1. Ejercicios de Correlación

Con `R` hacer scatterplots es muy sencillo. Además es tan útil lo que puede aprenderse de los datos que vale la pena entrenarse exponiéndose a muchos ejemplos. Con el tiempo se gana familiaridad con los tipos de patrones que se ven. De a poco uno aprende a reconocer cómo los diagramas de dispersión pueden revelar la naturaleza de la relación entre dos variables.

En esta ejercitación trabajaremos con algunos conjuntos de datos que están disponibles a través del paquete `openintro` de `R`. Brevemente:

`mammals`: El conjunto de datos de mamíferos contiene información sobre 62 especies diferentes de mamíferos, incluyendo su peso corporal, el peso del cerebro, el tiempo de gestación y algunas otras variables.

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

`smoking`: El conjunto de datos `smoking` contiene información sobre los hábitos de fumar de 1.691 ciudadanos del Reino Unido.

`cars`: El conjunto de datos `cars` está compuesto por la información de 54 autos modelo 1993. Se relevan 6 variables de cada uno (tamaño, precio en dólares, rendimiento en ciudad (millas por galón), tipo de tracción, cantidad de pasajeros, peso).

Para ver una documentación más completa, utilice las funciones `?` ó `help()`, una vez cargado el paquete. Por ejemplo, `help(mammals)`.

Ejercicio 1.1 *Mamíferos, Parte I. Usando el conjunto de datos de `mammals`, crear un diagrama de dispersión que muestre cómo el peso del cerebro de un mamífero (`BrainWt`) varía en función de su peso corporal (`BodyWt`).*

Ejercicio 1.2 *Medidas del cuerpo, Parte I. 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.*

Ejercicio 1.3 *Utilizando el conjunto de datos `smoking`, realizar un diagrama de dispersión que ilustre cómo varía la cantidad de cigarrillos que fuma por día una persona durante el fin de semana (`amtWeekends`), en función de su edad (`age`).*

Ejercicio 1.4 *Utilizando el conjunto de datos `cars`, realizar un scatter plot del rendimiento del auto en la ciudad (`mpgCity`) en función del peso del auto (`weight`).*

Ejercicio 1.5 *Para cada uno de los cuatro scatterplots anteriores describa la forma, la dirección y la fuerza de la relación entre las dos variables involucradas. Respuestas posibles:*

- *forma: lineal, no lineal (cuadrática, exponencial, etc.)*
- *dirección: positiva, negativa*
- *fuerza de la relación: fuerte, moderada, débil, no asociación. Tiene que ver con cuán dispersos están las observaciones respecto del patrón descripto en la forma.*

Ejercicio 1.6 *¿Para cuáles de los 4 conjuntos de datos tiene sentido resumir la relación entre ambas variables con el coeficiente de correlación muestral de Pearson? Para los casos en los cuales contestó que era apropiado,*

(a) calcúlelo usando R.

(b) Testee las siguientes hipótesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

para cada uno de esos conjuntos. Antes de hacerlo defina a ρ en palabras. Observe que en el ítem 1.6 (a) calculó un estimador de esta cantidad, para cada conjunto. ¿En qué casos rechaza la hipótesis nula, a nivel 0.05?

Ejercicio 1.7 *Mamíferos, Parte II. El conjunto de datos de* `mammals` *presenta un scatterplot que no es razonable resumir con el coeficiente de correlación muestral. El gráfico no es lindo por varios motivos, básicamente las observaciones parecen estar en escalas distintas, hay muchas observaciones superpuestas, necesitaríamos hacer un zoom del gráfico en la zona cercana al origen, a expensas de perder las dos observaciones con valores mucho más grandes que el resto. Podemos comparar lo que pasaría si no hubiéramos observado el diagrama de dispersión y quisiéramos resumir los datos con el coeficiente de correlación.*

(a) Calcule el coeficiente de correlación muestral de Pearson para los 62 mamíferos.

(b) Identifique las dos observaciones que tienen valores de peso corporal y cerebral más grandes que el resto. Realice un scatter plot de las restantes 60 variables. ¿Cómo podría describir este gráfico? Calcule el coeficiente de correlación muestral de Pearson para estas 60 observaciones.

(c) El gráfico hecho en el ítem anterior no corrige el problema original del todo. La forma general podría describirse como un abanico: claramente las variables están asociadas, la asociación es positiva (ambas crecen simultáneamente) pero la dispersión de los datos parece aumentar a medida que ambas variables aumentan. Esta forma es frecuente en los conjuntos de datos, suelen corresponder a observaciones que están medidas en escalas que no son comparables entre sí y suele corregirse al tomar logaritmo en ambas variables. Para ver el efecto de transformar las variables, realice un scatterplot con todas las observaciones, del logaritmo (en base 10, o en base e) del peso del cerebro en función del logaritmo del peso corporal. Observe el gráfico. ¿Cómo lo describiría? Calcule la correlación de Pearson para los datos transformados.

(d) Para ambos conjuntos de datos (transformados por el logaritmo y sin transformar) calcule la correlación de Spearman.

Ejercicio 1.8 *¿Con qué coeficiente de correlación, Pearson o Spearman, resumiría los datos de* `cars`*? (*`weight, mpgCity`*)*

2. Ejercicios Regresión Lineal Simple (primera parte)

Ejercicio 2.1 *Medidas del cuerpo, Parte II. Datos publicados en Heinz, Peterson, Johnson, y Kerk [2003], base de datos* `bdims` *del paquete* `openintro`*.*

(a) 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.*

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

(c) *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 una persona aumenta un cm. de contorno de cadera, en promedio su peso aumentará ... kilogramos".*

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

(e) *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.*

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

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

Ejercicio 2.2 *Medidas del cuerpo, Parte III. Base de datos* `bdims` *del paquete* `openintro`.

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

(b) *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.*

(c) *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.*

Ejercicio 2.3 *Mamíferos, Parte III. Base de datos* `mammals` *del paquete* `openintro`.

(a) *Queremos ajustar un modelo lineal para predecir el peso del cerebro de un mamífero (*`BrainWt`*) a partir del peso corporal (*`BodyWt`*) del animal. Habíamos visto en el Ejercicio 1.7 que si graficamos el peso del cerebro en función del peso corporal, el gráfico era bastante feo. Y que todo mejoraba tomando logaritmo (en cualquier base, digamos base 10) de ambas variables. Ajuste un modelo lineal para explicar a* log_{10} (`BrainWt`) *en función del* log_{10} (`BodyWt`) *. Como antes, escriba el modelo teórico y el ajustado. Una observación: en el help del* `openitro` *se indica que la variable* `BrainWt` *está medida en kg., sin embargo, esta variable está medida en gramos.*

(b) *Repita el scatterplot de las variables transformadas y superpóngale la recta ajustada.*

(c) *La observación 45 corresponde a un chancho. Prediga el peso del cerebro del chancho con el modelo ajustado, sabiendo que pesa 192 kilos. Recuerde transformar al peso corporal del chancho antes de hacer cálculos. Marque esa observación en el gráfico, con color violeta.*

(d) *La observación 34 corresponde a un ser humano. Prediga el peso del cerebro de un ser humano con el modelo ajustado, sabiendo que pesa 62 kilos. Recuerde transformar al peso corporal del chancho antes de hacer cálculos. Marque esa observación en el gráfico, con color rojo.*

3. Ejercicios de Regresión Lineal Simple (segunda parte)

Ejercicio 3.1 *Medidas del cuerpo, Parte IV. Base de datos* `bdims` *del paquete* `openintro`.

(a) Compare los ajustes realizados en los ejercicios 2.1 y 2.2. En ambos se ajusta un modelo lineal para explicar el peso medido en kilogramos (`wgt`*): en el ejercicio 2.1 por la circunferencia de la cadera medida en centímetros (*`hip.gi`*), en el ejercicio 2.2 por la altura media en centímetros (*`hgt`*). ¿Cuál de los dos covariables explica mejor al peso? ¿Qué herramienta utiliza para compararlos?*

(b) 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.

(c) 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?

(d) 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.

(e) 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.

(f) 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.

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

(h) 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).

Ejercicio 3.2 *(Del Libro de Weisberg [2005]) Uno de los primeros usos de la regresión fue estudiar el traspaso de ciertos rasgos de generación en generación. Durante el período 1893–1898, E. S. Pearson organizó la recolección de las alturas de $n = 1375$ madres en el Reino Unido menores de 65 años y una de sus hijas adultas mayores de 18 años. Pearson y Lee (1903) publicaron los datos, y usaremos estos datos para examinar la herencia. Los datos (medidos en pulgadas) pueden verse en el archivo de datos* `heights.txt` *del paquete* `alr3` *de* `R`*. Nos interesa estudiar el traspaso de madre a hija, así que miramos la altura de la madre, llamada* `Mheight`*, como la variable predictora y la altura de la hija,* `Dheight`*, como variable de respuesta. ¿Será que las madres más altas tienden a tener hijas más altas? ¿Las madres más bajas tienden a tener hijas más bajas?*

(a) Realice un scatterplot de los datos, con la altura de las madres en el eje horizontal.

i. Como lo que queremos es comparar las alturas de las madres con la de las hijas, necesitamos que en el scatterplot las escalas de ambos ejes sean las mismas (y que por lo tanto el gráfico sea cuadrado).

ii. Si cada madre e hija tuvieran exactamente la misma altura que su hija, ¿cómo luciría este scatterplot? Resuma lo que observa en este gráfico. Superpóngale la figura que describió como respuesta a la pregunta anterior. ¿Describe esta figura un buen resumen de la relación entre ambas variables?

iii. Los datos originales fueron redondeados a la pulgada más cercana. Si trabajamos directamente con ellos, veremos menos puntos en el scatterplot, ya que varios quedarán superpuestos. Una forma de lidiar con este problema es usar el jittering, es decir, sumar un pequeño número uniforme aleatorio se a cada valor. Los datos de la librería alr3 tienen un número aleatorio uniforme en el rango de -0.5 a +0.5 añadidos. Observemos que si se redondearan los valores del archivo `heights` *se recuperarían los datos originalmente publicados. En base al scatterplot, ¿parecería ser cierto que las madres más altas suelen tener hijas más altas y viceversa con las más bajas?*

(b) Ajuste el modelo lineal a los datos. Indique el valor de la recta ajustada. Superpóngala al scatter plot. ¿Presenta visualmente un mejor ajuste que la recta identidad postulada en el ítem anterior? Dé los estimadores de los coeficientes de la recta, sus errores estándares, el coeficiente de determinación, estime la varianza de los errores. Halle un intervalo de confianza de nivel 0.95 para la pediente. Testee la hipótesis $E(\texttt{Dheight} \mid \texttt{Mheight}) = \beta_0$ *versus la alternativa que* $E(\texttt{Dheight} \mid \texttt{Mheight}) = \beta_0 + \beta_1 \texttt{Mheight}$. *Escriba su conclusión al respecto en un par de renglones.*

(c) Prediga y obtenga un intervalo de predicción para la altura de una hija cuya madre mide 64 pulgadas. Observe que para que esta predicción sea razonable, hay que pensar que la madre vivía en Inglaterra a fines del siglo XIX.

(d) Una pulgada equivale a 2.54cm. Convierta ambas variables a centímetros (`Dheightcm` *y* `Mheightcm`*) y ajuste un modelo lineal a estas nuevas variables. ¿Deberían cambiar los estimadores de* β_0 *y* β_1*? ¿De qué manera? ¿Y los errores estándares? ¿Y los p-valores? ¿Y el coeficiente de determinación? ¿Y la estimación del desvío estándar de los errores? Compare ambos resultados, y verifique si sus conjeturas resultaron ciertas. En estadística, que un estimador se adapte al cambio de escala en las variables (covariable y respuesta) se dice: "el estimador es equivariante (afín y por escala)".*

Ejercicio 3.3 *Simulación 1. El objetivo de este ejercicio es generar datos para los cuales conocemos (y controlamos) el modelo del que provienen y la distribución que siguen.*

(a) Generar $n = 22$ *datos que sigan el modelo lineal simple*

$$Y = 10 + 5X + \varepsilon, \tag{1}$$

donde $\varepsilon \sim N\left(0, \sigma^2\right)$, *con* $\sigma^2 = 49$. *Las n observaciones las generamos independientes entre sí.*

i. Para hacer esto en R, conviene primero definir un vector de longitud 22 de errores, que tenga distribución normal. La instrucción que lo hace es `rnorm`. *Visualice los errores con un histograma de los mismos.*

ii. Inventamos los valores de X. Para eso, generamos 22 valores con distribución uniforme entre 0 y 10, con la instrucción `runif`. *Para no trabajar con tantos decimales, redondeamos estos valores a dos decimales, con la instrucción* `round()`.

iii. Ahora sí, definimos las Y usando todo lo anterior:

$$Y_i = 10 + 5X_i + \varepsilon_i,$$

para cada $1 \leq i \leq n = 22$. *Observar que nos hemos conseguido observaciones* $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ *independientes que siguen el modelo*

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

¿Cuánto valen los verdaderos β_0 *y* β_1*?*

(b) Haga un scatterplot de los datos generados.

(c) *Ajuste el modelo lineal, guarde el resultado obtenido en el objeto* `ajuste`*. Observe si los parámetros estimados son significativos. Calcule intervalos de confianza para la ordenada al origen y la pendiente, de nivel 0.95. Para esto recuerde los comandos:* `lm` *y* `confint`*. ¿Los verdaderos* β_0 *y* β_1 *pertenecen a dichos intervalos? ¿Cuánto dio la pendiente estimada,* $\widehat{\beta}_1$*? ¿En qué parte de la salida del ajuste lineal podemos encontrar el estimador de* σ*? ¿Cuánto debería valer?*

(d) *Pídamosle al R que chequee si el 5 pertenece al IC de nivel 0.95 calculado en base a la muestra. El* `R` *nos devolverá "*`TRUE`*" o "*`FALSE`*" como respuesta a esta pregunta. La computadora codifica los "*`TRUE`*" como 1 y los "*`FALSE`*" como 0 para poder operar numéricamente con respuestas de este tipo. También guardemos la pendiente estimada en un objeto que se llame* `beta1est`*.*

(e) *Superpóngale al scatterplot de los datos la recta verdadera (en azul) y la estimada en base a ellos (en rojo).*

Ejercicio 3.4 *Simulación 2. Ahora hacemos un upgrade del desafío. Vamos a repetir lo hecho en el ejercicio 3.3 muchas veces, digamos lo replicaremos* $B = 1000$ *veces. Llamaremos replicación a cada repetición del ejercicio anterior. ¿Qué replicamos? Repetimos generar* $n = 22$ *observaciones del modelo (1) con errores normales (lo que llamamos elegir una muestra), ajustamos el modelo lineal, guardamos la pendiente estimada y nos fijamos si el 5 pertenece al intervalo de confianza para la pendiente.*

(a) *¿Puede usted anticipar, desde la teoría las respuestas de las preguntas que siguen?*

   i. *Las pendientes estimadas en las* $B = 1000$ *replicaciones, ¿serán siempre iguales o cambiarán de replicación en replicación?*
   ii. *¿Alrededor de qué número variarán las pendientes estimadas en las 1000 replicaciones?*
   iii. *Si hacemos un histograma de estas* $B = 1000$ *replicaciones, ¿a qué distribución debería parecerse?*
   iv. *Aproximadamente, ¿qué porcentaje de los 1000 intervalos de confianza para la pendiente estimados a partir de las 1000 muestras cubrirá al verdadero valor de la pendiente?*
   v. *Observe que si usted tuviera 22.000 observaciones de un modelo, nunca las dividiría en 1000 tandas de 22 observaciones para analizarlas: las consideraría todas juntas. Es por eso que este ejercicio es irreal, es simplemente una herramienta de aprendizaje.*

(b) *Antes de empezar, definamos vectores donde guardaremos la información. De longitud* $B = 1000$ *cada uno, necesitamos un vector para los* $\widehat{\beta}_1$ *y otro para guardar las respuestas respecto de si el 5 pertenece o no al intervalo de confianza. Llamémoslos:* `beta1est` *e* `icbeta`*. Inicialmente ponemos un* `NA` *en cada coordenada de estos vectores (*`NA` *es, usualmente, la notación reservada para una observación faltante, son las siglas de not available). La instrucción* `rep` *del R (que repite un número o una acción un número fijo de veces resultará muy útil).*

(c) *Los valores de* $X_1, \ldots, X_{22}$ *los dejaremos siempre fijos, en los valores que tomamos en el ejercicio 3.3. En cada replicación elegimos nuevos valores para los errores, y consecuentemente, nuevos valores para la variable respuesta* $Y_1, \ldots, Y_{22}$*. No nos interesará guardar ni a los errores ni a las* Y*. Para cada muestra, corra el ajuste lineal y guarde la pendiente estimada y la respuesta en forma de true o false respecto de si el intervalo de confianza para la pendiente contiene al verdadero valor de la pendiente. Todo esto puede realizarse con la instrucción* `for` *del* `R`*, que no es la manera óptima de programar, pero sí es la más comprensible.*

(d) *Haga un histograma de las pendientes estimadas. ¿Qué distribución parecen tener los datos?*

(e) *¿Qué proporción de los intervalos de confianza construidos contiene al verdadero valor de la pendiente?*

Ejercicio 3.5 *Mamíferos, Parte IV. conjunto de datos* `mammals` *del paquete* `openintro`*. Vimos, en los ejercicios 1.7 y 2.3, que el scatter plot de los datos originales no tiene la forma elipsoidal (o de pelota de rugby, más o menos achatada) que podemos describir con un modelo de regresión lineal. Por ello, ajustamos un modelo lineal para explicar a* $\log_{10}(\texttt{BrainWt})$ *en función del* $\log_{10}(\texttt{BodyWt})$,

$$\log_{10}(\texttt{BrainWt}) = \beta_0 + \beta_1 \log_{10}(\texttt{BodyWt}) + \varepsilon. \tag{2}$$

Una observación: en el help del `openitro` *se indica que la variable* `BrainWt` *está medida en kg., sin embargo, esta variable está medida en gramos.*

(a) A partir de $\log_{10}(10) = 1$ *y de recordar que*

$$\log_{10}(ab) = \log_{10}(a) + \log_{10}(b),$$

podemos observar que en el modelo lineal (2) aumentar una unidad de $log_{10}(\texttt{BodyWt})$ *es lo mismo que multiplicar a* `BodyWt` *por 10. Si dos animales difieren en el* `BodyWt` *por un factor de diez, dé un intervalo del 95 % de confianza para la diferencia en el* $\log_{10}(\texttt{BrainWt})$ *para estos dos animales.*

(b) Para un mamífero que no está en la base de datos, cuyo peso corporal es de 100 kg., obtenga la predicción y un intervalo de nivel 95 % de predicción del $\log_{10}(\texttt{BrainWt})$. *Prediga el peso del cerebro de dicho animal. Ahora queremos convertir el intervalo de predicción del* $\log_{10}(\texttt{BrainWt})$ *en un intervalo de predicción para el* `BrainWt`*. Para eso, observemos que si el intervalo* (a, b) *es un intevalo de predicción de nivel 95 % para* $\log_{10}(\texttt{BrainWt})$, *entonces, un intervalo para el* `BrainWt` *está dado por* $\left(10^a, 10^b\right)$. *¿Por qué? Use este resultado para obtener un intervalo de predicción del peso del cerebro del mamífero cuyo peso corporal es 100kg. Mirando los valores numéricos obtenidos, ¿parece muy útil el resultado obtenido?*

(c) Observe que si quisiéramos construir el intervalo de confianza de nivel 95 % para el peso del cerebro esperado de un mamífero cuyo peso corporal es es 100kg, no es posible hacer la conversión del ítem anterior de manera automática, ya que para cualquier funcion g en general

$$E[g(Y)] \neq g(E[Y]).$$

Si se quiere construir dicho intervalo, habrá que apelar a otras herramientas, por ejemplo el desarrollo de Taylor de la función g.

Ejercicio 3.6 *(Del Libro de Weisberg [2005]) La perca americana o lubina (smallmouth bass) es un pez que vive en lagos y cuya pesca constituye una actividad bastante difundida. En Estados Unidos, para garantizar un equilibrio saludable entre la conservación del medio ambiente y la explotación humana se implementan distintas políticas de regulación de su pesca. Entender los patrones de crecimiento de los peces es de gran ayuda para decidir políticas de conservación de stock de peces y de permisos de pesca. Para ello, la base de datos* `wblake` *del paquete* `alr3` *registra la longitud en milímetros al momento de la captura (*`Length`*) y la edad (*`Age`*) para* $n = 439$ *percas medidas en el Lago West Bearskin en Minnesota, EEUU, en 1991. Ver help(*`wblake`*) para más información de los datos. Las escamas de los peces tienen anillos circulares como los árboles, y contándolos se puede determinar la edad (en años) de un pez. La base de datos también tiene la variable* `Scale` *que mide el radio de las escamas en mm., que no utilizaremos por ahora.*

(a) Hacer un scatter plot de la longitud (`Length`*) en función de la edad (*`Age`*). ¿Qué observa? La apariencia de este gráfico es diferente de los demás gráficos de dispersión que hemos hecho hasta ahora. La variable predictora* `Age` *sólo puede tomar valores enteros, ya que se calculan contando los anillos de las esacamas, de modo que realmente estamos graficando ocho poblaciones distintas de peces. Como es esperable, la longitud crece en general con la edad, pero la longitud del pez más largo de un año de edad excede la longitud del pez más corto de cuatro años de edad, por lo que conocer la edad de un pez no nos permitirá predecir su longitud de forma exacta.*

(b) *Calcule las medias y los desvíos estándares muestrales para cada uno de las ocho subpoblaciones de los datos de las percas. Dibuje un boxplot de la longitud para cada edad de las percas, todos en la misma escala. Describa lo que ve. La longitud, ¿parece aumentar con la edad? La dispersión de la longitud, ¿parece mantenerse más o menos constante con la edad? ?¿O crece? ¿O decrece?*

(c) *Ajuste un modelo lineal para explicar la longitud (`Length`) en función de la edad (`Age`). ¿Resulta significativa la pendiente? Resuma la bondad del ajuste con el R^2. Superponga la recta estimada al gráfico de dispersión, y también las medias muestrales por grupos. Halle el estimador de σ que proporciona el modelo lineal. ¿A qué valor debiera parecerse? ¿Se parece? Observar que no debiera parecerse a `sd(Length)`. ¿Le parece que el ajuste obtenido por el modelo lineal es satisfactorio?*

(d) *Obtenga intervalos de confianza de nivel 95 % para la longitud media a edades 2, 4 y 6 años (no simultáneos). ¿Sería correcto obtener IC para la longitud media a los 9 años con este conjunto de datos?*

4. Ejercicios de Diagnóstico para Regresión Lineal Simple

Ejercicio 4.1 *Madres e hijas II. Archivo de datos `heights.txt` del paquete `alr3`. Continuando con el ejercicio 3.2, en el que proponemos ajustar el modelo lineal simple para explicar la altura de la hija, `Dheight`, a partir de la altura de la madre, llamada `Mheight`, como la variable predictora.*

(a) *Hacer gráficos para evaluar la adecuación del modelo linel para explicar los datos.*

(b) *Compare el ajuste clásico con el ajuste robusto propuesto.*

(c) *Concluya respecto de la adecuación del modelo lineal en este caso.*

Ejercicio 4.2 *Medidas del cuerpo V. Base de datos `bdims` del paquete `openintro`.*

(a) *Realice gráficos de que le permitan evaluar los ajustes realizados en los ejercicios 2.1 y 2.2 con esta base de datos, tanto para explicar el peso por el contorno de cintura como el ajuste para explicar el peso por la altura. ¿Lo conforman estos modelos ajustados?*

(b) *Compare el ajuste clásico del modelo lineal con el ajuste robusto. ¿Cambian mucho los modelos ajustados? ¿Qué indica esto? No se desanime, este ejercicio sigue en el capítulo próximo.*

Ejercicio 4.3 *Mamíferos, Parte V. Base de datos `mammals` del paquete `openintro`.*

(a) *En el ejercicio 1.7 observamos que el scatter plot del peso del cerebro de un mamífero (`BrainWt`) en función de su peso corporal (`BodyWt`) no se podía describir como una pelota de rugby más o menos achatada. Supongamos que no hubiéramos hecho el gráfico de dispersión, e intentemos ajustar un modelo lineal a los datos. Ajuste el modelo lineal simple que explica `BrainWt` en función de `BodyWt`. Luego realice el gráfico de residuos versus valores predichos. El gráfico de residuos estandarizados versus valores predichos. El de residuos estudentizados versus valores predichos. ¿Difieren mucho entre sí?*

(b) *Use el test de outliers basado en los residuos estudentizados. Indique cuáles son las observaciones candidatas a outliers.*

(c) *Calcule los leverages. Identifique las observaciones candidatas a más influyentes según este criterio. Calcule las distancias de Cook, vea cuáles son las observaciones influyentes.*

(d) *Compare con el ajuste robusto.*

(e) *Finalmente, para el modelo de regresión propuesto en el ejercicio 3.5 para vincular los logaritmos en base 10 de ambas variables, haga un gráfico de residuos versus valores predichos, y algunos otros gráficos de diagnóstico. ¿Le parece que este modelo ajusta mejor a los datos?*

Ejercicio 4.4 *Hacer un ajuste robusto a los datos de perímetro cefálico y edad gestacional. Comparar con el ajuste clásico. Identificar la presencia de outliers. ¿Son muy influyentes en el ajuste? Recordar que de todos modos este no es el último modelo que probaremos sobre estos datos. Datos en el archivo* `low birth weight infants.txt`

5. Ejercicios Regresión Lineal Múltiple (primera parte)

Ejercicio 5.1 *Medidas del cuerpo V. Base de datos* `bdims` *del paquete* `openintro`.

(a) En el ejercicio 2.1 explicamos el peso de las personas registradas en esta base de datos, por el contorno de la cadera y en el ejercicio 2.2 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.*

(b) 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?

(c) 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 2.1 y 2.2. ¿Es correcto comparar los R^2 de los tres ajustes? ¿Qué valores puedo comparar? ¿Es mejor este ajuste múltiple?

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

(e) 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.

(f) 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.

Referencias

Heinz, G., Peterson, L. J., Johnson, R. W., y Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education*, *11*(2).

Weisberg, S. (2005). *Applied linear regression* (3rd. ed. ed.). John Wiley & Sons.