Universidad de Puerto Rico
Departamento de Matemáticas
Humacao, Puerto Rico 00791

Laboratorio I: Modelos de Regresión

Un problema bien común en las ciencias es el de aproximar o buscar un modelo que describa un conjunto de datos provenientes de un experimento. Con esto se quiere determinar las relaciones entre las distintas variables del problema y a su ves para hacer predicciones de estas a partir de ciertos datos. Veamos un caso concreto. En una cierta fabrica de madera les interesa medir la cantidad (en términos de volumen) de un cargamento de madera que acaban de recibir. La madera no esta procesada, es decir que lo que se recibe es un cargamento de árboles del mismo tipo pero de distintos tamaños. El proceso de medir el volumen de un árbol es tedioso de modo que no es practico medir el volumen de cada árbol. La idea detrás del proceso de *regresión* es seleccionar una "muestra" de los árboles y medir para estos la altura, diámetro, y volumen. Cabe señalar que la altura y diámetro de un árbol son fáciles de medir mientras que el volumen como dijimos es difícil de medir. Usando estos datos podemos ahora "construir" una formula para el volumen de un árbol como función del diámetro ó altura ó ambos. Esta formula se usa para "estimar" volúmenes de árboles a partir del diámetro ó altura.

Vamos ahora a abrir el archivo `tree_data` en MATLAB que contiene los datos de diámetros en pulgadas (`diam`), alturas en pies (`height`), y volúmenes en pies cúbicos (`vol`) para 31 árboles del mismo tipo. Esto lo hacemos mediante la instrucción:

```
tree_data
```

(Estamos suponiendo aquí que el usuario esta en el directorio donde se encuentran los archivos del laboratorio.) Con la instrucción `whos` puede ver una lista de las variables en su espacio de trabajo hasta el momento. Si quisiera ver digamos los datos de volúmenes, escriba `vol` y oprima la tecla de *ENTER* y le aparecerá en la pantalla la lista de los volúmenes. Usamos ahora la instrucción `plot` de MATLAB para hacer una gráfica de los volúmenes como función del diámetro:

```
plot(diam,vol,'.')
```

La parte `'.'` de la instrucción marca cada dato con un punto. Puede usar otros símbolos para marcar los datos. Escriba `help plot` en la ventana de MATLAB para obtener mas información sobre la instrucción `plot`. Debemos pensar ahora en diferentes "modelos" para describir los datos:

1. Modelo lineal en los coeficientes y los datos: $y = a + bx$.

2. Modelo lineal en los coeficientes y cuadrático en los datos: $y = a + bx + cx^2$.

3. Modelo nolineal en los coeficientes y exponencial en los datos: $y = a\mathrm{e}^{bx}$.

4. Modelo lineal en los coeficientes y periódico en los datos: $y = a\cos(x) + b\,\mathrm{sen}(x)$.

5. Modelo lineal en los coeficientes y exponencial en los datos: $y = a\mathrm{e}^{x} + b\mathrm{e}^{2x}$.

Así podemos continuar proponiendo diferentes modelos. Cual escogemos depende de consideraciones relacionadas al problema. Mirando la gráfica de volumen como función del diámetro, podemos pensar en un modelo con una recta como en (1) arriba.

Ejercicio 0.1. Usando la copia que se le entregó de la gráfica de volúmenes como función de los diámetros, dibuje la mejor recta según su criterio para aproximar estos datos. Calcule su pendiente e intercepto en el eje de las y.

Regresión Lineal

Vamos a concentrarnos ahora en modelos lineales en los coeficientes y los datos como el (1) arriba. ¿Cómo podemos seleccionar la mejor recta? Necesitamos un criterio o forma de "medir" cuan buena es una recta aproximando los datos. Vamos a representar los datos como el conjunto de pares ordenados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Aquí puede pensar que las x's representan digamos los diámetros en el problema de los árboles y las y's representan los volúmenes correspondientes. La recta que buscamos tiene la forma $y = a + bx$. Tenemos que determinar los coeficientes a, b. Primero observamos que la *diferencias o residuos entre los datos y el modelo* están dadas por:

$$y_i - (a + bx_i) \quad , \quad i = 1, 2, \ldots, n. \tag{0.1}$$

Como medida de cuan bien el modelo se ajusta a los datos, introducimos el *error cuadrado medio*, denotado RMSE, el cual esta dado por:

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - (a + bx_i)\right)^2\right]^{1/2}. \tag{0.2}$$

El error cuadrado medio es una función de a, b. Queremos seleccionar a, b que hagan a RMSE lo más pequeño posible. Usando técnicas del calculo, se puede verificar que el RMSE más pequeño se obtiene con

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad , \quad a = \bar{y} - b\bar{x}, \tag{0.3}$$

donde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{0.4}$$

Para un valor de z dado que no tiene que ser igual a ninguno de los x_i, la predicción del modelo dada por $a + bz$ tiene un error de $\pm$RMSE con una *probabilidad bien alta.*

Los coeficientes a, b de la mejor recta los podemos calcular en MATLAB con la instrucción `polyfit`. Esta instrucción es lo equivalente a calcular con las formulas (0.3), (0.4). Por ejemplo, para calcular a, b para la mejor recta para los datos de volumen como función del diámetro, escribimos:

```
coef=polyfit(diam,vol,1)
```

Observará ahora en la pantalla los valores de a, b: la b primero y luego la a. Note que el tercer argumento de la función `polyfit` especifica el grado del mejor polinomio en este caso uno porque usamos una recta para aproximar los datos. Si en lugar de uno usa un dos como tercer argumento entonces estaría buscando el mejor polinomio de grado dos que aproxima a los datos (modelo (2) arriba). Los resultados de `polyfit` los podemos usar junto con el programa `reg_line` para hacer gráficas de los datos, la mejor recta, y los residuos y el calculo de RMSE. Trate la siguiente secuencia de instrucciones a ver que resulta:

```
coef=polyfit(diam,vol,1);
reg_line(diam,vol,coef)
```

Ejercicio 0.2. Compare los estimados que obtuvo a mano de los valores de a, b para los datos de volúmenes como función del diámetro con los obtenidos con la función `polyfit`. Estime el volumen de un árbol cuyo diámetro es 15 pulgadas. Estime los parámetros a, b para los volúmenes como función de la altura. ¿Cúal usted cree están mejor correlacionados: volumen y diámetro ó volumen y altura? Trate con un modelo cuadrático para los volúmenes como función de la altura.

Ejercicio 0.3. Calcule a, b, y RMSE para los datos en el archivo `weight_data` que contiene los pesos en libras para mujeres embarazadas al momento de concepción (`weightc`) y al momento de parir (`weightd`). Estime el peso al momento de parir para una mujer que pesaba 142 libras al momento de concepción.

Gráficas de Residuos y un Modelo Exponencial

Consideremos ahora los datos del archivo `cancer_data` que representan el crecimiento de células cancerosas como función del tiempo. Escriba la siguientes instrucciones para ver que resulta:

```
cancer_data
coef=polyfit(time,cells,1);
reg_line(time,cells,coef)
```

Estamos aquí buscando la mejor recta que aproxima a los datos. Observamos aquí algo particular en la gráfica de los residuos: tienen un patrón bien definido. Esto es indicativo de que el modelo lineal no representa correctamente la relación entre la cantidad de células y el tiempo transcurrido.

Ejercicio 0.4. Examine las gráficas de residuos para volumen como función de diámetro en los datos de los árboles y para los datos de los pesos de mujeres embarazadas. ¿Qué patrón observa en los residuos?

Para crecimiento de células o poblaciones en general, los modelos exponenciales son más apropiados. Así que consideramos un modelo de la forma $y = ae^{bx}$. Note que la b aparece en forma nolineal en este modelo. Podemos mediante una transformación arreglar esto en este caso. Note que si $y = ae^{bx}$ entonces $\log y = \log a + bx$, i.e., tenemos un modelo lineal para los datos $(x_i, \log y_i)$, $i = 1, \ldots, n$ con coeficientes $\log a, b$. Para hacer la regresión usamos las instrucciones:

```
coef=polyfit(time,log(cells),1);
a=exp(coef(2));
b=coef(1);
```

Esto es lo que hace el programa `reg_exp` además de que genera unas gráficas.

Ejercicio 0.5. Calcule los RMSE para el modelo exponencial y el modelo lineal para los datos de células cancerosas. ¿Cúal usted cree es el mejor modelo? Utilice los RMSE y las gráficas de residuos.

Referencias

[1] Straffin, P., Applications of Calculus, MAA Notes Number 29, The Mathematical Association of America, 1993.