Ejercicios para la clase 5

1. Cargar los datos de niños de bajo peso (`low birth weight infants.txt`) y repetir el ajuste de mínimos cuadrados para explicar el perímetro cefálico con la edad gestacional.

   a) Compruebe, ajustando un modelo lineal simple, que momage no resulta significativa para explicar a headcirc.

   b) Proponga un modelo lineal con dos variables explicativas: toxemia y gestage para explicar a headcirc.

      1) Ajústelo con los datos. Evalúe la significatividad de los coeficientes. ¿Puede interpretar el modelo propuesto? Para hacerlo, escriba ecuaciones para la esperanza condicional para este caso, como las dadas en clase, condicionando a
         - toxemia y gestage,
         - toxemia $= 0$ y gestage,
         - toxemia $= 1$ y gestage.
      2) Grafique en 2D headcirc versus gestage, con dos colores distintos (azul y rojo) según toxemia. Superpóngale al gráfico los valores predichos (usando `fitted.values`), en colores celeste y rosa de acuerdo al valor de toxemia. Interprete.

2. El conjunto de datos `bdims` del paquete `openintro` contiene medidas de circunferencia del cuerpo y diámetro esquelético para 507 individuos físicamente activos.

   a) Utilizando el conjunto de datos `bdims`, realizar un diagrama de dispersión que muestre cómo el peso de una persona (`wgt`) varía en función de su altura (`hgt`). Identifique el género de las observaciones en el scatterplot, para ello pinte de rojo a las mujeres y de azul a los hombres, use la instrucción `col` de `R`. Observar que en esta base de datos, `sex = 1` para los hombres y `sex = 0` para las mujeres.

      (i) Realizar un diagrama de dispersión que muestre la relación entre el peso medido en kilogramos (`wgt`) y la circunferencia de la cadera medida en centímetros (`hip.gi`), ponga el peso en el eje vertical. Describa la relación entre la circunferencia de la cadera y el peso.

      (ii) ¿Cómo cambiaría la relación si el peso se midiera en libras mientras que las unidades para la circunferencia de la cadera permanecieran en centímetros?

      (iii) Ajuste un modelo lineal para explicar el peso por la circunferencia de cadera, con las variables en las unidades originales. Escriba el modelo (con papel y lápiz, con betas y epsilones). Luego, escriba el modelo ajustado (sin epsilones). Interprete la pendiente estimada en términos del problema. Su respuesta debería contener una frase que comience así: "Si comparamos el peso esperado para las personas cuyo contorno de cadera es de x cm con el peso esperado para las personas con $x + 1$ centímetros de contorno de cadera, según el modelo ajustado...".

      (iv) Superponga la recta ajustada al scatterplot. Observe el gráfico. ¿Diría que la recta describe bien la relación entre ambas variables?

      (v) Elegimos una persona adulta físicamente activa entre los estudiantes de primer año de la facultad. Su contorno de cadera mide 100 cm. Prediga su peso en kilogramos.

      (vi) Esa persona elegida al azar pesa 81kg. Calcule el residuo.

      (vii) Estime el peso esperado para la población de adultos cuyo contorno de cadera mide 100 cm.

   b) (i) Realizar un diagrama de dispersión que muestre la relación entre el peso medido en kilogramos (`wgt`) y la altura (`hgt`).

      (ii) Ajuste un modelo lineal para explicar el peso por la altura. Escriba el modelo (con papel y lápiz, con betas y epsilones). Luego, escriba el modelo ajustado (sin epsilones). Interprete la pendiente estimada en términos del problema. Interprete la pendiente. ¿Es razonable el signo obtenido para la pendiente estimada? Superponer al scatterplot anterior la recta estimada.

      (iii) La persona elegida en el ejercicio anterior, medía 187 cm. de alto, y pesaba 81 kg. Prediga su peso con el modelo que tiene a la altura como covariable. Calcule el residuo de dicha observación.

   c) Intervalos de confianza y predicción

(i) Compare los ajustes realizados en los ejercicios $2a$ y $2b$. En ambos se ajusta un modelo lineal para explicar el peso medido en kilogramos (`wgt`): en el ejercicio $2a$ por la circunferencia de la cadera medida en centímetros (`hip.gi`), en el ejercicio $2b$ por la altura media en centímetros (`hgt`). ¿Cuál de los dos covariables explica mejor al peso? ¿Qué herramienta utiliza para compararlos?

(ii) Para el ajuste del peso usando la circunferencia de cadera como única covariable, halle un intervalo de confianza de nivel 0.95 cuando el contorno de cadera mide 100 cm. Compárelo con el intervalo de predicción para ese mismo contorno de cadera.

(iii)) Para el ajuste del peso usando la altura como única covariable, halle un intervalo de confianza de nivel 0.95 cuando la altura es de 176 cm. Compárelo con el intervalo de predicción para esa misma altura. ¿Cuál de los dos modelos da un intervalo de predicción más útil?

(iv) Construya un intervalo de confianza para el peso esperado cuando el contorno de cintura es de 80cm., 95cm., 125cm. de nivel 0.95. Estos tres intervalos, ¿tienen nivel simultáneo 0.95? Es decir, la siguiente afirmación ¿es verdadera o falsa? Justifique. En aproximadamente 95 de cada 100 veces que yo construya los IC basados en una (misma) muestra, cada uno de los 3 IC contendrán al verdadero valor esperado del peso.

(v) Construya los intervalos de predicción para el peso esperado cuando de nivel (individual) 0.95 cuando el contorno de cintura es de 80cm., 95cm. y 125cm. Compare las longitudes de estos tres intevalos entre sí. Compárelos con los IC de nivel individual.

(vi) Construya los intervalos de confianza para el peso esperado cuando de nivel simultáneo 0.95 cuando el contorno de cintura es de 80cm., 95cm. y 125cm.

(vii) Estime la varianza del error (σ^2) en ambos modelos.

(viii) Realice un scatterplot del peso en función del contorno de cintura. Superponga los IC y los IP al gráfico, de nivel 0.95 (no simultáneo).

d) Modelo lineal múltiple.

(i) En el ejercicio $2a$ explicamos el peso de las personas registradas en esta base de datos, por el contorno de la cadera y en el ejercicio $2b$ la explicamos con un modelo con la altura como covariable. Proponga un modelo de regresión múltiple que explique el peso medido en kilogramos (`wgt`) utilizando el contorno de la cadera medida en centímetros (`hip.gi`) y la altura media en centímetros (`hgt`) como covariables. Escriba el modelo que está ajustando. Realice el ajuste con el R.

(ii) Interprete los coeficientes estimados. ¿Resultan significativos? Cambian sus valores respecto de los que tenían los coeficientes que acompañaban a estas variables en los modelos de regresión lineal simple?

(iii) Evalúe la bondad del ajuste realizado, a través del R^2. Indique cuánto vale y qué significa. Se quiere comparar este ajuste con el que dan los dos modelos lineales simples propuestos en los ejercicios $2a$ y $2b$. ¿Es correcto comparar los R^2 de los tres ajustes? ¿Qué valores puedo comparar? ¿Es mejor este ajuste múltiple?

(iv) Estime la varianza de los errores. Compare este estimador con los obtenidos en los dos ajustes simples.

(v) Estime el peso esperado para la población de adultos cuyo contorno de cadera mide 100 cm y su altura es de 174cm. Dé un intervalo de confianza de nivel 0.95 para este valor esperado.

(vi) Prediga el peso de un adulto cuyo contorno de cadera mide 100 cm y su altura es de 174cm. Dé un intervalo de predicción de nivel 0.95 para este valor. Compare las longitudes de los tres intervalos de predicción que se obtienen usando el modelo que solamente tiene al contorno de cadera como explicativa, al que solamente usa la altura y al modelo múltiple que contiene a ambas.