Modelos Lineales

Ejercicios propuestos

1. Este ejercicio tiene por objetivo comprender el modelo de regresión simple. Genera $n = 100$ observaciones de la variable independiente X según $\mathcal{U}(0, 10)$: $x_1, \ldots, x_n$. Genera los residuos y las variables dependientes según las tres situaciones que te presentan:

Situación A	Situación B	Situación C
$\epsilon \sim N(0, 4)$	$\epsilon \sim N(0, 4)$	$\epsilon \sim N(0, 4X/9)$
$E(Y\|X) = 2 + 3X$	$E(Y\|X) = 2 + 3X - X^2$	$E(Y\|X) = 2 + 3X$
$VAR(Y\|X) = 4$	$VAR(Y\|X) = 4$	$VAR(Y\|X) = 4X^2/9$

   (a) ¿Cuál de estas tres situaciones está dentro del modelo de regresión lineal estudiado en clase?

   (b) Realiza gráficos de dispersión de Y sobre X para cada situación. ¿Puedes identificar en cada gráfico qué aspectos del modelo no se cumplen?

2. Genera n observaciones según:

   - $X \sim \mathcal{U}(0, 10)$
   - $\epsilon \sim N(0, 4)$
   - $Y = 2 + 3X + \epsilon$

   Toma el conjunto de datos que has generado $(x_1, y_1), \ldots, (x_n, y_n)$, ajusta el modelo de regresión simple y guarda las estimaciones que has obtenido. Repite el proceso 1000 veces y rellena la siguiente tabla:

Modelo teórico:
$E(Y|X) = 2 + 3X\,, \quad VAR(Y|X) = 4$

n	$(b_{0(0.025)}, b_{0(0.975)})$	$(b_{1(0.025)}, b_{1(0.975)})$	$(s^2_{(0.025)}, s^2_{(0.975)})$
20			
50			
100			
1000			

   ¿Te parece que las estimaciones que obtienes por mínimos cuadrados son razonables?

   Nota: Entendemos por $b_{0(\alpha)}$ el percetil $\alpha \times 100$ de la distribución de las estimaciones b_0 que has obtenido ($\alpha \in (0, 1)$). De la misma manera para estimaciones $b_{1(\alpha)}$ y $s^2_{(\alpha)}$.

3. Consideraremos el ejemplo propuesto por Anscombe(1973). Los datos los encontrarás en el fichero *anscombe.txt*. Se consideran los datos como 4 pares de variables cuantitativas y se quiere estudiar la relación lineal entre cada par de variables. Es decir, se quiere estudiar la asociación entre los pares $x123$ e $y1$, $x123$ e $y2$, $x123$ e $y3$ y finalmente $x4$ e $y4$.

   (a) Importa los datos y calcula la correlación lineal entre cada par de variables. ¿Qué par de variables dirías que están correlacionadas? ¿En qué sentido?

   (b) Calcula la recta de regresión para cada par de variables (Considera $y*$ como la variable dependiente y $x*$ como la independiente). ¿Cuáles son las estimaciones de los coeficientes de regresión?

   (c) Calcula el porcentaje de variabilidad explicada (R^2) por cada modelo.

   (d) Representa gráfricamente los residuos del modelo frente a las predicciones del modelo. ¿ Qué observas?

   (e) Representa gráficamente cada par de variables. ¿Qué conclusiones obtienes?

4. Queremos estudiar la relación entre el peso (Y) y la altura (X). Para ello hemos recogido una muestra de 10 chicas de 18 anõs. Los datos son los siguientes:

Altura (cm)	169.6	166.8	157.1	181.1	158.4	165.6	166.7	156.5	168.1	165.3
Peso (Kg)	71.2	58.2	56.0	64.5	53.0	52.4	56.8	49.2	55.6	77.8

   (a) Representa gráficamente estos datos y decide si la regresión lineal puede ser un modelo adecuado para estudiar la posible asociación entre el peso y la altura.

   (b) Dibuja conjuntamente el gráfico de dispersión de los datos así como la recta de regresión estimada. (`abline()`)

   (c) Estima los coeficientes de regresión. Calcula $\bar{x}$ y $\bar{y}$ y comprueba que $(\bar{x}, \bar{y})$ es un punto de la recta de regresión.

   (d) Dibuja conjuntamente, los datos, la recta de regresión y en centro de gravedad $(\bar{x}, \bar{y})$.

   (e) Comprueba la igualdad de la descomposición de la variabilidad que has visto en clase
   $$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2.$$

   (f) En base a esta descomposición calcula R^2 y el valor del estadístico F. Comprueba que coinciden con los calculados directamente por R.

5. En la librería `MASS` está el conjunto de datos `Animals`. Cárgalo en memoria y contesta a las siguientes preguntas:

   (a) Haz los diagramas de dispersión del *peso del cerebro (Y)* frente al *peso del cuerpo(X)* y del logaritmo del peso del cerebro ($log(X)$) frente al logaritmo del peso del cuerpo ($log(Y)$). ¿Qué situación crees que se ajustará mejor con el modelo de regresión lineal? ¿Por qué?

   (b) Calcula el modelo $M1 : log(Y) \sim log(X)$. Estudia los residuos. Qué observas? Estudia las distancias de Cook. ¿Destacan algunos individuos? ¿Tienen alguna coincidencia? ¿Crees que el modelo $M1$ es adecuado?

   (c) Crea el factor que indica si un animal es dinosaurio o no (Z). Ajusta el modelo $M1 : log(Y) \sim log(X) + Z$ Estudia los residuos de este modelo. Calcula las distancias de Cook. ¿Destacan algunos individuos? ¿Crees que el modelo $M2$ es adecuado?

   (d) Para animales con $log(X) = 10$, que valor medio de $log(Y)$ se espera si no son dinosaurios? Y si fueran dinosaurios?

6. El conjunto de datos *edss.dat* recoge información de 32 pacientes con esclerosis múltiple. Las variables recogidas son:

   - Sexo (M: hombre, F: mujer)
   - Edad
   - Fecha de nacimiento
   - Edad de la aparición de la enfermedad
   - Puntuación EDSS medida 5 años después de la aparición de la enfermedad

     Nota: EDSS, Expanded Disability Score System, es un método muy utilizado para medir la progresión de la enfermedad a partir de criterios clínicos de fácil acceso para los neurólogos.

     `http://www.mult-sclerosis.org/expandeddisabilitystatusscale.html`

   El objetivo es estudiar las relaciones entre el sexo, la edad de aparición de la enfermedad y el EDSS. Por ejemplo,

   (a) ¿qué tipo de análisis deberíamos hacer para estudiar la relación entre EDSS y la edad de aparición de la enfermedad?

   (b) ¿qué tipo de análisis deberíamos hacer para estudiar la relación entre EDSS y el sexo?

   (c) ¿qué tipo de análisis deberíamos hacer para estudiar la relación entre EDSS, la edad de aparición de la enfermedad y el sexo? Según este modelo, ¿cuál es la media estimada de EDSS para una mujer que empezó con la enfermedad a los

29 años? ¿Y para un hombre para el que la enfermedad también surgió a los 29 años?