptos.)

**Ejercicio 3.**  Para analizar la bondad de ajuste a una distribución normal, existen diversas técnicas de pruebas de hipótesis. El siguiente ejercicio busca comparar las distintas conclusiones que se pueden hacer a partir de ellos. Proceda con los siguientes pasos:

1. Cree una variable x con 50 números aleatorios una distribución uniforme con parámetros mínimo=1 y máximo=10. Además, cree dos variables que les llamará error1 y error2 de la siguiente forma:
   1. error1=rt (50, df=3) (O sea, números aleatorios de una distribución t de Student con v=3)
   2. error2=rnorm(50) (O sea, números aleatorios de una distribución normal estándar) (2 ptos.).
2. Investigue cuál es la diferencia entre números aleatorios de una distribución t y de una distribución normal estándar, en términos de la forma de la distribución (2 ptos.)
3. Las variables anteriores equivalen a los residuos de los modelos de regresión. Pruebe la hipótesis nula de que error1 y error2 se distribuyen normalmente con media=0 y desviación estándar=1, con:
   1. Prueba de Shapiro
   2. Prueba de Kolmogorov-Smirnov
   3. Prueba de Jarque Bera
   4. No escriba si se rechaza H0 o no en cada una, sino que en 5 líneas describa las diferencias generales entre las tres pruebas para las dos variables (4 ptos).
4. Ahora estime dos variables y1 y y2 que sean iguales a:
   1. y1=10+3\*x+error1
   2. y2=10+3\*x+error2 (2ptos.)
5. Estime dos modelos de regresión, y1 en función de x, y y2 en función de x. Para cada uno de ellos, calcule el intervalo de confianza para la pendiente (el coeficiente beta1) (5 ptos.)
6. (Repasar de Métodos). Con bootstrap y 1000 réplicas, calcule intervalos de confianza no paramétricos (los percentiles) para beta1 del modelo 1 y para beta1 del modelo 2 (5 ptos.)
7. (Repasar de Métodos). Suponga que los intervalos de bootstrap son los correctos. Compare los intervalos de confianza de cada modelo con los intervalos de confianza bootstrapeados y diga si se parecen o no. Además, con los resultados, argumente si la falta de normalidad afecta los intervalos de confianza de beta (10 ptos)