Primer trabajo práctico: Modelo lineal

Para la entrega debe crearse un informe explicando cada ítem, con los gráficos y justificaciones que crea pertinentes. La entrega debe realizarse en un archivo formato pdf, a jmgarcia@fi.uba.ar. Fecha límite de entrega: domingo 29 de mayo.

El dataset *diabetes2.csv* es original del National Institute of Diabetes and Digestive and Kidney Diseases. Se creó con el objetivo de predecir si un paciente tiene o no diabetes, basado en diferentes medidas contenidas en el dataset. En particular, todos los pacientes son de sexo femenino mayores de 21 años de herencia Pima. Para este primer estudio, lo que buscaremos es encontrar relaciones entre otras de las variables contenidas en el dataset.

1. Cargar los datos de *diabetes2.csv*. La variable Outcome indica si la persona es diabética (1) o no (0). Transformarla en un factor. Finalmente revisar que todas las variables contenidas en el dataframe estén correctamente definidas.

2. Se desea ajustar un modelo de regresión múltiple para predecir la variable BMI en función del resto de las variables en el data set. Escribir el modelo propuesto, indicando los supuestos del mismo.

3. Realizar un scatterplot de las variables con la función $ggpairs$.

4. A partir de la tabla de correlaciones estimadas entre las variables, si tuviera que elegir una sola variable para proponer un modelo de regresión simple, ¿cuál eligiría y porqué?

5. Realizar un ajuste de regresión lineal múltiple. A partir de la tabla de coeficientes estimados, ¿Qué variables resultan significativas? ¿A qué nivel? ¿Cuál es el valor de la estimación para σ^2? Especificar las hipótesis nulas y alternativas para *alguno* de los test t reportados en la tabla, el estadístico del test y la regla de decisión. ¿Cómo se calcula el p-valor para este test?

6. Evaluar la bondad del ajuste realizado, a través del coeficiente de determinación. Indicar cuánto vale y qué significa.

7. ¿Es la regresión significativa? Especificar las hipótesis nula y alternativa de este test. ¿Cómo se calcula el p-valor en este caso? ¿Rechazaría a un nivel de significación de 0.05?.

8. ¿Cuál sería la estimación de la esperanza del BMI de una mujer tuvo 2 embarazos y tiene una concentración de glucosa de 100, una presión sistólica de 70, un valor de piel de triceps de 20, no tiene diabetes, un valor de la función pedigree de 0.24 y 30 años de edad?

9. Hallar un intervalo de confianza y de predicción de nivel 0.95 para la estimación hallada en el ítem anterior.

10. *Selección de modelos.* Plantear un nuevo modelo en el que intervengan aquellas variables que contribuyen significativamente y estimar los parámetros por mínimos cuadrados. ¿Qué modelo elegiría finalmente? Utilizar medidas de bondad de ajuste y de predicción, tal como la estimación del error cuadrático medio.

11. Validar los supuestos expresados en el ítem 2 a partir del análisis de los residuos, para el modelo seleccionado.

12. Mediante bootstrap no paramétrico, estimar la densidad del estimador para el coeficiente de regresión de la variable elegida en el punto 4. ¿Qué concluye?