Regresión lineal con R
Mathieu Kessler
Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena

Regresión lineal múltiple con R

1 Regresión lineal simple

1.1 Old Faithful

Para ilustrar los comandos de `R` que están asociados a la regresión lineal, utilizaremos los datos de las erupciones del geyser Old Faithful, que vimos en la primera sesión.

```
> geyser.dat <- read.table(file="geyser.txt", dec=",",sep=";", header=F, col.names
= c("duracion", "intervalo"))
> attach(geyser.dat) # permite aceder directamente a las variables de geyser.dat
```

Empecemos por una representación gráfica de `intervalo` en función de `duracion`

```
> plot(duracion, intervalo)
```

Añadimos la recta ajustada:

```
> abline(lm(intervalo ~ duracion))
```

Hemos utilizado en esta última instrucción el comando fundamental para la regresión lineal en `R`, que es `lm` (por "linear model"): utiliza como argumento principal una **fórmula** de `R`:

```
lm(intervalo ~ duracion)
```

Con esto indicamos que queremos llevar a cabo el ajuste de `intervalo` respecto a `duracion`.
En el caso en que, por razones físicas, queremos ajustar la nube de puntos por una recta que pasa por el origen (ecuación $y = ax$), podemos indicarlo en la fórmula con `intervalo ~ duracion - 1`
La salida de `lm` es un objeto que contiene mucha información, podemos empezar por almacenar este objeto, dandole el nombre `geyser.lm` por ejemplo.

```
> geyser.lm = lm (intervalo ~ duracion)
```

Podemos obtener un resumen del ajuste,

```
> summary(geyser.lm)
```

```
Call:
lm(formula = intervalo ~ duracion)
Residuals:
      Min       1Q   Median       3Q      Max
 -14.1130  -4.4802  -0.4712   4.0370  16.8153
Coefficients:
             Estimate  Std.  Error  t value    Pr(>|t|)
 (Intercept)   33.9668       1.4279    23.79  <2e-16 ***
    duracion   10.3582       0.3822    27.10  <2e-16 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  6.159 on 220 degrees of freedom
Multiple R-Squared:  0.7695, Adjusted R-squared:  0.7685
F-statistic:  734.6 on 1 and 220 DF, p-value:  < 2.2e-16
```

Deducimos que la ecuación de la recta ajustada es

$$intervalo = 33.97 + 10.36 \cdot duracion,$$

que el valor de coeficiente de determinación múltiple R^2 es 0.77, que los dos coeficientes son significativos.
Podemos aceder a los valores ajustados, los residuos, y los coeficientes con

```
> geyser.lm$fitted
> geyser.lm$resid
> geyser.lm$coef
```

Finalmente, podemos comprobar que los errores son aprox. normales, que la varianza es constante, e identificar posibles observaciones influyentes con las gráficas de diagnóstico. Una de las gráficas más útiles corresponde a la representación de los residuos en función de los valores ajustados. Podemos comprobar con ello la linealidad y la homoscedasticidad (los residuos se deben repartir en una banda de amplitud más o menos constante alrededor del eje Ox).
Para ello, podemos aplicar la función plot al objeto resultado del ajuste lineal:

```
plot(geyser.lm,which=1)
```

Cambiando el valor del argumento `which`, obtenemos distintos gráficos de diagnóstico, aunque en estas prácticas nos centremos en el de residuos frente a valores ajustados.

Problemas

1. **Nivel del mar en Venecia** Queremos estudiar la evolución del máximo anual del nivel del mar (en cm) en Venecia. Los datos de los que disponemos corresponden a los años 1931-1981, y están contenidos en el fichero `venecia.txt` (Datos reales,

publicados en Smith R.L, "Extreme value theory based on the r largest annual events", *Journal of Hydrology*, 86 (1986))

2. **Resistencia del cemento.** Se quiere estudiar la resistencia de unas piezas de cemento en función de su edad en días.

Edad (días)	Resistencia(kg/cm^2)				
1	13.0	13.3	11.8		
2	21.9	24.5	24.7		
3	29.8	28.0	24.1	24.2	26.2
7	32.4	30.4	34.5	33.1	35.7
28	41.8	42.6	40.3	35.7	37.3

Los datos están en el fichero `cemento.txt`. Proponer un modelo que relacione la resistencia con el tiempo de secado.

3. **Producción mundial de petroleo.**

Se quiere estudiar la evolución de la producción mundial de petroleo de 1880 a 1973. Los datos se encuentran en el fichero `petroleo.txt`

4. **Hidrólisis del éster** La hidrólisis de un cierto éster tiene lugar en medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial de 30 mM del éster, se han medido las concentraciones del mismo a diferentes tiempos obteniéndose los resultados siguientes (fichero `ester.txt`)

T (mn)	3	4	10	15	20	30	40	50	60	75	90
C 10^{-3}(M)	25.5	23.4	18.2	14.2	11	6.7	4.1	2.5	1.5	0.7	0.4

   (a) Realice una nube de puntos de las dos variables. ¿Le parece adecuado un modelo lineal para escribir este conjunto de datos?

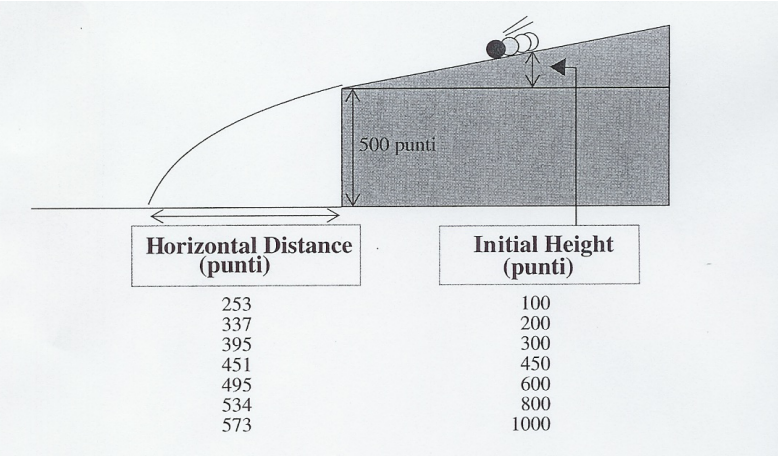
   (b) Defina una nueva variable Y' que sea $Y' = ln(conc)$ y realizar la nube de puntos Y' en función de t.

   (c) Realizar un ajuste por mínimos cuadrados de Y' sobre t con un modelo del tipo: $y = ax + b$. ¿Cuál es el modelo teórico que propone para C en función del tiempo?

   (d) Nos dan la información adicional de que se sabe con exactitud que la concentración inicial para $T = 0$ era igual a $30 \cdot 10^{-3}M$. ¿Cómo podemos incluir esta información en nuestro modelo?

2 Regresión lineal múltiple

En el caso en que tenemos más de una variable explicativa, los comandos son los mismos que los vistos en el apartado anterior para regresión lineal simple. La única diferencia radica en la expresión de la fórmula que especifica el modelo de interés en `lm`.
Ilustramos el análisis de regresión lineal múltiple con un ajuste polinomial con los datos históricos de Galileo.
En 1609 Galileo demostró matemáticamente que la trayectoria de un cuerpo que cae con un componente de velocidad horizontal es una parábola. Su descubrimiento tuvo su origen en observaciones empíricas que realizó casi un año antes. Para estas observaciones, ideó un experimento en el que una bola empapada de tinta rodaba en un plano inclinado para luego caer desde una altura de 500 punti (1 punti= 169/189mn). Galileo estudió la distancia horizontal que alcanza la bola en función de la altura desde la que sale. Un diagrama ilustrativo, extraido de Ramsey, Schafer (2002), "The statistical Sleuth" p 268, se enseña a continuación:



Empezamos por importar los datos del fichero `galileo.txt`,

```
> galileo <- read.table("galileo.txt", header=F, col.names =c("d","h"))
```

Un gráfica de `d` en función de `h` nos convence que debemos probar ajustar un polinomio de grado 2 (¿ó 3?). El modelo sería

$$d = a_0 + a_1 h + a_2 h^2,$$

o

$$d = a_0 + a_1 h + a_2 h^2 + a_3 h^3.$$

Después de `attach galileo`, creamos el objeto galileo.lm que contenga el análisis de regresión lineal con un polinomio de grado 2:

```
> galileo.lm = lm(d~ h+I(h^2))
```

Hemos utilizado en la fórmula la notación `I(h^2)` para indicar que queremos incluir en el modelo la potencia de grado 2 de `h`.
Obtenemos

```
> summary(galileo.lm)
Call:
lm(formula = d ~ h + I(h^2))

Residuals:
       1      2       3      4       5        6      7
 -13.235  8.018  12.996  2.071  -5.724  -12.249  8.123

Coefficients:
               Estimate  Std.  Error  t value      Pr(>|t|)
 (Intercept)   2.028e+02    1.580e+01   12.832  0.000213 ***
 h             6.983e-01    7.055e-02    9.899  0.000585 ***
 I(h^2)       -3.362e-04    6.296e-05   -5.340   0.005926 **
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

Residual standard error:  12.86 on 4 degrees of freedom
Multiple R-Squared:  0.9913, Adjusted R-squared:  0.9869
F-statistic:  227.6 on 2 and 4 DF, p-value:  7.586e-05
```

La interpretación de la salida es la misma que en el caso de la regresión lineal simple.

Problemas

1. **Estimación del volumen de madera de un arbol.**

   En ingeniería forestal existe la necesidad evidente de poder predicir el volumen de madera disponible de un tronco de un arbol todavía en pie. El método más sencillo consiste en medir el diámetro cerca del suelo y la altura del tronco y estimar el volumen utilizando estas dos cantidades. En el fichero `cerezos.txt` están los datos de un experimento realizado en un parque nacional de Pennsylvania donde se midió con cuidado el volumen después de cortar el tronco de (v : volumen, d : diametro y a : altura)

   (a) Realizar el análisis de regresión lineal del volumen sobre el diámetro y la altura. Estudie en particular los residuos.

   (b) Si se supone que el tronco es un cílindro perfecto, ¿cuál sería la relación entre v, a y d ? Proponer una transformación sobre los datos que sea acorde con esta relación física Realizar el ajuste lineal correspondiente con especial interés en el análisis de los residuos.

   (c) Si se supone que el tronco es un cono perfecto, ¿cuáles deberían ser los valores de los parámetros del apartado anterior?.

2. **Calor emitido por el fraguado de cemento.** Se estudia la relación entre la composición de un cemento tipo Portland y el calor desprendido durante la fase de fraguado. Los datos se pueden encontrar en el fichero `hald.txt`. La variable Y es la cantidad de calor desprendido en calorías por gramos de cemento, mientras que las variables X1, X2 X3 y X4 representan el contenido en porcentaje de cuatro productos A, B, C y D.

   (a) Obtener la matriz de correlaciones de las distintas variables. (*Si* `x` *es un data.frame, la matriz de correlaciones de sus variables se obtiene con* `cor(x)`).

   (b) Realizar un ajuste lineal.

3. **Consumo de helados** Se quisó identificar los factores más influyentes en el consumo de helados. Para ello se midió en una familia durante 30 semanas entre el 18 de marzo de 1953 hasta 11 de julio 1953 el consumo semanal de helado por persona (y), junto con las cantidades siguientes que se pensaba podían tener alguna influencia sobre el consumo : p el precio de una pinta de helado, i los ingresos semanales de la familia, temp : la temperatura media de la semana. También aparece el número de la semana. Los datos están en el fichero `helados.txt`

   (a) Represente gráficamente el consumo de helados en función de las semanas.

   (b) Determinar la matriz de correlación de las variables `y, p, i` y `temp`. ¿Cuál es la variable que parece tener más influencia en `y`?

   (c) Realizar un ajuste lineal de y sobre p,i y temp. ¿Qué vale la varianza residual y R^2?

   (d) Realizar un ajuste lineal de y sobre i y temp. Misma pregunta que en el apartado anterior

4. **Perdida de peso.** Se sabe que un determinado producto pierde peso después de ser producido. En el archivo `peso.txt` se ha recogido la diferencia (peso nominal-peso real) para varias unidades en distintos tiempos.

   (a) Ajustar un modelo de regresión lineal simple para explicar la evolución de la diferencia de peso en función del tiempo.

   (b) Realizar la gráfica de los residuos en función de los valores ajustados. ¿Le parece adecuado nuestro modelo para analizar estos datos? ¿Tiene alguna idea para mejorarlo?