Regresión Avanzada
Práctica 1: Regresión Lineal Simple

Ejercicios teóricos.

1. (a) Cuál es la diferencia entre y_i e $\hat{y}_i$ en el análisis de regresión?
   (b) Qué es el término ϵ en el modelo de regresión. Por qué ocurre?
   (c) Mostrar que la recta de regresión lineal estimada por mínimos cuadrados siempre pasa por $(\bar{x}, \bar{y})$.
   (d) Asumiendo que $\bar{x} = 0$ y que $\bar{y} = 0$, mostrar que R^2 y $\hat{\rho}_{XY}$ coinciden.
   (e) Mostrar que $\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0$.
   (f) Mostrar que $\sum_{i=1}^{n} x_i r_i = 0$
   (g) Mostrar que $\sum_{i=1}^{n} \widehat{y}_i r_i = 0$
2. Hallar el sesgo de $\widehat{\beta}_0$ y $\widehat{\beta}_1$.
3. Mostrar que $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son estimadores consistentes.
4. Calcular $cov(\widehat{\beta}_0, \widehat{\beta}_1)$.
5. Considerar el modelo lineal con las variables centradas

$$y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \epsilon_i$$

   (a) Hallar los estimadores de β_0 y β_1
   (b) Mostrar que son insesgados. Especificá que supuestos usaste.
6. Bajo los supuestos usuales y asumiendo normalidad de los errores encontrar los estimadores de máxima verosimilitud para β_0 y para β_1. Compararlos con los estimadores de mínimos cuadrados.

Ejercicios Prácticos

1. Un economista del Departamento de Recursos Humanos de Florida State está preparando un estudio sobre el comportamiento del consumidor. Él recolectó los datos que aparecen en miles de dólares para determinar si existe una relación entre el ingreso del consumidor y los niveles de consumo. Teniendo en cuenta los datos de `florida.txt`

(a) Cuál es la variable dependiente e independiente?

(b) Hacer un diagrama de dispersión para los datos.

(c) Calcular e interpretar el modelo de regresión. Dar intervalos de confianza de nivel 95% para los parámetros estimados. Qué dice el modelo sobre la relación consumo - ingreso? Qué proporción de cada dólar adicional que se gana se invierte en consumo?

(d) Qué consumo pronosticaría para alguien que tiene un ingreso de $u\$s27500$?

(e) Cuál es el error estándar de los residuos para el Departamento de Recursos Humanos de Florida State? Cómo se interpretan los resultados?

(f) Calcular e interpretar el coeficiente de correlación entre el ingreso del consumidor y los niveles de consumo.

2. En este ejercicio trabajaremos con el conjunto de datos `Auto` de la librería `ISLR` de R. El objetivo final ser predecir el consumo de combustible a travs de alguna de las variables que presenta el conjunto de datos.

   (a) Visualizar el conjunto de datos utilizando el comando `fix`.

   (b) Realizar un análisis explorio de las variables. Decidir que tipo de variables (cualitativas, cuantitativas), mostrar gráfica y analíticamente su vinculación.

   (c) Ajustar un modelo de regresión lineal simple para modelar la variable `mpg` (milles per gallon) teniendo como variable regresora aquella que consideres más adecuada del item anterior.

      i. Hay una relación entre estas variables?

      ii. Cuán fuerte es esta relación? Es positiva o negativa?

      iii. Cuál es el valor predicho para `mpg` asociado al valor mediano de la variable regresora?

   (d) Realizar un gráfico de dispersión para las variables y en el mismo graficar la recta de regresión ajustada. Notas algún inconveniente en el gráfico?

3. Consideramos el conjunto de datos `grasas.txt` que corresponden a tres variables medidas en 25 individuos: `edad`, `peso` y `cantidad de grasas` en sangre.

(a) Realizar un análisis exploratorio de las variables. Ver como son las interacciones entre ellas y que características principales se observan.

(b) Ajustar un modelo de regresión lineal simple para modelar la variable `grasas` teniendo como variable regresora aquella que consideres más adecuada del item anterior.

   i. Hay una relación entre estas variables?

   ii. Cuán fuerte es esta relación? Es positiva o negativa?

   iii. Cuál es el valor predicho para `grasas` asociado al valor mediano de la variable regresora?

(c) Realizar un gráfico de dispersión para las variables y en el mismo graficar la recta de regresión ajustada.

(d) Dar el intervalo de confianza de nivel 0.95% para la pendiente e intercept de la recta de regresión e interpretarlo.

4. Se consideran los datos relativos al fracaso escolar en las diferentes comunas de Madrid, los datos se encuentran en el archivo `fracaso.txt` corresponden al año 2003 y ciclo lectivo 2003-2004. Para cada comuna se tiene los datos correspondientes a la renta per capita bruta media (en euros) y el fracaso escolar (en porcentaje).

   (a) Realizar un análisis exploratorio de los datos.

   (b) Ajustar un modelo lineal para explicar el fracaso escolar en función de la renta.

   (c) En un diagrama de dispersión graficar los datos y la recta ajustada en el item anterior. Comentar que se observa.

   (d) Calcular intervalos de confianza a nivel 0.95% para la pendiente y para el intercept. Interpretarlos.

   (e) A nivel 5% podemos afirmar que a niveles más altos de renta están asociados a niveles más bajos de deserción escolar?

   (f) Cuánto vale el coeficiente de correlación entre el nivel de renta y el porcentaje de fracaso escolar?

   (g) Qué porcentaje de fracaso escolar se predice en una poblacion cuya renta es $x_0 = 13000$ euros?

   (h) Cuál es el residuo correspondiente a Colmenar Viejo?

5. Considerar el modelo lineal

$$y = \beta_0 + \beta_1 x + \epsilon$$

   (a) Generar 100 datos que verifiquen la ecuación 1 con $\beta_0 = 1$ y $\beta_1 = 3$, con errores independientes y normales con $\mu = 0$ y $\sigma = 2$

   (b) Graficar estos datos.

   (c) Estime los valores de β_0 y β_1 utilizando el método de mínimos cuadrados y guarde el resultado de $\hat{\beta}_0$ y $\hat{\beta}_1$ en dos vectores. En el gráfico anterior agregar en diferentes colores la recta *verdadera* y la recta estimada.

   (d) Repita los items (a) y (b) 1000 veces.

   (e) Estudie los resultados obtenidos. ***Hint***: use los comandos *hist*, *mean*, *var*.

6. Considerar el modelo lineal del ejercicio 5. En este ejercicio estudiaremos que ocurre con los desvíos estandares de los estimadores.

   (a) Generar 100 datos que verifiquen la ecuación 1 con $\beta_0 = 1$ y $\beta_1 = 3$, con errores independientes y normales con $\mu = 0$ y $\sigma = 10$.

   (b) Graficar y comparar con el gráfico del ejercicio 5.

   (c) Repetir 300 veces el item a. en todos los casos guardar las estimaciones de los parámetros.

   (d) Estimar el desvo estándar de los estimadores de β_0 y β_1.

   (e) Repetir este procedimiento para $\sigma = 1, 2, \ldots, 20$. Para cada valor de σ guardar el desvío estándar correspondiente.

   (f) Hacer un gráficos de σ versus esvío estándar. Qué conclusión se saca?