Tarea 3

Propedeutico Maestría en Ciencias de datos
Prof. Mauricio García Tec

1. Parte teórica

Esta parte del proyecto será sobre regresión lineal. Supongamos que quieren explicar una variable estadística Y(por ejemplo altura) utilizando la información de p variables $X^1, \ldots, X^p$ (peso, ancho de huesos, etc.). Si se toma una muestra de N individuos, cada variable está representada por un vector de tamaño N. La información de las variables explicativas se pueden juntar en una matriz

$$X = [X^1|\ldots|X^p],$$

de tamaño $n \times p$ donde cada columna es una variable y cada fila uno de los individuos de la muestra. Tienen que contestar lo siguiente:

1. Plantear el problema de regresión como un problema de mínimos cuadrados, encontrar el vector $\hat{\beta} = [\hat{\beta}_1, \ldots, \hat{\beta}_p]^\top$ que resuelva

   $$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2$$

   y encontrar la solución teórica. ¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos? ¿Podríamos usarlo para ajustar polinomios (ej $y = x^2$)?

2. Argumentar la relación entre la solución encontrada y un problema de proyección en subespacios vectoriales de álgebra lineal. ¿Cuál es la relación particular con el teorema de Pitágoras?

   Solución:

   Para la matriz de $n \times p$, X, se define el espacio columna de X como,

   $$R(X) = \{W : W = X\beta\}$$

   El mínimo de $||Y - W||$ para $W \in R(X)$ se alcanza en $\hat{W}$ tal que $(Y - \hat{W}) \perp R(X)$, i.e. cuando $Y - \hat{W}$ es ortogonal a todos los vectores en $R(X)$, que es cuando $\hat{W}$ es la proyección ortogonal de Y en $R(X)$. Dicho $\hat{W}$ existe y es único, además tiene la representación

   $$\hat{W} = PY = X(X^\top X)^- X^\top Y,$$

   donde $P = X(X^\top X)^- X^\top$ es el operador de proyección ortogonal sobre $R(X)$

3. ¿Qué logramos al agregar una columna de unos en la matriz X? Es decir, definir mejor

   $$X = [\mathbf{1}_n|X^1|\ldots|X^p],$$

   con $\mathbf{1}_n = [1, 1, \ldots, 1]^\top$.

   Solución:

   Bajo este planteamiento ahora se tendrá que Y debiese tomar la forma

   $$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i^1 + \ldots + \hat{\beta}_p X_i^p$$

   es decir habrá un valor $\hat{\beta}_0 = \hat{Y}$ cuando los demás valores son 0 (se le conoce como intercepto u ordenada al origen).

4. Plantear el problema de regresión ahora como un problema de estadística

$$Y_i = \beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \epsilon_i,$$

donde los errores son no correlacionados con distribución

$$\epsilon_i \sim N(0, \sigma^2)$$

Solución:

Como se está haciendo un análisis condicional se puede suponer X constante. Entonces

$$\begin{aligned}
\mathbb{E}(Y_i) &= \mathbb{E}(\beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \epsilon_i) \\
&= \beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \mathbb{E}(\epsilon_i) \\
&= \beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + 0 \\
&= \beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p
\end{aligned}$$

Y también

$$Var(Y_i) = Var(\beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

Entonces, como $\epsilon_i \sim N(0, \sigma^2)$, entonces

$$Y_i \sim N(\beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p, \sigma^2)$$

Además, como para $i \neq j$, ϵ_i y ϵ_j no están correlacionados entonces Y_i e Y_j no están correlacionados. Pero como tienen distribución normal, entonces además Y_i e Y_j son independientes. Entonces ahora el problema de determinar β se convirtió en un problema de inferencia.

5. ¿Cuál es la función de verosimilitud del problema anterior?

Solución:

Como $Y_1, \ldots, Y_n$ son independientes, entonces

$$\begin{aligned}
f_Y(y_1, \ldots, y_n) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 X_i^1 - \ldots - \beta_p X_i^p)^2 \right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i^1 - \ldots - \beta_p X_i^p)^2 \right\} \\
&= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0, \beta_1, \ldots . \beta_p) \cdot (1, X_i^1, X_i^2, \ldots, X_i^p))^2 \right\} \\
&= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta \cdot (1, X_i^1, X_i^2, \ldots, X_i^p))^2 \right\}
\end{aligned}$$

Sin embargo, el exponente de esta expresión se puede escribir como

$$\sum_{i=1}^{n} (y_i - \beta \cdot (1, X_i^1, X_i^2, \ldots, X_i^p))^2 =$$

$$[y_1 - \beta \cdot (1, X_1^1, X_1^2, \ldots, X_1^p), \ldots, y_n - \beta \cdot (1, X_n^1, X_n^2, \ldots, X_n^p)] *$$
$$[y_1 - \beta \cdot (1, X_1^1, X_1^2, \ldots, X_1^p), \ldots, y_n - \beta \cdot (1, X_n^1, X_n^2, \ldots, X_n^p)]^\top =$$
$$||y - X\beta||^2.$$

Entonces, la función de verosimilitud está dada por

$$L(\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}||Y - X\beta||^2\right\}$$

6. Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados.

   Solución:

   Nótese que la aplicación $t \mapsto \exp(-ct)$ es decreciente, entonces para maximizar $L(\beta, \sigma^2)$ con respecto a β simplemente se debe minimizar $\frac{1}{2\sigma^2}||Y - X\beta||^2$, i.e. simplemente se debe minimizar $||Y - X\beta||^2$, que es precisamente el problema de optimización de mínimos cuadrados.

   Para $\hat{\beta} = \hat{\beta}_{MLE}$ se tiene que maximizar con respecto a σ^2

$$L(\sigma^2) = L(\hat{\beta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}||Y - X\hat{\beta}||^2\right\}$$

   En este caso la log-verosimilitud está dada por

$$\ell(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}||Y - X\hat{\beta}||^2$$

   De aquí que

$$\frac{\partial}{\partial\sigma^2}\ell(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}||Y - X\hat{\beta}||^2$$

   Por lo tanto, $\frac{\partial}{\partial\sigma^2}\ell(\sigma^2) = 0$ si y solamente si

$$\sigma^2 = \frac{1}{n}||Y - X\hat{\beta}||^2 = \frac{1}{n}||Y - \hat{Y}||^2$$

2. Parte aplicada

1. ¿Qué tan bueno fue el ajuste? Una buena respuesta incluye argumentaciones teóricas y visualizaciones. Puntos adicionales si investigan como usar alguna de las librerias ggplot2 o plotly para sus gráficas.

   Solución:

   Sólo se realizó la regresión con respecto a las variables numéricas: carat, depth, table, x, y y z.

   Aplicando la función lm (se anexa código en archivo adjunto) se obtuvo

$$Price = 20849 + 10686 \cdot carat - 203 \cdot depth - 102 \cdot table - 1315x + 66y + 41z + \epsilon,$$

   donde $\epsilon \sim N(0, \hat{\sigma}^2)$ con $\hat{\sigma}^2 = 1497$. Con una R^2 de 0.8592 (cercana a 1) y un p-value menor a 2.2e-16. Por lo tanto se puede decir que el ajuste es bueno cerca de la media.

2. ¿Qué medida puede ayudarnos a saber la calidad del ajuste? ¿Cuál fue el valor de σ que ajustó su modelo y que relación tiene con la calidad del ajuste?

   Solución:

   Se puede demostrar que

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$$

   Equivalentemente,

$$1 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

   Definiendo,

$$1 = R^2 + \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

   El segundo término es el error, que se espera sea pequeño, así que se espera que R^2 sea cercano a 1.

3. ¿Cuál es el ángulo entre Y y $\hat{Y}$?. Hint: usen la y el arcocoseno.

   Solución:

   En este caso

$$0.8592 = R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{||\hat{y} - \bar{y}||^2}{||y - \bar{y}||^2}$$

   De aquí que el ángulo sea de 30.77 grados.

4. Definan una funcion que calcule la logverosimilitud de unos parámetros β y σ^2

   Solución:

   Podriamos definir una función

```
verosimilitud<- function(b){

prod(dnorm(diamonds$price),b[1] + beta[2]*diamonds$carat +

b[3]*diamonds$depth + b[4]*diamonds$table + b[5]*diamonds$x +

b[6]*diamonds$y + b[7]*diamonds$z, b[8]))

}
```

   sin embargo el cálculo del producto la haría ineficiente. Por eso se usará la equivalencia con la minimización de mínimos cuadrados.

5. Utilicen la función optim de R para numéricamente el máximo de la función de verosimilitud. Si lo hacen correctamente, su solución debe coincidir con la del método lm.

   Solución:

```
sumacuadrados<- function(beta){
sum((diamonds$price - beta[1]- beta[2]*diamonds$carat -

beta[3]*diamonds$depth - beta[4]*diamonds$table -

beta[5]*diamonds$x - beta[6]*diamonds$y - beta[7]*diamonds$z)^2)
}

optim(c(20000,10000,-200,-100,-1300,50,50), sumacuadrados,hessian=TRUE)
```

Obteniendose el vector $\hat{\beta} = (20953, 10746, -204, -100, -1358, 31, 115)$ que no es precisamente el que se obtuvo con lm ya que esta optimización con optim() se hace en dimensión 7, lo que dificulta cualquier algoritmo de optimización.