2.15 Selección de variables independientes.

Para la regresión se asumía: variables **independientes** conocidas y utilizadas en el modelo, linealidad en sus p parámetros y, que la ecuación ajustada se usaría dentro de los rangos de valores de donde se extrajo el modelo. Dentro de estos conceptos es posible hacer modificaciones sobre todo en el número de variables, algunas de las cuales pueden ser eliminadas del modelo.

El análisis de regresión presupone que el conjunto de variables por incluir en la ecuación ha sido bien decidido. Sin embargo, en muchas de sus aplicaciones el conjunto no es predeterminado por lo cual la primera parte del análisis, a menudo, debe seleccionar las variables y para algunos procesos quizás esto llegue a ser lo más importante (Chatterjee & Price 1977).

En muchos procesos subjetivos en los cuales es escasa una formulación teórica de un problema, aparece un gran número de variables de cuya lista algunas pueden descartarse por: 1) no ser fundamentales al problema, 2) estar sujetas a grandes errores de medición y 3) por resultar colineales con otra u otras independientes (Neter *et al*. 1983), 4) (también ley de Pareto).

Anotan estos mismos autores que variando entonces los propósitos de la regresión entre describir, controlar y predecir, seguramente ningún subconjunto del original podrá ser el mejor para todos los usos, por lo cual las ayudas en el proceso de selección no deberán ser de uso mecánico, sino que deben servir de base para el juicio del investigador. Además de las imperfecciones obvias de los datos, que podrían desde luego rectificarse antes de elaborar un modelo, se dan otras por el modelo mismo, pero no será posible dar un procedimiento estándar a seguir en todos los casos (Daniel & Wood 1980).

La determinación de una ecuación dada, basada en un subconjunto del conjunto de variables originales, involucra tres aspectos básicos:
1- Criterio para seleccionar y analizar dicho subconjunto;
2- Estimación de la ecuación final, y,
3- Técnica computacional, como se aborda en Mateo & Miguel (1979), Chatterjee & Price (1977) y Draper & Smith (1966).

Acá se entremezclarán buscando dar una idea lo más sucinta del tema que debe dejar de lado ciertas complejidades que escapan al alcance de este texto.

Como principio se debe hacer una revisión de los datos ya que se pueden presentar tres tipos de defectos, algunos sólo visualizables luego de los análisis como: - medidas remotas y su influencia en el modelo, -conjunto de datos que una vez colectados no representan a la población por cambios drásticos en ésta, y - datos agrupados o encasillados en variables que cambian poco frente a otras que lo hacen más frecuentemente (Daniel & Wood 1980).

2.15.1 Número posible de modelos de regresión.

Para el modelo estudiado con $(p-1)$ variables independientes son posibles $2^{(p-1)}$ modelos, desde el modelo nulo (sin variables $\hat{Y} = \overline{Y}$), hasta el modelo pleno, con los subconjuntos

configurables así, $(p-1)$ con una sola variable, $(p\text{-}1)(p\text{-}2)/2$ con 2 variables, y otros que se buscan con diversas estrategias que van desde el proceso de seleccionar sobre todas las posibles regresiones, hasta procesos con adición o retirada de variables como los procesos paso a paso: "*stepwise*".

2.15.2 Formulación del problema.

Partiendo del modelo analizado:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i \quad (2.195)$$

es posible omitir algunas variables especialmente si $p-1$ es grande. Se denotarán las variables retenidas por $X_1, X_2, \ldots, X_q$ y las omitidas por $X_{q+1}, X_{q+2}, \ldots, X_{p-1}$, para lo cual se debe examinar el efecto de supresión de variables bajo dos condiciones: 1ª, que todos los $\beta_j\ (j = 0, \ldots, p-1)$ sean $\beta_j \neq 0$; 2ª, que los $\beta_j\ (j = 0, \ldots, q)$ sean $\beta_j \neq 0$ y los $\beta_j\ (j = q+1, \ldots, p-1)$ sean ceros, para analizar el efecto de dejar variables en el modelo que podrían quitarse y viceversa.

2.15.3 Consecuencias de retirar variables.

Sean $b_0^*, b_1^*, \ldots, b_{p-1}^*$ los estimados de los β_j para el modelo completo y $b_0, b_1, \ldots, b_q$ cuando se le hacen recortes. Asimismo sean $\hat{Y}^*$ y $\hat{Y}$ bajo las mismas connotaciones. Sucede entonces que $b_0, b_1, \ldots, b_q$ son estimadores sesgados de $\beta_0, \beta_1, \ldots, \beta_q$ excepto que los restantes $\beta_j\ (j = q+1, \ldots, p-1)$ sean cero o las variables sean ortogonales al conjunto de las que se retiran. Los estimados $b_0^*, b_1^*, \ldots, b_q^*$ * tienen menor precisión que $b_0, b_1, \ldots, b_q$, o sea que:

$$\sigma^2(b_j^*) \geq \sigma^2(b_j), \quad (2.196)$$

lo que significa que las varianzas en los modelos reducidos no resultan mayores que las del modelo completo, y que el retiro de variables nunca incrementará las varianzas de los coeficientes estimados retenidos (Chartterjee & Price 1977). Estos mismos autores reportan que dado que los b_j son sesgados y los b_j^* no, la mejor visualización de la precisión de los estimados se obtiene comparando los cuadrados medios del error de los $b_j\left(MSE(b_j)\right)$ contra las varianzas de los b_j^*:

$$MSE(b_j) << \sigma^2(b_j^*) \quad (2.197)$$

lo cual es cierto únicamente si las variables retenidas tienen coeficientes de regresión más pequeños en magnitud que la desviación estándar de ellos, o sea que $b_j < s(b_j)$. Además el estimado de σ^2 del modelo reducido también resulta casi siempre sesgado. Similar es el efecto de la predicción: $\sigma^2(Y) \leq \sigma^2(Y^*)$. El precio por
omitir variables es el sesgo introducido, de forma que sí:

$$MSE(b_{sesg}) \leq \sigma^2(b_{insesg}) \quad (2.198)$$

la ganancia en la precisión no es compensada por el cuadrado de los sesgos.

Pero sucede también que al retener variables extrañas al modelo o poco importantes, de tal manera que los $\beta_j \approx 0,\ \sigma(\beta_j) << \sigma(b_j\ estimados)$, se dan pérdidas de precisión tanto en estimación como en predicción (Chatterjee & Price 1977). Se expondrán los criterios más simples al respecto, haciendo notar que la prueba de F es ampliamente usada para juzgar la necesidad de adicionar o retirar variables.

De todos modos, es mucha la complejidad en este campo, desde criterios simples para juzgar el fenómeno como la robustez hasta sofisticadas teorías al respecto, pero recalcando que algunos procesos, como los "stepwise" son riesgosos cuando hay multicolinealidad, pues el efecto que se mide es marginal (Daniel & Wood 1980).

Es posible dividir estos métodos en dos grupos, uno con base en todos los modelos de regresión y otro con base en rutinas "*stepwise*".

2.15.3.1. Análisis sobre todos los posibles modelos. Existen varias propuestas.

2.15.3.1.1 Criterio del R_p^2. Consiste en un análisis de los coeficientes de determinación múltiple R^2 ya estudiados, con el fin de seleccionar uno o varios subconjuntos de variables independientes. R_p^2 indica que hay p parámetros o $p-1$ variables predictoras en la ecuación. Puesto que:

$$R_p^2 = \frac{SSR_p}{SSTO} = 1 - \frac{SSE_p}{SSTO} \quad (2.199)$$

tiene denominador $(SSTO)$ constante para todas las posibles regresiones, R_p^2 varia inversamente con SSE_p, que no puede aumentar con el incremento del número de variables, por lo cual R_p^2 será un máximo cuando las $(p-1)$ variables potenciales estén presentes en el modelo.

No se trata de maximizar R_p^2, sino de la búsqueda de unos puntos en los cuales la ganancia obtenida por encima de ellos no sea importante.

Ejemplo, se quiere evaluar el turno económico de plantaciones forestales con base en tres parámetros: $X_1 = edad\ de\ la\ plantación$, $X_2 = índice\ de\ manejo$, que incluye índice de sitio a los 5 años y, $X_3 = indicador\ de\ facilidad\ de\ manejo\ silvicultural$ que, incluye pendiente, cercanía a vías, estado del terreno y especie, previamente definido por otro tipo de estudios, para lo cual se tomaron los datos de la Tabla 2.1.

El siguiente ejemplo se motrará resuelto usando R, para lo cual se deberá bajar la librería *leaps*

Tabla 2.1. Datos tomados para evaluar el turno económico de plantaciones forestales con base en tres parámetros

Y	X1	X2	X3

p1	20.4	7.4	5.7	21.6
p2	8.0	5.8	4.6	19.1
p3	20.1	6.0	6.7	25.0
p4	21.5	7.5	6.6	35.6
p5	56.9	11.2	7.6	55.9
p6	5.8	8.7	4.5	25.2
p7	11.8	2.6	7.4	20.5
p8	7.5	3.6	2.8	13.0
p9	31.1	4.5	7.3	30.5
p10	19.1	3.4	7.7	14.8
p11	29.5	5.8	7.0	32.5
p12	57.4	11.2	7.5	55.9
p13	48.3	6.8	8.2	64.0
p14	34.0	3.7	5.4	16.1
p15	9.5	5.0	4.8	18.3
p16	7.1	3.1	6.5	7.4
p17	17.2	5.3	5.5	30.2

```
library(leaps)
turno<-read.table("clipboard")
attach(turno)
names(turno)
[1] "Y"  "X1" "X2" "X3"
```

El modelo siguiente es para mostrarle la pseudoparadoja estadística, una prueba de F significativa y ningún valor de t aparentemente significativo (*p-values>0.05*):

```
mode<-lm(Y~X1*X2*X3)#Modelo con todas las interacciones
summary(mode2)

Call:
lm(formula = Y ~ X1 * X2 * X3)

Residuals:
   Min     1Q Median     3Q    Max
-5.900 -4.425 -1.014  2.466 18.217

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.6064   104.9705  -0.044    0.966
X1           -1.4448    21.3528  -0.068    0.948
X2            0.8748    14.9886   0.058    0.955
X3            2.9287     5.4929   0.533    0.607
X1:X2         0.6600     3.0257   0.218    0.832
X1:X3        -0.4936     0.9762  -0.506    0.625
X2:X3        -0.3576     0.7557  -0.473    0.647
X1:X2:X3      0.0658     0.1321   0.498    0.630

Residual standard error: 7.881 on 9 degrees of freedom
Multiple R-squared:  0.8779,    Adjusted R-squared:  0.7829
F-statistic: 9.243 on 7 and 9 DF,  p-value: 0.001732
```

Dado que ninguna de las interacciones fue significativa, se corre el modelo con solo las variables simples originales. El modelo general inicial muestra que en apariencia X3 es la única variable significativa al modelar en R:

```
 mode1<-lm(Y~X1+X2+X3)

summary(mode1)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-12.283  -5.195  -1.685   3.651  22.286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.2070    12.6487  -1.360   0.1968
X1            0.5863     1.4374   0.408   0.6900
X2            2.9563     2.0667   1.430   0.1762
X3            0.6700     0.2743   2.442   0.0296 *
```

En *leaps*, presentaremos algunos de los siguientes criterios presentados para la eliminación de variables.

```
leaps(x=turno[,2:4], y=turno[,1], names=names(turno)[2:4], method="r2")
$which
     X1    X2    X3
1 FALSE FALSE  TRUE
1 FALSE  TRUE FALSE
1  TRUE FALSE FALSE
2 FALSE  TRUE  TRUE
2  TRUE FALSE  TRUE
2  TRUE  TRUE FALSE
3  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "X1"          "X2"          "X3"

$size
[1] 2 2 2 3 3 3 4#número de variables del modelo

$r2
[1] 0.7279941 0.4071647 0.3865116 0.7624868 0.7285751 0.6578783 0.7654885
```

El primer modelo solo tiene a X3 como VI. Con ella sola el R2= 0.728.

```
leaps( x=turno[,2:4], y=turno[,1], names=names(turno)[2:4], method="adjr2")
$which
     X1    X2    X3
1 FALSE FALSE  TRUE
1 FALSE  TRUE FALSE
1  TRUE FALSE FALSE
2 FALSE  TRUE  TRUE
2  TRUE FALSE  TRUE
2  TRUE  TRUE FALSE
3  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "X1"          "X2"          "X3"

$size
[1] 2 2 2 3 3 3 4

$adjr2
[1] 0.7098604 0.3676423 0.3456124 0.7285563 0.6898001 0.6090038 0.7113704
```

En la Tabla 2.2 se dan los diversos valores obtenidos de este y otros parámetros que serán revisados. En estos modelos se observa que X_3 sola casi explica el 72%/76,3% de la reducción de la variación total y se observan pocos incrementos para regresión con 2 y 3 variables después de que X_3 esté presente. Prácticamente con X_2 y X_3 se obtiene casi el mismo R^2 que con X_1, X_2 y X_3 y además se nota que X_1 a pesar de tener un R^2

mayor que X_2 no aparece en el modelo. En las figuras respectivas se llamará $X_1 = A,\ X_2 = B,\ X_3 = C$.

Tabla 2.2. Cálculos de diversos criterios de selección de variables

X´s	p	R2p	SSEp	gl	MSEp	Cp	R2aj	PRESS
0	1	0	4578	16	286.1			
X1	2	0.38651	2808	15	187.2	21.0083	0.34561	4447.21
X2	2	0.40716	2714	15	180.9	19.8635	0.36764	3496.04
X3	**2**	**0.72799**	**1245**	**15**	**83**	**2.07848**	**0.70986**	**1627.51**
X1, X2	3	0.65788	1566	14	111.9	7.9653	0.609	2573.93
X1, X3	3	0.72858	1242	14	88.7	4.04627	0.6898	2169.01
X2, X3	3	0.76249	1087	14	77.7	2.1664	0.72856	1566.78
X1,X2, X3	4	0.76549	1073.5	13	82.6	4.0000	0.71137	1927.14

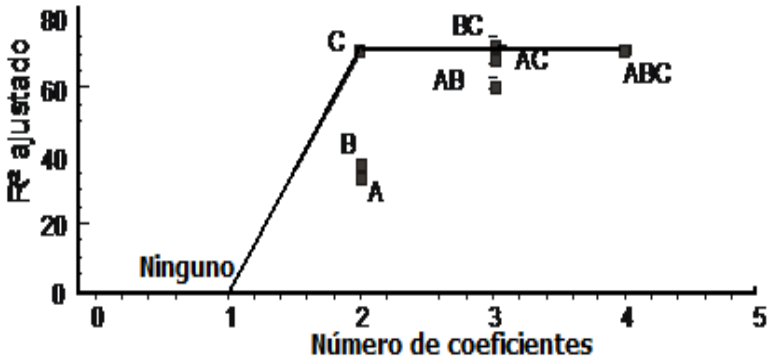


Figura 2.28. Gráfica del criterio R_a^2 .

2.15.3.1.2 Criterio R_2 **ajustado o** MSE_p **.** En vista de que el R_p^2 no toma en cuenta el número de parámetros en el modelo y que es un valor que alcanza su máximo con todas las variables, se propone el R_a^2 definido por (Neter *et al*. 1983) como:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} = 1 - \frac{MSE}{\frac{SSTO}{n-1}} \quad (2.200)$$

R_a^2 aumenta cuando MSE decrece ya que $SSTO/(n-1)$ es un número fijo. Este Indicador resulta equivalente al llamado MSE_p, cuyos valores mínimos se pueden incrementar a medida que p lo hace, especialmente cuando la reducción en SSE_p llega a ser tan pequeña que no alcanza a compensar la pérdida de un grado de libertad adicional. Este criterio (method=r2adj,anterior) permite ubicar el conjunto de variables que minimizan MSE_p o que se sitúan muy próximos a este valor. Se observa que el conjunto X_2, X_3 alcanza el menor valor incluso por debajo de X_1, X_2, X_3. Fig2.29.

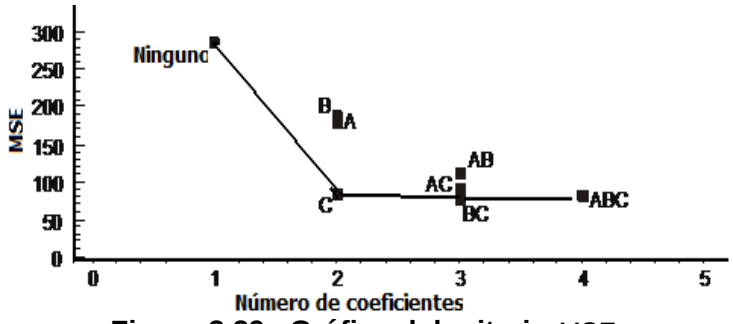


Figura 2.29. Gráfica del criterio *MSE*.

2.15.3.1.3 Criterio del C_p de Mallows. Para comparación de ecuaciones alternativas, se hace necesario un criterio de bondad de ajuste. Daniel & Wood (1980) recomiendan el concepto de Error Cuadrado Total o C_p de $Malllows$, el cual mide la suma de sesgos cuadrados más los errores cuadrados aleatorios de Y^2 en todas las observaciones. El C_p con n observaciones usando una ecuación ajustada con p parámetros es:

$$C_p = \sum_{i=1}^{n}\left(\text{sesgos}\right)^2 + \sum_{i=1}^{n}\left(\text{errores aleatorios}\right)^2 = \sum_{i=1}^{n}\left(V_{ei} - V_{eqi}\right)^2 + \sum_{i=1}^{n} Var\left(Y_{pi}\right) \quad (2.201)$$

en donde: $V_{ei} =$ valor esperado de Y en la ecuación verdadera, $V_{eq} = \beta_0 + \sum \beta_j X_{ij}$, valor esperado con la ecuación obtenida, $\left(V_{ei} - V_{eqi}\right) =$ sesgo de la observación i, $p = k+1$ cuando existe intercepto; pero $p = k$ cuando no lo hay, siendo p el número de variables en el modelo. Sean:

$$\left.\begin{aligned} SSB_p &= \sum_{i=1}^{n}\left(V_{ei} - V_{eqi}\right) \ \text{Suma de cuadrados de los sesgos} \\ \frac{C_p}{\sigma^2} &= STSE = \frac{SSB_p}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} Var\left(Y_{pi}\right) \end{aligned}\right\} \quad (2.202)$$

en que $STSE =$ error total cuadrático estandarizado o C_p/σ^2. Se recuerda que: $\frac{(n-1)s^2}{\sigma^2} = \chi^2$; o sea que toda $\frac{SS\ldots}{\sigma^2}$ se comporta igual y también por el presupuesto de homocedasticidad:

$$\sum_{i=1}^{n} Var\left(Y_{ij}\right) = p\sigma^2 \therefore STSE = \frac{SSB_p}{\sigma^2} + p \quad (2.203)$$

de donde la esperanza de la suma de cuadrados residual SSE_p para un modelo con p términos, salidos de un ANAVA, es:

$$\left.\begin{aligned} &E\left(SSE_p\right) = SSB_p + (n-p)\sigma^2 \quad \therefore \quad SSB_p = E\left(SSE_p\right) - (n-p)\sigma^2 \\ &\rightarrow STSE = \frac{E\left(SSE_p\right)}{\sigma^2} - \frac{(n-p)\sigma^2}{\sigma^2} + p = \frac{E\left(SSE_p\right)}{\sigma^2} - n + 2p; \end{aligned}\right\} \quad (2.204)$$

Pero, con un buen estimador de la σ^2, como la s^2, C_p se vuelve un buen estimador de la $STSE$. Neter *et al*. 1983, Chatterjee & Price 1977 anotan que la mejor estimación de σ^2,

es la obtenida con el modelo completo o sea con $p-1$ variables independientes, lo que equivale a aceptar que el modelo tiene el conjunto de variables bien elegidas y por tanto de (2.204):

$$C_p = \frac{E\left(SSE_p\right)}{s^2} - n + 2p \quad (2.205)$$

Cuando una ecuación tiene un buen ajuste entonces la ecuación verdadera y el modelo estimado tienen $E\left(V_{ei} - V_{eqi}\right) = 0$, con lo cual:

$$E\left(SSE_p\right) = 0 + \left(n-p\right)\sigma^2 \text{ y } C_p = \frac{0 + \left(n-p\right)\sigma^2}{s^2} - n + 2p = p \rightarrow E\left(C_p\right) = p \quad (2.206)$$

Si no hay sesgos, al cancelarse s^2 con σ^2 de (2.206).

Un criterio para utilizar este concepto en forma gráfica, usa la línea de 45° como referencia (a veces los gráficos no lo parecen, por conceptos de escala) de modo que los modelos más cercanos a la recta presentarán el mejor ajuste, lo cual hipotéticamente se puede mostrar así: Para el ejemplo propuesto el mejor estimador de $\sigma = 83.45$ es el del modelo completo. Fig 2.30.

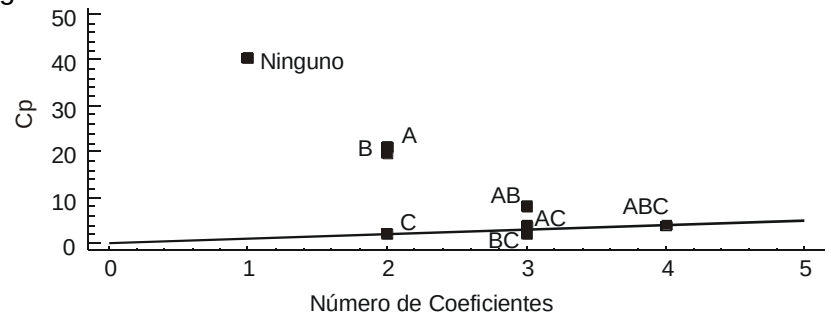


Figura 2.30. Gráfico del criterio del Cp de Mallows.

Sin embargo cuando el modelo tiene muchas variables sin poder explicatorio, o sea con $\beta_k \approx 0$, la estimación de σ^2 resulta muy grande y la pérdida en grados de libertad para el divisor no será balanceada por una reducción en la suma de cuadrados del error. Si σ^2 es muy grande C_p resulta muy pequeño y hay que buscar un mejor estimativo de σ^2 pues de lo contrario se produce un uso de limitada confiabilidad.

Al usar el criterio del C_p se busca identificar los subconjuntos de variables para los cuales este es pequeño o lo más cercanos a p. Los que tienen C_p pequeño tienen un error total medio cuadrático pequeño pero pueden tomar grandes sesgos y cuando C_p es cercano a p los sesgos del modelo resultan pequeños. Los puntos que quedan por encima de la línea tienen mayor sesgo, pero los más cercanos a ella tienen menor valor total del error. Entonces como anotan Daniel & Wood (1980) la adición de términos puede reducir sesgos al modelo, pero al costo de incrementar la varianza total de predicción para los n puntos y en consecuencia la varianza promedia por punto.

Si se necesita una ecuación para interpolación, en el espacio de los datos, es posible eliminar unos pocos términos para obtener ecuaciones más simples, en otras palabras, se

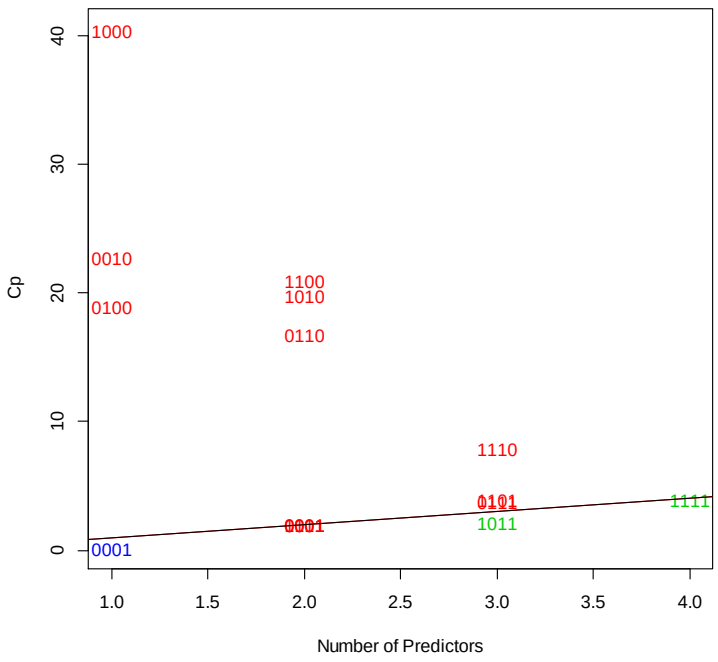
justificará aceptar algún sesgo con el fin de obtener un error promedio más bajo de predicción.

En R existen algunas librerías para calcular o visualizar el Cp en los mejores modelos.

Posibilidad 1: cálculo y visualización *library (wle)*

cp<-mle.cp(Y~X1+X2+X3) # que # que nos da el mejor modelo y su Cp

```
> cp

Call:
mle.cp(formula = Y ~ X1 + X2 + X3)

Mallows Cp:
     (Intercept) X1 X2 X3     cp
[1,]           0  0  0  1 0.2072 #modelo con solo X3 pero sin intrecepto
[2,]           1  0  1  1 2.1660 #modelo con intercepto X2 y X3
[3,]           1  1  1  1 4.0000 #modelo pleno
Printed the first  3  best models

cp$cp
      (Intercept) X1 X2 X3         cp
 [1,]           1  0  0  0 40.4343705
  ….
  [8,]           0  0  0  1  0.2071987
 [9,]           1  0  0  1  2.0784761
[10,]           0  1  0  1  2.0797329
…
 [15,]           1  1  1  1  4.0000000

plot(cp)
```



`cp$cp` # da las soluciones, plot(cp) # los grafica y pinta en azul la solución.

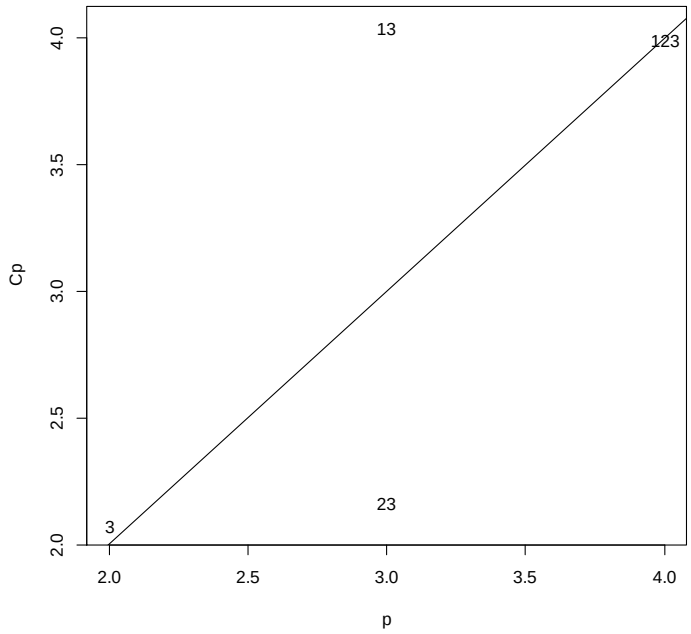
Posibilidad 2: cálculo en los r mejores modelos. library(leaps)

```
ajuste.leaps<-leaps(x=turno[,-1],y=turno[,1],method="Cp")
ajuste.leaps
$which
```

```
      1     2     3
1 FALSE FALSE  TRUE
1 FALSE  TRUE FALSE
1  TRUE FALSE FALSE
2 FALSE  TRUE  TRUE
2  TRUE FALSE  TRUE
2  TRUE  TRUE FALSE
3  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"

$size
[1] 2 2 2 3 3 3 4

$Cp
[1]  2.078476 19.863452 21.008341  2.166396  4.046270  7.965299  4.000000

cbind(ajuste.leaps$which,p=ajuste.leaps$size,Cp=ajuste.leaps$Cp)
  1 2 3 p        Cp
1 0 0 1 2  2.078476
1 0 1 0 2 19.863452
1 1 0 0 2 21.008341
2 0 1 1 3  2.166396
2 1 0 1 3  4.046270
2 1 1 0 3  7.965299
3 1 1 1 4  4.000000
```

Posibilidad 3: visualización library(faraway)

```
library(faraway)#Cp con línea de 45 grados
Cpplot(ajuste.leaps)
```



2.15.3.1.4 Sumas de cuadrados de la predicción: PRESS. Otra forma de examinar habilidades predictivas entre varias regresiones, especialmente para muestras pequeñas es el *PRESS* un acrónimo para predicción de sumas de cuadrados que da una indicación de la capacidad predictiva del modelo (Allen *et al*. 1973, citado por Green 1983).

Se calcula por remoción de una de las n observaciones de los datos y estimación de los coeficientes de regresión con los $n-1$ restantes datos. Los valores de las variables independientes de las observaciones removidas se insertan y calculan con el modelo obtenido y se estima la variable dependiente. Las n diferencias entre los valores estimados y observados al cuadrado, se suman y se obtiene el:

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 \quad (2.207),$$

en que $y_1 =$ la i-*ésima* observación en la variable dependiente, $Y_{(i)} =$ el estimador de $E(Y_i)$ excluida la *i*-ésima observación, y n tamaño de muestra. Este estadístico efectivamente analiza cada modelo en n datos puntuales independientes. Por fortuna se puede simplificar la obtención de él así:

$$PRESS = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - Q_i} \right)^2 \quad (2.208)$$

$\hat{y}_i =$ estimador de $E(Y_i)$ incluyendo n observaciones,

$$\mathbf{Q}_i = \mathbf{x}_i' \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{x}_i \quad (2.209)$$

$\mathbf{x}_i =$ vector $(p*1)$ de variables independientes correspondiente al *i*-ésimo dato puntual, $\mathbf{X} =$ matriz $(n*p)$ de variables independientes y $Y_i =$ *i*-ésima observación de la variable dependiente.

En 2.209, volvemos a la matriz *hat* ya vista por lo cual para llegar al *PRESSp* para cada submodelo. Escribimos en R, la función:

```
PRESS<-sum((modelo$residuals/(1-hatvalues(modelo)))^2)
```

Por ejemplo, para el modelo mode1:

```
mode1<-lm(Y~X3)
PRESS1<- sum((mode1$residuals/(1-hatvalues(mode1)))^2)
[1] 1627.505
```

A diferencia de otros modelos el *PRESS* no constituye una opción ventajosa en sí mismo para modelos de muchos coeficientes. El modelo que obtenga el menor *PRESS* deberá ser el que mejor prediga, aunque se comporte en un orden diferente a otros estadísticos.

2.15.4 Procesos para selección de variables por pasos.

Hasta ahora se ha asumido que las variables fueron bien escogidas de antemano, pero se dan situaciones en que el conjunto de variables explicatorias no fue predeterminado, por lo cual seleccionarlas hace parte del análisis.
Como se vio hay varias aproximaciones siendo las principales, la vista, con *todos los posibles modelos de regresión* (que pueden volverse ineficientes) y, los que podríamos llamar *automáticos*, útiles especialmente con un número alto de variables independientes, que no evalúen todos los posibles modelos, sino una secuencia de ellos en que, a cada paso, se adicionan y/o eliminan variables conocidos como "*stepwise*", con procesos equivalentes en términos a la reducción de la suma de cuadrados del error, correlación parcial y al estadístico F^* (Neter *et al*. 1983), etc.

Con $p-1$ variables explicadoras se evalúan por lo menos p modelos en contraste con los 2^{p-1} necesarios en los métodos anteriores, ubicándose en dos categorías básicas, los procesos hacia adelante (*Forward*) y la eliminación hacia atrás (*Backward*), y además combinaciones de éstos.

2.15.4.1 Procesos hacia delante, adición de variables. Se empieza con la ecuación sin variables y se le adiciona la variable de más alto R^2 simple con la dependiente. Si el β_k respectivo difiere de cero significativamente se le retiene y se busca una segunda variable: la de más alto R^2 con Y cuando se ha retenido la primera y ajustado de acuerdo con su efecto. Se prueba el β_k respectivo. Si este es significativo se busca una variable en forma similar. El proceso termina cuando la última variable adicionada tiene un coeficiente de regresión insignificante o han entrado todas las variables al modelo.

La significación del β_k de esta última variable se hace con una prueba de t computada en la última ecuación, generalmente un valor de t cortado por lo bajo para analizar el coeficiente de las nuevas variables entradas. El proceso recorre el conjunto pleno de variables y propicia $p-1$ ecuaciones posibles.

2.15.4.2 Proceso por eliminación de variables. Comienza con el modelo completo y sucesivamente elimina variables cada vez, con base en la contribución a la suma de cuadrados del error.

La primera que se elimina es la que menos aporte a la reducción de dicha suma, o sea la de menor razón de t (relación entre el coeficiente de regresión y el error estándar de su β_k). Si todas las razones de t son significantes se retienen las $p-1$ variables. Si hay razones insignificantes las $p-2$ restantes se ajustan de nuevo y se examinan las razones de t hasta terminar un proceso con todas las razones que resultan significantes. Este proceso también ajusta al menos p ecuaciones de regresión.

2.15.4.3 Procesos combinados por pasos. Se parecen más al proceso hacia adelante, pero con la posibilidad de borrar variables en otros pasos. Son similares los conceptos a los dos anteriores, pero se asumen diferentes niveles de significación para la inclusión y exclusión de variables al modelo. En términos de F^*, (Neter *et al*. 1983) describen este proceso, que es también un *stepwise*, con su algoritmo de búsqueda así:

1. Ajuste de modelos de regresión simple para cada una de las $p-1$ variables potenciales X. El estadístico será:

$$F_k^* = \frac{MSR(X_k)}{MSE(X_k)} \quad (2.210)$$

La X con mayor valor de F es candidata a la primera adición si el valor de F^* excede algún valor predeterminado, de lo contrario el modelo no la considera.

2. Se ajustan modelos con dos variables, dada la primera X_j seleccionada en el paso 1, entonces el estadístico F será:

$$F_k^* = \frac{MSR\left(X_k \mid X_j\right)}{MSE\left(X_j,\, X_k\right)} = \left[\frac{b_k}{s\left(b_k\right)}\right]^2 \quad (2.211)$$

para verificar si o no $\beta_k = 0$ cuando X_j y X_k son las variables del modelo. La variable X_k cuyo valor F^* sea más alto y supere un F predeterminado es adicionada, de lo contrario el programa termina.

3. Suponiendo que X_1 es adicionada en el paso anterior, el modelo ahora examina si algunas variables presentes en el modelo deben eliminarse. En el presente caso se analizaría F_j^*:

$$F_j^* = \frac{MSR\left(X_k \setminus X_l\right)}{MSE\left(X_l,\, X_k\right)} \quad (2.212)$$

En posteriores pasos habrá un número de estos F^* para cada una de las variables en el modelo junto a las últimas adicionadas. La variable para la cual este F^* es más pequeño resulta candidata a la supresión, si cae debajo del límite predeterminado de él, de lo contrario será retenida.

Suponiendo que X_j es retenida tal que $\left(X_j,\, X_1\right)$ están en el modelo, la rutina "*stepwise*" de nuevo examina la próxima candidata a la adición, la incorpora si es el caso y analiza el comportamiento de X_j y X_1 para borrarlas o dejarlas, hasta cuando ya no sea posible adicionar más, en cuyo caso termina.

2.15.4.4 Regresión paso a paso (stepwise) en R. Son posibles varias opciones: Paso a paso hacia adelante (*Forward stepwise*), hacia atrás (*backward stepwise*) o una combinación de ambas, en cuyo caso R usa el criterio de información de Akaike (AICp) a cada paso en lugar de los vistos: $AIC = -2 * logVerosimilitud + k \times gl$, que veremos luego.

Forward requiere del ajuste de: un modelo base, con una sola variable predictora y, un modelo completo, con todas las variables predictoras deseadas. Para ajustar el modelo base se escoge una variable que consideraríamos como posible de ser incluida en nuestro modelo final. Con el ejemplo del turno financiero ajustamos algunos de los ya vistos:

```
base<-lm(Y~X3)#Modelo base
modecom<-lm(Y~X1+X2+X3)#Modelo completo

modfw<-step(base, scope = list( upper=modecom, lower=~1), direction = "forward", trace=T)
Start:  AIC=76.99

Y ~ X3

       Df Sum of Sq    RSS    AIC
+ X2    1   157.891 1087.2 76.689
<none>              1245.1 76.994
+ X1    1     2.659 1242.5 78.958

Step:  AIC=76.69
Y ~ X3 + X2

       Df Sum of Sq    RSS    AIC
<none>              1087.2 76.689
+ X1    1     13.74 1073.5 78.473
```

Este resultado significa que el modelo con una sola variable predictora X3 tiene un AIC=76.99, muy parecido al modelo con las variables X3 y X2, AIC=76.69, cuya disminución no justifica este modelo por cuanto al correrlo se detecta multicolinealidad, y al adicionar X1 vuelve y aumenta como se ve:

```
summary(modfw)

Call:
lm(formula = Y ~ X3 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.972  -5.979  -1.424   3.212  21.765

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.1057     9.8032  -1.439 0.172163
X3            0.7576     0.1655   4.576 0.000431 ***
X2            2.6190     1.8368   1.426 0.175818
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.812 on 14 degrees of freedom
Multiple R-squared:  0.7625,    Adjusted R-squared:  0.7286
F-statistic: 22.47 on 2 and 14 DF,  p-value: 4.264e-05
extractAIC(base)
[1]  2.00000 76.99405
```

Por lo anterior el modelo con solo X3 se considera adecuado para la modelación solicitada

Eliminación hacia atrás backward. El comando es más simple, pues inicia con el modelo completo (modecom) como base:

```
modecom<-lm(Y~X1+X2+X3)#Modelo completo

modbw<- step( modecom, direction = "backward", trace=T)#Modelo stepwise-backward
Start:  AIC=78.47
Y ~ X1 + X2 + X3

       Df Sum of Sq    RSS    AIC
- X1    1     13.74 1087.2 76.689
<none>              1073.5 78.473
- X2    1    168.97 1242.5 78.958
- X3    1    492.59 1566.1 82.893

Step:  AIC=76.69
Y ~ X2 + X3

       Df Sum of Sq    RSS    AIC
<none>              1087.2 76.689
- X2    1    157.89 1245.1 76.994
- X3    1   1626.49 2713.7 90.239
```

Se observa que al retirar X1 se llega a un AIC=76.69, en el siguiente paso con X2 y X3 al retirar X2 se llegaría a 76.99. Como ve resultado muy similar al anterior

Proceso combinado adelante-atrás. Se parte del forward y se cambia la dirección:

```
modsaa<-step(base, scope = list( upper=modecom, lower=~1), direction = "both", trace=T)
Start:  AIC=76.99
Y ~ X3

       Df Sum of Sq    RSS    AIC
+ X2    1     157.9 1087.2 76.689
```

```
<none>              1245.1 76.994
+ X1    1       2.7 1242.5 78.958
- X3    1    3332.4 4577.5 97.127

Step:  AIC=76.69
Y ~ X3 + X2

       Df Sum of Sq    RSS    AIC
<none>              1087.2 76.689
- X2    1    157.89 1245.1 76.994
+ X1    1     13.74 1073.5 78.473
- X3    1   1626.49 2713.7 90.239
```

2.15.5 Resumen de procesos de modelación por pasos y graficación en R.

Aunque se reiteran conceptos ya usados se presenta este resumen. Para el criterio de todos los posibles modelos (subconjuntos de variables) el mejor ajuste se basó en algunos criterios como R^2, y también AIC y BIC con los cuales se asignan puntajes que permiten escoger el mejor modelo. Esto lo haremos con la librería *leaps* y la función *regsub()* así:

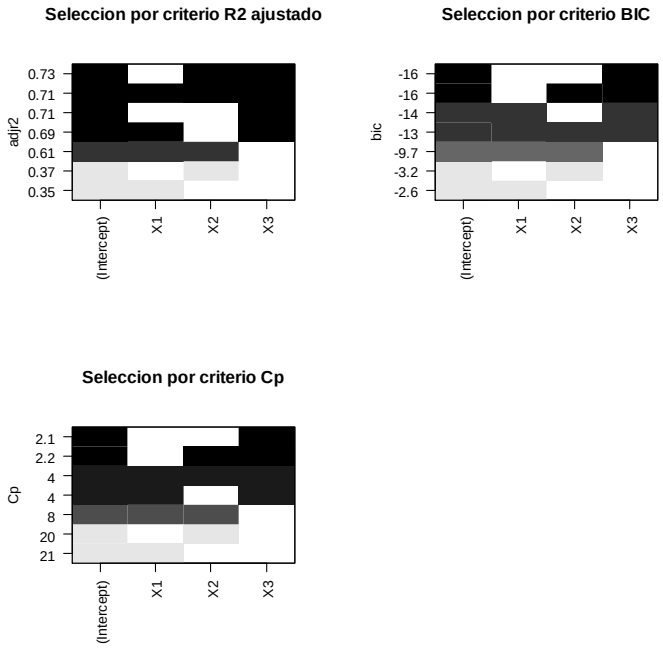
```
library(leaps)
subco<-regsubsets(Y~X1+X2+X3, data=turno,nbest=10)
plot(subco,scale="adjr2",main="Seleccion por criterio R2 ajustado", cex=1.5, cex.main=1.5,
cex.axis=1.5,cex.lab=1.5)
```

En esta gráfica, el valor máximo de la ordenada tiene la clave, el color negro indica cual variable es incluida en el modelo mientras blanco indica cuáles no.

Por ejemplo, con *r2aj*, el modelo con intercepto, X2 y X3, descarta X1.

Con el criterio BIC se muestra que con intercepto y X3 se llega al menor valor BIC.

Con el cp, se muestra que con X2 y X3 se obtienen valores de cp 2.2 y 2.1y, entonces con solo X3 bastaría para un buen modelo.



2.16 Regresión no lineal (RNL).

Cuando ya no es posible encontrar respuestas lineales entre las variables o no es factible por cualquier criterio transformar variables o parámetros, aparece la regresión no lineal, *RNL.* Existen algunos modelos no lineales reconocidos en diversas ramas de la ciencia, especialmente en modelaciones del crecimiento, distribuciones de proporciones, distribuciones diamétricas en dasometría, etc., como los mostrados a continuación:

Funciones sigmoidales, en forma de s

Logistica biparametrica $y = \frac{e^{a+bx}}{1+e^{a+bx}}$

Logistica triparametrica $y = \frac{a}{1+be^{-cx}}$

Funcion de Weibull $y = a - be^{-\left(cx^{d}\right)}$

Curvas arqueadas

Forma de campana $y = a\exp\left(-\left|bx\right|^{2}\right)$

Doble exponenecial $y = ae^{bx} - ce^{-dx}$

Funciones asintóticas

Exponencial asintotica biparametrica $y = a\left(1 - e^{-bx}\right)$

Exponencial asintotica triparametrica $y = a - be^{-cx}$

$\left(2.216\_a\right)$

Una buena definición de RNL es una, función no lineal "conocida, propuesta o supuesta" de parámetros desconocidos, en que los valores esperados de las observaciones Y_i:

$$E(Y_i) = f(X_i, \boldsymbol{\theta})(i = 1, \ldots, n) \quad (2.213)$$

en que $\boldsymbol{\theta}$ es un vector de *p* parámetros desconocidos y las X_i son constantes reales conocidas y, además, las diferencias:

$$y_i - E(y_i) = e_i \text{ deben ser no correlacionadas con} \begin{cases} E(e_i) = 0 \rightarrow \bar{e} = 0 \\ \sigma^2(e_i) = \sigma^2 \end{cases} \quad (2.214)$$

con lo cual se puede establecer un modelo:

$$Y_i = f(X_i, \boldsymbol{\theta}) + e_i \quad (2.215)$$

El método clásico para estimar el vector desconocido $\mathbf{E}(\mathbf{Y}_i)$ usa el principio de los mínimos cuadrados el cual consiste en encontrar un estimado de $\boldsymbol{\theta}$ que minimice la expresión:

$$\Phi(\boldsymbol{\theta}) = \sum_{i=1}^{n} (Y_i - f(x_i, \boldsymbol{\theta}))^2 \quad (2.216)$$

Para iniciar, se toma un ejemplo para tratar de ajustar alturas contra diámetro; con respecto a la solución de una ecuación propuesta como:

$$h = \beta_0 (1 + k_0)^d + \beta_1 (1 + k_1)^d + e \quad (2.217)$$

en la cual d un diámetro de referencia y h una altura para diámetro d y $\beta_0, \beta_1, k_0, k_1$ parámetros desconocidos. Aún si se supiera que aumentos desconocidos de la altura fueran insignificantes y cubiertos por el error e, quedaría la duda si h_d, correlaciona con d para proporcionar estimados de los parámetros $\beta_0, \beta_1, k_0, k_1$. Como no existe ninguna transformación lineal, se ajustará el valor de $\hat{h}$ así:

$$\hat{h}' = \hat{\beta}_0 (1 + \hat{k}_0)^d + \hat{\beta}_1 (1 + \hat{k}_1)^d \quad (2.218)$$

tal que la suma de cuadrados residuales entre los valores observados y estimados sea mínima:

$$Q = (\hat{\beta}_0, \hat{\beta}_1, \hat{k}_0, \hat{k}_1) = \sum_{i=1}^{n} \left[h - \left\{ \hat{h}' = \hat{\beta}_0 (1 + k_0)^d + \hat{\beta}_1 (1 + \hat{k}_1)^d \right\} \right]^2 \quad (2.219)$$

que siguiendo el proceso normal, se hace derivando parcialmente con respecto a los parámetros $\beta_0, \beta_1, k_0, k_1$, e igualando a cero:

$$\left.\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= \sum_{i=1}^{n} \left[h - \hat{b}_0 (1 + \hat{k}_0)^d - \hat{b}_1 (1 + \hat{k}_1)^d \right] \left[-2(1 + \hat{k}_0)^d \right] = 0 \\ \frac{\partial Q}{\partial \beta_1} &= \sum_{i=1}^{n} \left[h - \hat{b}_0 (1 + \hat{k}_0)^d - \hat{b}_1 (1 + \hat{k}_1)^d \right] \left[-2(1 + \hat{k}_1)^d \right] = 0 \\ \frac{\partial Q}{\partial k_0} &= \sum_{i=1}^{n} \left[h - \hat{b}_0 (1 + \hat{k}_0)^d - \hat{b}_1 (1 + \hat{k}_1)^d \right] \left[-2\hat{b}_0 d (1 + \hat{k}_0)^{d-1} \right] = 0 \\ \frac{\partial Q}{\partial k_1} &= \sum_{i=1}^{n} \left[h - \hat{b}_0 (1 + \hat{k}_0)^d - \hat{b}_1 (1 + \hat{k}_1)^d \right] \left[-2\hat{b}_1 d (1 + \hat{k}_1)^{d-1} \right] = 0 \end{aligned}\right\} \quad (2.220)$$

proceso de cuatro ecuaciones con cuatro incógnitas, no lineales, además difícil de resolver, aun con múltiples soluciones y con un gran trabajo de cálculo, transformaciones de

variables y de pronto con multicolinealidad. Para este caso particular se acude a los procesos de aproximaciones lineales sucesivas y concretamente similar a la de la figura 2.1.

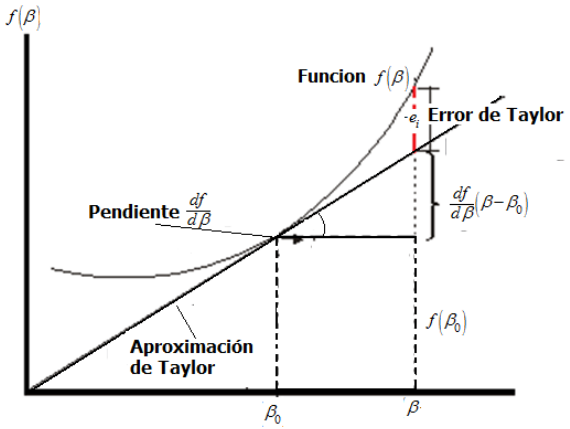


Figura 2.1 Proceso de Linealización de Taylor.

Esta aproximación es conocida como la linealización de Taylor (con k variables y p parámetros), en el cual se propone un punto de partida β_0 para β ; de la cual se aprecia que:

$$f(\beta) = f(\beta_0) + \frac{df}{d\beta}(\beta - \beta_0) + e_i \quad (2.221)$$

Para generalizar se trabajará el modelo:

$$Y = f(X_1, X_2, \ldots, X_k;\ \beta_1, \beta_2, \ldots, \beta_p) + e \quad (2.222)$$

en que $e_t =$ error de Taylor. Se trata de acercar paulatinamente β a β_0 para minimizar este error.

Supuesta, como en este caso, o conocida la forma de dicha ecuación, se tienen datos observados Y , $X_1, X_2, \ldots X_k$ y se procederá a estimar $\beta_1, \beta_2, \ldots, \beta_p$, por medio de unas conjeturas iniciales: $\beta_{10}, \beta_{20}, \ldots, \beta_{p0}$, aplicando en las vecindades de la conjetura, la linealización de Taylor, que constituye una buena aproximación:

$$Y = f(X_1, X_2, \ldots, X_k;\ \beta_1, \beta_2, \ldots, \beta_p) = f(X_1, X_2, \ldots, X_k;\ \beta_{10}, \beta_{20}, \ldots, \beta_{p0}) + \frac{\partial f}{\partial \beta_{10}}(\beta_1 - \beta_{10}) + \frac{\partial f}{\partial \beta_{20}}(\beta_2 - \beta_{20}) + \ldots + \frac{\partial f}{\partial \beta_{p0}}(\beta_p - \beta_{p0}) + e_i \quad (2.223)$$

cuyas derivadas parciales $\partial f / \partial \beta_i$ se evalúan en $\beta_{10}, \beta_{20}, \ldots, \beta_{p0}$. Se sustituye (2.223) en (2.222) y se reúnen los errores aleatorios e_t en un término ε_t con lo cual se tendrá:

$$Y = f(\beta_{10}, \beta_{20}, \ldots, \beta_{p0}) + \frac{\partial f}{\partial \beta_{10}}(\beta_1 - \beta_{10}) + \ldots + \frac{\partial f}{\partial \beta_{p0}}(\beta_p - \beta_{p0}) + \boldsymbol{\varepsilon} \quad (2.224)$$

que se puede reescribir como:

$$Y - f(\beta_{10}, \beta_{20}, \ldots, \beta_{p0}) + \frac{\partial f}{\partial \beta_{10}}(\beta_{10}) + \ldots + \frac{\partial f}{\partial \beta_{p0}}(\beta_{p0}) = (\beta_{10})\frac{\partial f}{\partial \beta_{10}} + \ldots + (\beta_{p0})\frac{\partial f}{\partial \beta_{p0}} + \boldsymbol{\varepsilon} \quad (2.225)$$

cuyo término izquierdo se constituye de valores observados de Y y de los cálculos de f y, sus derivadas en los valores de ensayo, con lo cual se podría asimilar este lado con un valor $Y_{nueva} = Y_n$. Asimismo en el lado derecho, las derivadas parciales se podrán denotar como X_{n1}, X_{n2},…, X_{np}, con lo cual se tendrá:

$$Y_n = \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_p X_{np} + \mathcal{E} \quad (2.226)$$

ecuación linealizada que permite estimar β_1, β_2,…, β_p. Estos estimados resultan notablemente mejores que los supuestos iniciales y se seguirán usando en vez de los viejos, en un siguiente paso, repitiendo el proceso completo, hasta lograr la convergencia, que no siempre da el mínimo absoluto, sino que puede converger a mínimos locales (usualmente lo hace) o aun no converger.

Con respecto a la convergencia Hartley, citado por Orozco & Magidin (1976) habla de ciertas condiciones como que, tanto la función como la primera y segunda derivadas con respecto a **θ** deberán ser continuas para todo X, y que no haya degradación en el rango de la matriz de coeficientes de los sistemas de ecuaciones configurados.

Ejemplo: suponga que desea ajustar el modelo (subjetivo) $y = aX^b$ para los datos cuyo primer par cartesiano fuera (7, 2), con valores de partida a=1, b=2. Se buscan sus derivadas:

$\frac{\partial y}{\partial a}\left(aX^b\right) = X^b$; $\frac{\partial y}{\partial b}\left(aX^b\right) = aX^b \ln X$

Entonces: $y_{1n} = 2 - 1\times 7^2 + 7^2 + 1\times 7^2 \ln 7$; $x_{1n} = 1x^2 = 1\times 7^2$; $x_{2n} = 2\times 1\times 7^2 \times \ln 7$

Etc.

Neter *et al*. (1983) dividen los modelos en lineales, intrínsecamente lineales y no linealizables por métodos ordinarios, aunque algunos pueden ser de ambas categorías finales, como el modelo exponencial:

$$Y_i = \left[\beta_0 e^{\beta_1 X_i}\right]\mathcal{E}_i \quad (2.227)$$

que es no lineal en los parámetros β_0 y β_1, pero que puede linealizarse por transformación logarítmica como:

$$\ln Y_i = \beta_0 + \beta_1 X_i + \mathcal{E}_i \quad (2.228)$$

para quedar intrínsecamente lineal. Cuando esto es posible, se hace aún más imprescindible el estudio de la aptitud del modelo ya que podría presentar los $\mathcal{E}_i$ no normalmente distribuidos. Pero sí, el modelo anterior se hubiera escrito:

$$Y_i = \beta_0 e^{\beta_1 X_i} + \mathcal{E}_i \quad (2.229)$$

con los términos del error aditivos, se saldría del contexto de los intrínsecamente linealizables. Otros modelos utilizados en Ingeniería Forestal para estudios de crecimiento, curvas de mortalidad y supervivencia, etc., como la curva logística con error aditivo, son no linealizables, ejemplo:

$$Y_i = \frac{\beta_0}{1 + \beta_1 e^{\beta_2 X_i}} + \mathcal{E}_i \quad (2.230)$$

A manera de Ejemplo: Para un inventario forestal de gran área (50.000 Ha) se hizo un estudio para determinar previamente el error estadístico (Y) posible, obtenido de acuerdo

con el número de bloques muestreados (X). Si se permiten errores diferenciales de acuerdo con el tipo de bosques, se desea producir una estimación aproximada del error para un número dado de bloques posibles, Tabla 2.3.

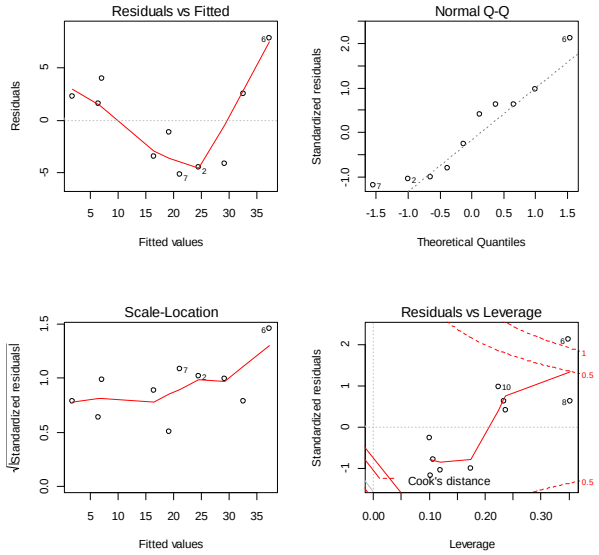
Tabla 2.3. Datos obtenidos de un inventario forestal para determinar el número de bloques.

	X	Y	Mo lin	Mo 1	Mo 2
1	14	35	32.48	31.34	33.748
2	26	20	24.45	17.22	21.193
3	38	13	16.42	11.93	13.308
4	53	8	6.37	8.65	7.44
5	19	25	29.14	23.33	27.8
6	7	45	37.17	61.27	44.27
7	31	16	21.11	14.53	17.458
8	60	4	1.70	7.67	5.672
9	34	18	19.10	13.29	15.541
10	52	11	7.05	8.81	7.734

Se le ajustó en primera instancia el modelo $Y_i = \beta_0 + \beta_1 X_i$, con un $R^2 = 0{,}883$, pero sin ajustes satisfactorios, aunque parecido a situaciones ya comentadas en el análisis de residuales de temas anteriores.

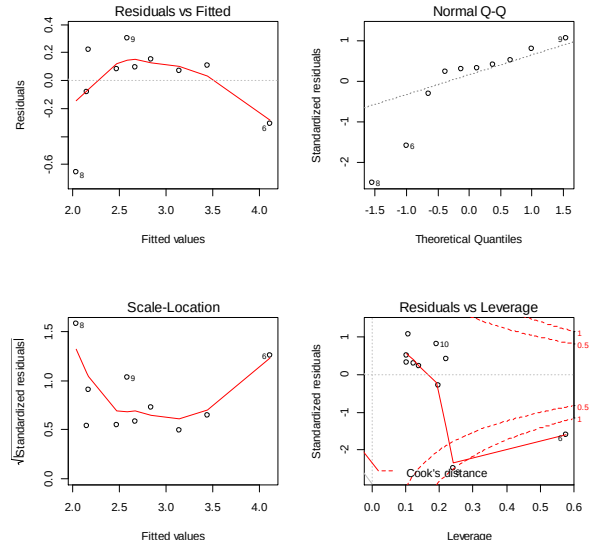
Se utilizó enseguida el modelo: (2.227) y se linealizó como (2.228), con $R^2 = 0{,}83$ y los estimados mostrados en la Tabla, pero el análisis de residuales mostró heterocedasticidad, entonces se ensayó el modelo no lineal (2.229) para el cual se estimaron entonces β_0 y β_1, así como sus respectivos ajustes, línea roja de la Fig 2.2.

```
mod1<-lm(Y~X)
summary(mod1)
Call:
lm(formula = Y ~ X)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.86288    3.22636  12.975 1.18e-06 ***
X           -0.66955    0.08636  -7.753 5.47e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.57 on 8 degrees of freedom
Multiple R-squared:  0.8825,    Adjusted R-squared:  0.8678
F-statistic:  60.1 on 1 and 8 DF,  p-value: 5.471e-05

 names(mod1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```



Los residuales parecen mostrar una forma de bañera.

```
mod2<-lm(log(Y)~log(X)); plot(mod2)
```




```
summary(mod2)
Call:
lm(formula = log(Y) ~ log(X))

Residuals:
     Min       1Q   Median       3Q      Max
-0.65104 -0.04107  0.09119  0.13979  0.30368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9974     0.5067  11.837 2.38e-06 ***
log(X)       -0.9672     0.1490  -6.492  0.00019 ***

Residual standard error: 0.3006 on 8 degrees of freedom
Multiple R-squared:  0.8404,    Adjusted R-squared:  0.8205
F-statistic: 42.14 on 1 and 8 DF,  p-value: 0.0001898

> exp(mod2$fitted.values)
        1         2         3         4         5         6         7         8
31.340164 17.221541 11.930710  8.647932 23.325169 61.271679 14.527425  7.670145
```

```
        9        10
13.285777  8.808734
```

Residuales poco atractivos. Existen muchas otras propuestas y métodos como el del gradiente, propuesto por Cauchi en 1847, el de Gauss-Newton, que usa las series de expansión de Taylor ya vistas, o el de *Marquardt*, considerada una interpolación entre los anteriores citados (Orozco & Magidin 1976).

2.16.1 Método de Gauss Newton.

Por el alcance dado a este texto sólo se expondrá este método que, usa las series de Taylor para aproximar regresiones no lineales con términos lineales, emplea los mínimos cuadrados para estimar los parámetros y se somete al proceso iterativo ya enunciado.

Inicia, como vimos, con unos valores de partida para $\beta_1, \beta_2,\ldots, \beta_{p-1}$, que se denotarán $g_0^{(0)}, g_1^{(0)},\ldots, g_{p-1}^{(0)}$. Los números entre paréntesis definen el número de la interacción: Estos valores de arranque, pueden obtenerse de estudios previos u otras expectativas. Enseguida se aproximan las respuestas medias para $f\left(X_i,\boldsymbol{\theta}\right)$ para las n observaciones por medio de los términos lineales en la serie de Taylor con los valores $g_r^{(0)}$, obteniéndose para la *i*-ésima observación en la iteración cero, similar a la ecuación (2.225).

$$f\left(X_i,\,\boldsymbol{\theta}\right)-f\left(X_i,\,g^{(0)}\right)+\sum_{r=0}^{p-1}\left\{\frac{\partial f\left(X_i,\,\boldsymbol{\theta}\right)}{\partial\beta_r}\right\}_{r=g^{(0)}}\left(\beta_r-g_r^{(0)}\right)\quad(2.231)$$

en la cual: $g^{(0)}$ es el vector de los parámetros de partida. Los términos entre llaves de (2.231) son las mismas derivadas parciales de la función de regresión encontradas en sus ecuaciones normales pero evaluadas para: $\beta_r=g_r^{(0)}$ para $r=0,\,1,\ldots,\,p-1$.

$$\mathbf{g}^{(\boldsymbol{o})}=\begin{vmatrix} g_0^{(0)} \\ g_1^{(0)} \\ \vdots \\ g_{p-1}^{(0)} \end{vmatrix}\quad(2.232)$$

Neter *et al*. (1983) simplifican la notación así:

$$\left.\begin{aligned} f_i^{(0)} &= f\left(X_i,\,\mathbf{g}^{(0)}\right) \\ \beta_r^{(0)} &= \beta_r-\mathbf{g}_r^{(0)} \\ D_{ir}^{(0)} &= \left|\frac{\partial f\left(X_i,\,\boldsymbol{\theta}\right)}{\partial\beta_r}\right|_{r=g^{(0)}} \end{aligned}\right\}\quad(2.233)$$

La aproximación de Taylor para la respuesta media en la *i*-ésima observación se puede notar así:

$$f\left(X_i,\theta\right)\approx f_i^{(0)}+\sum_{r=0}^{p-1}D_{ir}^{(0)}\beta_r\quad(2.234)$$

y la aproximación al modelo de regresión no lineal $\mathbf{Y}_i = f\left(\mathbf{X}_i,\ \boldsymbol{\theta}\right) + \boldsymbol{\mathcal{E}}_i$ es:

$$Y_i \approx f_i^{(0)} + \sum_{r=0}^{p-1} D_{ir}^{(0)} \beta_r^{(0)} + \mathcal{E}_i \quad (2.235)$$

Cuando se traslada $f_i^{(0)}$ a la izquierda y se obtiene la diferencia $Y_i - f_i^{(0)} = Y_i^{(0)}$ se llega al modelo de aproximación lineal, similar a la ecuación (2.226):

$$Y_i^{(0)} \approx \sum_{r=0}^{p-1} D_{ir}^{(0)} \beta_r^{(0)} + \mathcal{E}_i;\ \ (i = 1,\ 2, \ldots,\ n) \quad (2.236)$$

En el cual $Y_i^{(0)} = Y_i - f_i^{(0)}$ se puede notar en la forma matricial conocida $\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{\mathcal{E}}$:

$$\mathbf{Y^{(0)}} \approx \mathbf{D^{(0)}\beta^{(0)}} + \boldsymbol{\mathcal{E}} \quad (2.237)$$

presentable así vectorial y matricialmente:

$$\mathbf{Y}_{\boldsymbol{n*1}}^{\mathbf{(0)}} = \begin{bmatrix} Y_1 - f_1^{(0)} \\ Y_2 - f_2^{(0)} \\ \vdots \\ Y_n - f_n^{(0)} \end{bmatrix};\ \ \mathbf{D}_{\boldsymbol{n*p}}^{\mathbf{(0)}} = \begin{bmatrix} D_{1_0}^{(0)} & D_{1_1}^{(0)} & \ldots & D_{1_{p-1}}^{(0)} \\ D_{2_0}^{(0)} & D_{2_1}^{(0)} & \ldots & D_{2_{p-1}}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ D_{n_0}^{(0)} & D_{n_1}^{(0)} & \ldots & D_{np-1}^{(0)} \end{bmatrix};\ \ \boldsymbol{\beta}_{\boldsymbol{p*1}} = \begin{bmatrix} \beta_0^{(0)} \\ \beta_1^{(0)} \\ \vdots \\ \beta_{p-1}^{(0)} \end{bmatrix} \quad (2.238)$$

Este modelo es similar al $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\mathcal{E}}$, ya evaluado en el que simplemente se reemplaza la matriz $\mathbf{X}$ por la $\mathbf{D}$ de derivadas parciales, con lo cual es posible entonces estimar los parámetros $\boldsymbol{\beta}^{(0)}$ por medio de las ecuaciones normales como se había visto $\boldsymbol{b} = \left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'Y}$:

$$\mathbf{b^{(0)}} = \left(\mathbf{D'^{(0)}D^{(0)}}\right)^{-1}\mathbf{D'^{(0)}Y^{(0)}} \quad (2.239)$$

En que $\mathbf{b}^{(0)}$ es el vector de estimadores mínimo cuadrático de los coeficientes de regresión estimados. Los demás coeficientes de regresión estimados revisados $\mathbf{g}_r^{(1)}$ se obtendrán usando estos estimadores:

$$\mathbf{g_r^{(1)}} = \mathbf{g_R^{(0)}} + \mathbf{b_r^{(0)}} \quad (2.240)$$

en que $\mathbf{g}_r^{(1)}$ denota el estimado revisado de $\boldsymbol{\beta}_r$ al final de la primera iteración. En forma matricial el proceso de revisión se presenta como:

$$\mathbf{g^{(1)}} = \mathbf{g^{(0)}} + \mathbf{b^{(0)}} \quad (2.241)$$

En cada punto se revisa si los ajustes van en la dirección apropiada, de acuerdo con la medida denotada $\mathbf{SSE}^{(0)}$:

$$\mathbf{SSE^{(0)}} = \sum_{i=}^{n}\left[Y_i - f\left(X_i,\ g^{(0)}\right)\right]^2 = \sum_{i=1}^{n}\left(Y_i - f_i^{(0)}\right)^2 \quad (2.242)$$

Al final de la primera iteración los coeficientes de regresión estimados son $\mathbf{g}^{(1)}$ y su respectiva medida de ajuste $\mathbf{SSE}^{(1)}$

$$\mathbf{SSE^{(1)}} = \sum_{i=}^{n}\left[Y_i - f\left(X_i,\ g^{(1)}\right)\right]^2 = \sum_{i=1}^{n}\left(Y_i - f_i^{(1)}\right)^2 \quad (2.243)$$

Si el método resulta efectivo, en la primera iteración $\mathbf{SSE}^{(1)} << \mathbf{SSE}^{(0)}$. Se debe observar que las funciones de regresión no lineales son las usadas para evaluar $\mathbf{SSE}^{(i)}$ y no las aproximaciones lineales de las series de Taylor. El proceso se continua hasta que $\mathbf{SSE}^{(s+1)} - \mathbf{SSE}^{(s)}$ sea insignificante. Los últimos estimadores $\mathbf{g}$ serán los coeficientes de regresión y la última $\boldsymbol{SSE}$ la suma de cuadrados del error.

Como aplicación para el ejemplo dado, se usarán, como puntos de partida, los valores de $\beta_0 = 400$ y $\beta_1 = -0.3$ cercanos a los obtenidos por la transformación logarítmica. El criterio de los mínimos cuadrados requiere en este estado evaluación de la función de regresión no lineal, usando los parámetros $\mathbf{g}_\mathbf{0}^{(\mathbf{0})}$ y $\mathbf{g}_\mathbf{1}^{(\mathbf{1})}$. Para ello $f(X_1, g_0) = f_1^{(0)} = g^{(0)} \exp\left(g_1^{(0)} X_1\right)$, con lo cual se calculará: $SSE(0) = \sum\left(Y_i - f_i(0)\right)^2 = (0,81163)^2 + (2,93216)^2 = 35,042$. Los vectores son entonces:

$$\mathrm{Y}_{10*1}^{(0)} = \begin{bmatrix} \mathrm{y}_1 - \mathrm{f}_1^{(0)} \\ \mathrm{y}_2 - \mathrm{f}_2^{(0)} \\ \mathrm{y}_3 - \mathrm{f}_3^{(0)} \\ \mathrm{y}_4 - \mathrm{f}_4^{(0)} \\ \mathrm{y}_5 - \mathrm{f}_5^{(0)} \\ \mathrm{y}_6 - \mathrm{f}_6^{(0)} \\ \mathrm{y}_7 - \mathrm{f}_7^{(0)} \\ \mathrm{y}_8 - \mathrm{f}_8^{(0)} \\ \mathrm{y}_9 - \mathrm{f}_9^{(0)} \\ \mathrm{y}_{10} - \mathrm{f}_{10}^{(0)} \end{bmatrix} = \begin{bmatrix} 35 - 34,1884 \\ 20 - 21,6691 \\ 13 - 13,7341 \\ 8 - 7,7670 \\ 25 - 28,2724 \\ 45 - 44,6068 \\ 16 - 17,9194 \\ 4 - 5,9529 \\ 18 - 15,9888 \\ 11 - 8,0680 \end{bmatrix} = \begin{bmatrix} +0,81164 \\ -1,66906 \\ -0,73415 \\ +0,23299 \\ -3,27238 \\ +0,39324 \\ -1,91943 \\ -1,95294 \\ +2,01125 \\ +2,93216 \end{bmatrix} \quad (2.244)$$

y la matriz de derivadas en el ítem 0:

$$\mathrm{D}_{10*2}^{(0)} = \begin{bmatrix} \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_1\right) & \mathrm{g}_0^{(0)} \mathrm{x}_1 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_1\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_2\right) & \mathrm{g}_0^{(0)} \mathrm{x}_2 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_2\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_3\right) & \mathrm{g}_0^{(0)} \mathrm{x}_3 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_3\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_4\right) & \mathrm{g}_0^{(0)} \mathrm{x}_4 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_4\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_5\right) & \mathrm{g}_0^{(0)} \mathrm{x}_5 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_5\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_6\right) & \mathrm{g}_0^{(0)} \mathrm{x}_6 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_6\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_7\right) & \mathrm{g}_0^{(0)} \mathrm{x}_7 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_7\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_8\right) & \mathrm{g}_0^{(0)} \mathrm{x}_8 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_8\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_9\right) & \mathrm{g}_0^{(0)} \mathrm{x}_9 \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_9\right) \\ \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_{10}\right) & \mathrm{g}_0^{(0)} \mathrm{x}_{10} \mathrm{EXP}\left(\mathrm{g}_1^{(0)}, \mathrm{x}_{10}\right) \end{bmatrix} = \begin{bmatrix} 0.58743 & 478.6371 \\ 0.37232 & 563.3955 \\ 0.23560 & 521.8976 \\ 0.13345 & 411.6516 \\ 0.48580 & 537.1752 \\ 0.76644 & 312.4273 \\ 0.30789 & 555.5022 \\ 0.10228 & 357.1765 \\ 0.27472 & 543.6175 \\ 0.13862 & 419.5275 \end{bmatrix} \quad (2.245)$$

con ella se obtendrían los estimadores mínimo cuadráticos $b_{(0)}$ de acuerdo con (2.239). Los siguientes son los valores obtenidos por iteración del proceso anterior como el salido de R,. Aparece entonces la regresión ajustada:

$$\hat{Y} = 58,0847 * EXP(-0,0387815 * X) \quad (2.246)$$

que mejora mucho más los estimados anteriores. En R:

```
modelo3<- nls(Y~a*exp(b*X),start=list(a=400.2,b=-0.3))
summary(modelo3)

Formula: Y ~ a * exp(b * X)
```

```
Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 58.07313    2.98648   19.45 5.08e-08 ***
b -0.03877    0.00252  -15.39 3.16e-07 ***

Residual standard error: 2.044 on 8 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 4.113e-06

av<-seq(0,60,0.1)#para graficar la línea X
bv<-predict(modelo3,list(X=av))

plot(X,mod1$fitted.values,main="Modelo lineal y no lineal", col="blue", cex.main=1.5,cex.lab
=1.5, cex.axis=1.5,pch=20)
lines(av,bv,col="red")
points(X,Y,add=T,pch=20,col="black")
```

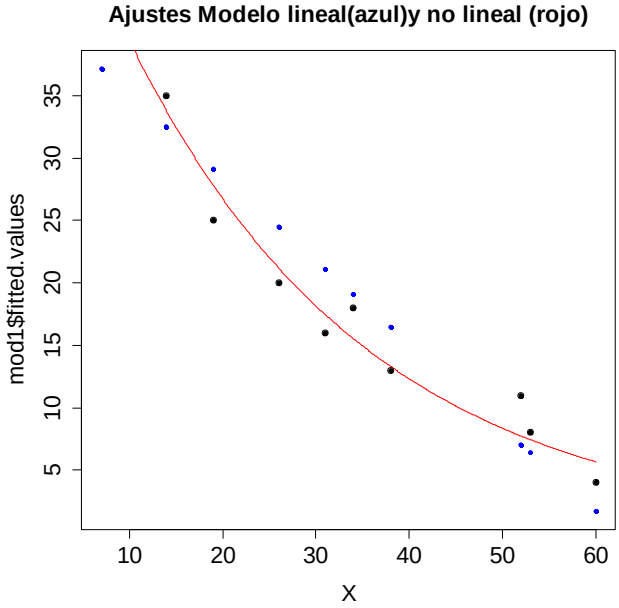


Figura 2.2. Ajuste de la regresión no lineal para error de muestreo contra número de bloques, línea roja.

Visualmente se ve la correspondencia entre los datos originales y el modelo no lineal.

2.16.1.1 Inferencias acerca de los parámetros de regresión. Requieren de un estimado de σ^2 que se comporta similar al visto para la regresión lineal:

$$MSE = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-p} = \frac{\sum_{i=1}^{n}\left(Y_i - f\left(X, \mathbf{g}\right)\right)^2}{n-p} \quad (2.247)$$

con $\mathbf{g} =$ vector final de parámetros estimados.

```
SSE<-sum((cv-Y)^2)
SSE
[1] 33.40923
MSE<-SSE/(8)
 MSE
[1] 4.176153
raizMSE<-MSE^.5
raizMSE
```

```
[1] 2.043564

SSTO<-sum((Y-mean(Y))^2)
> SSTO
[1] 1422.5

SSR
[1] 1389.091

R2<-SSR/SSTO
R2
[1] 0.9765137
```

Se puede a la manera vista para la *RL* buscar unas razones estimadas de *pseudo* t de student, en la forma acostumbrada $t = \frac{g}{s_{\bar{g}}}$, así para a, seria $t_a = 19.38$ y para b: $t_b = 15.24$.

La *MSE* no resulta insesgada, pero esto carece de importancia cuando el tamaño de muestra es grande. Un estimador, para $s^2(g) = MSE(\mathbf{D}'\mathbf{D})^{-1}$, es insesgado, en que $\mathbf{D}$ es la matriz de derivadas parciales evaluada al final, y $s^2(g)$ tiene la misma forma que se había encontrado para $s^2(b) = MSE(\mathbf{X}'\mathbf{X})^{-1}$ en la regresión lineal.

2.16.1.2. Estimación de intervalos para un θ_k .Si no existen problemas de residuales, especialmente si hay independencia y normalidad, el siguiente resultado es cierto:

$$\frac{g_k - \theta_k}{s(g_k)} \approx t_{(1-\alpha/2;n-p)};\ k = 0,\ 1,\ldots,\ p-1 \quad (2.248)$$

de donde saldrían los límites de confianza asintóticos, más ajustados cuando el tamaño de la muestra es grande, para cualquier parámetro θ_k . En la forma usual:

$$\theta_k = g_k \pm t_{(1-\alpha/2;n-p)} * s(g_k) \quad (2.249)$$

En la medida en que los intervalos de algún coeficiente cubran el valor 0, serán candidatos a removerlos del modelo. Es posible, así mismo, encontrar regiones de confianza conjunta pero son difíciles de interpretar cuando el número de parámetros es alto, para lo cual se acude a estimados de Bonferroni (B) de forma:

$$\theta_k = g_k \pm B * s(g_k);\ \text{en que } B = t_{\left(1-\frac{\alpha}{2m};\ n-p\right)} \quad (2.250)$$

siendo m el número de parámetros estimados para la región buscada. Se puede apreciar en la Figura 2.2, el ajuste logrado para los datos del problema presentado.

2.16.1.3 Uso de "fuerza bruta" en R. Existe una librería ***nls2***, que utiliza un algoritmo llamado "*brute force*" empleado para encontrar los valores iniciales, es decir, no se emplea para realizar el modelo, sino para resolver la problemática de dichos valores. Creamos nuestra función objetivo:

fo<-Y~a*exp(b*X)#Funcion objetivo propuesta
Podríamos ir realizando diversas pruebas para encontrar los valores iniciales, a veces sin mucho éxito, por lo cual es mejor elaborar una lista de posibles datos de partida como:

```
part1 <- expand.grid(a = seq(10, 400, len = 10),b=seq(-0.01,-0.9,len=10))
```

```
mod2 <- nls2(fo, start = part1, algorithm = "brute-force")#para generar otros valores de partida, no para modelar,
```

mod2#No importa como modelo, sino como un Nuevo punto de partida:

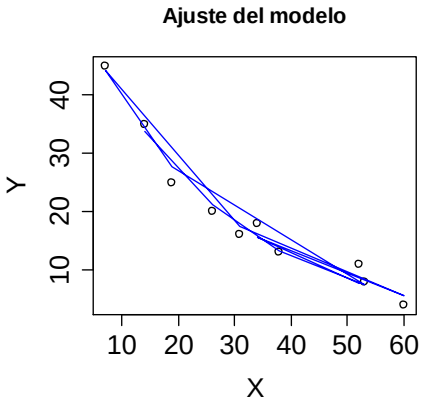
```
Nonlinear regression model
  model: Y ~ a * exp(b * X)
   data: parent.frame()
       a        b
140.0000  -0.1089
 residual sum-of-squares: 1260

Number of iterations to convergence: 100
Achieved convergence tolerance: NA
```

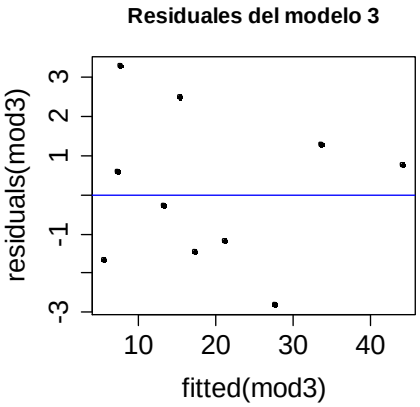
A partir de este, de nuevo con nls2, generamos el modelo siguiente ya explorado:

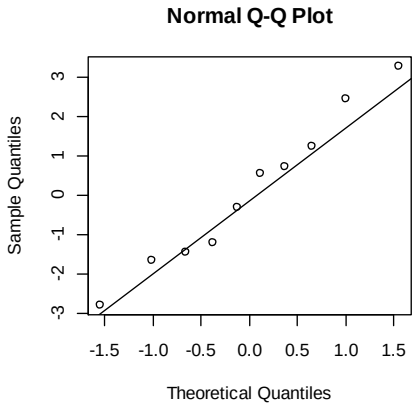
```
mod3 <- nls2(fo, start = mod2)

mod3
Nonlinear regression model
  model: Y ~ a * exp(b * X)
   data: <environment>
       a        b
58.07312 -0.03877
 residual sum-of-squares: 33.41

Number of iterations to convergence: 6
Achieved convergence tolerance: 6.697e-06

summary(mod3)

Formula: Y ~ a * exp(b * X)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 58.07312    2.98648   19.45 5.08e-08 ***
b -0.03877    0.00252  -15.39 3.16e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.044 on 8 degrees of freedom
Number of iterations to convergence: 6
Achieved convergence tolerance: 6.697e-06
```

El algoritmo uso 6 iteraciones, y los parámetros significativos, entonces se realizará un diagnóstico del modelo. Comenzamos por graficar el resultado del modelo:

```
plot(fitted(mod3),residuals(mod3),main="Residuales del modelo 3", cex.lab=1.5, cex.axis=1.5, pch=20)
abline(a=0,b=0,col="blue")
qqnorm(residuals(mod3))
qqline(residuals(mod3))
deviance(mod3)
[1] 33.40923

foo = function(x,a,b){
a*exp(b*x)
}
```








```
shapiro.test(residuals(mod3))#normalidad de residuales

        Shapiro-Wilk normality test

data:  residuals(mod3)
W = 0.97151, p-value = 0.9045

#Test de Leneve
library(car)

leveneTest(Y,as.factor(X))#para homocedasticidad#Prueba no paramétrica

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  9
       0

confint(mod3)#Intervalos e confianza para los parámetros

Waiting for profiling to be done...
         2.5%       97.5%
a 51.37784456 65.36309100
b -0.04497741 -0.03309087
```

2.17. Aplicaciones prácticas de la regresión a ciertos procesos comunes de la medición forestal.

No es necesariamente para este sector, cualquiera tendrá posibilidades de ello, de acuerdo con el entorno. Acá se expone para un sector que necesita un gran dominio en procesos

regresivos, pues muchas variables dependen del comportamiento de otras, como en el caso de los inventarios forestales y otros procesos silviculturales.

2.17.1 Comparación de líneas de regresión.

Hay situaciones en las cuales se pretende juntar en un solo modelo dos ecuaciones que parecen ser muy similares, especialmente para resolver problemas de tamaño de las muestras en cada modelo reducido. Se presentan dos formas de lograrlo, una por un incipiente análisis de covarianza y otra por análisis de significación de variables condicionales (FAO 1980).

2.17.1.1 Análisis de covarianza. Un análisis de covarianza, incipiente, permite comparar pendientes y términos independientes con base en la combinación de los datos para obtención de sumas y productos corregidos por separado para unos datos combinados. Se analiza el ejemplo paso a paso para su mejor comprensión, con base en datos diámetro vs altura comercial obtenidos en un bosque húmedo tropical en dos posiciones fisiográficas diferentes de cativo (*Prioria copaifera*), cinco árboles por posición Tabla 2.4.

Tabla 2.4. Datos de diámetro (dap) vs altura comercial (alco) obtenidos en un bosque húmedo tropical en dos posiciones fisiográficas diferentes de cativo *(Prioria copaifera*), con cálculos posteriores incorporados.

	dap	alco	posfis	1pred/pos	2pred gen	1(y-yes)^2p	2(y-yes)^2p	sumas	
1	30	7.3	coliba	7.00	7.05	0.090	0.062		
2	20	5	coliba	5.03	5.06	0.001	0.004		
3	60	13	coliba	12.93	13.02	0.016	0.049		
4	80	17	coliba	16.88	17.00	0.016	0.000		
5	40	8.7	coliba	8.98	9.04	0.076	0.116	0.198	0.231
6	51	11	vegin	11.20	11.23	0.040	0.053		
7	58	13	vegin	12.58	12.62	0.379	0.333		
8	30	7	vegin	7.05	7.05	0.002	0.003		
9	71	15	vegin	15.15	15.21	0.065	0.096		
10	58	13	vegin	12.58	12.62	0.173	0.142	0.659	0.627
							sumas	**0.857**	**0.858**

```
cativo<-read.table("clipboard")
attach(cativo)
names(cativo)
[1] "dap"    "alco"   "posfis"
mode1<-lm(alco~dap,cativo[which(posfis=="coliba"),])
mode1

Call:
lm(formula = alco ~ dap, data = cativo[which(posfis == "coliba"),])

Coefficients:
(Intercept)          dap
     1.0750       0.1975

mode2<-lm(alco~dap,cativo[which(posfis=="vegin"),])
mode2

Call:
lm(formula = alco ~ dap, data = cativo[which(posfis == "vegin"),])

Coefficients:
(Intercept)          dap
     1.1178       0.1997
```

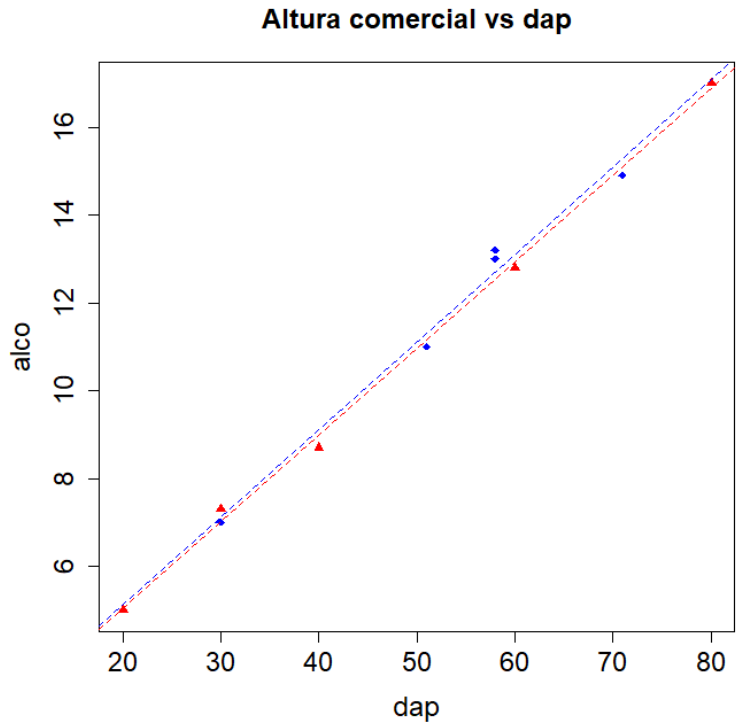
```
mode3<-lm(alco~dap)
mode3

Call:
lm(formula = alco ~ dap)

Coefficients:
(Intercept)          dap
      1.081        0.199
```

Con estos datos se calcularon las respectivas regresiones:

$$\left.\begin{aligned} \text{alco}_1 &= 1.0750 + 0.1975\text{dap}_1 \\ \text{alco}_2 &= 1.1178 + 0.1997\text{dap}_2 \\ \text{alco}_c &= 1.0810 + 0.1990\text{dap}_c \end{aligned}\right\} \quad (2.251)$$

en que $alco_i\,(i=1,\ 2)$ son las regresiones de cada posición y $alco_c$ es el modelo con los datos conjuntos. Para verlo se grafica en R:

```
plot(dap,alco, pch=16+as.numeric(posfis), col=c("red","blue")[as.numeric(posfis)], cex.axis=
1.5, cex.lab=1.5,main="Altura comercial vs dap",cex.main=1.5)
abline(lm(alco[posfis=="coliba"]~dap[posfis=="coliba"]),lty=2,col="red")
abline(lm(alco[posfis=="vegin"]~dap[posfis=="vegin"]),lty=2,col="blue")
abline(lm(alco~dap),lty=1,col="black")
```



Por lo menos gráficamente no pareciera haber diferencias por estratos. Los datos para el ancova incipiente se organizan en la Tabla 2.5

```
colba <- subset(cativo, posfis=="coliba",select=dap:alco)
 colba
  dap alco
1  30  7.3
2  20  5.0
3  60 12.8
4  80 17.0
5  40  8.7
vegai <- subset(cativo, posfis=="vegin",select=dap:alco)
```

```
 vegai
   dap alco
6   51 11.0
7   58 13.2
8   30  7.0
9   71 14.9
10  58 13.0
```

Tabla 2.5. Organización de los datos de la 2.5¡Error! No se encuentra el origen de la referencia. **para el análisis de covarianza**.

Estadisticos	Colinas bajas	Vega inundable	Datos combinados
$\sum H$	50.80	59.10	109.90
$\sum H^2$	606.82	735.25	1342.07
$\sum d$	230.00	268.00	496.00
$\sum d^2$	12900.00	15270.00	26170.00
$\sum dH$	2795.00	3348.50	6143.50
n	5	5	10
SCCH	90.69	36.69	134.27
SCCd	2320.00	905.20	3369.50
SPCdH	485.00	180.74	670.48
b_1	0.1975	0.1997	0.1989
b_0	1.075	1.118	1.081
SCR	90.495	36.088	133.412

El análisis de covarianza se puede efectuar a partir de estos datos, calculando las siguientes cantidades adicionales:

$$1)\ SCC\ b_1 = \sum_{i=1}^{2} SCR_i - \frac{\left(\sum_{i=1}^{2} SPCd_i h_i\right)^2}{\sum_{i=1}^{2} SCCd_i} \quad (2.252)$$

$$2)\ SCC\ b_0 = \sum_{i=1}^{2} (SCCH_i) - \sum_{i=1}^{2} SCR_i \quad (2.253)$$

$$3)\ SCE = SCCH_c - (SCCb_0) - SCCb_1 + SCR_c \quad (2.254)$$

Se usa la notación de sumas de cuadrados corregidas para las distintas combinaciones de variables. Los grados de libertad de las cantidades anteriores son: $SCCH_c = n_c - 1;\ SCE = n_1 + n_2 - 2r;\ SCR_c = 1;\ SCCb_1 = r - 1;\ SCCb_0 = r - 1$, en los cuales: n_1 y $n_2 =$ números de datos de regresiones separadas; $r =$ número de regresiones ajustadas; $n_c = n_1 + n_2 =$ número combinado de datos. Para el ejemplo:

$$SCC\ b_1 = (90.495 + 36.089) - (458.2 + 180.74)^2 / (2320 + 905.2) = 0.0030617$$

$$SCC\ b_0 = (90.692 + 36.688) - (90.495 + 36.089) = 0.7975$$

$$SCE = 134.269 - (0.7975 + 0.003061731 + 133.412) = 0.0569264$$

($SSC =$ sumas de cuadrados corregidos, =SS.. acostumbrados), con las cuales se procede al análisis de covarianza,Tabla 2.6)

Tabla 2.6 Análisis de covarianza.

Fuente de variación	*gl.*	Suma de cuadrados corregidos	Cuadrados medios	Razón de variación
Regr. combinada	1	$SCRC = 133.4116$	133.4115	14061.47*
Entre pendientes	1	$SCCb_1 = 0.00306$	0.00306	0.3227
Entre interceptos	1	$SCCb_0 = 0.79750$	0.79750	84.0559**
Residual	6	$SSE = 0.05693$	0.00948	
Total	9	$SSYC = 134.2690$		

Se comparan las "*F*" para $(1, 6)$ grados de libertad, y se obtienen las siguientes conclusiones: La regresión combinada es altamente significativa, no se debe entonces al azar. La diferencia de pendientes no es significativa, o sea que son muy similares y ello puede ser debido a efectos aleatorios del muestreo en ambos conjuntos de datos. Los términos independientes en cambio difieren significativamente por lo cual se trata de regresiones realmente diferentes a pesar de tener pendientes significativamente iguales. Entonces los datos no podrían combinarse en una sola ecuación sin perder exactitud para cada estrato.

Otra forma de validarse es a través de la función anova del R:

```
anova(modelo1,modelo)
Analysis of Variance Table

Response: h1
          Df Sum Sq Mean Sq F value    Pr(>F)
d1         1 90.495  90.495  1374.6 4.316e-05 ***
Residuals  3  0.197   0.066

Warning message:
In anova.lmlist(object, ...) :
  models with response '"h"' removed because response differs from model 1

anova(modelo2,modelo)
Analysis of Variance Table

Response: h2
          Df Sum Sq Mean Sq F value    Pr(>F)
d2         1 36.088  36.088  180.47 0.0008918 ***
Residuals  3  0.600   0.200

Warning message:
In anova.lmlist(object, ...) :
  models with response '"h"' removed because response differs from model 1
```

que permite las mismas conclusiones del análisis de covarianza de la regresión combinada.

En R existen otras opciones para hacerlo:

```
ancova1<-lm(alco~posfis*dap)
summary(ancova1)

Call:
lm(formula = alco ~ posfis * dap)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39423 -0.23750 -0.06641  0.25625  0.50146

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.075000   0.384440   2.796   0.0313 *
```

```
posfisvegin     0.042764   0.772125   0.055   0.9576
dap             0.197500   0.007569  26.094 2.09e-07 ***
posfisvegin:dap 0.002169   0.014286   0.152   0.8843

Residual standard error: 0.3646 on 6 degrees of freedom
Multiple R-squared:  0.9941,    Adjusted R-squared:  0.9911
F-statistic: 334.8 on 3 and 6 DF,  p-value: 4.572e-07
```

Las interacciones no son significativas

```
anova(ancova1)
Analysis of Variance Table

Response: alco
           Df  Sum Sq Mean Sq F value    Pr(>F)
posfis      1   6.889   6.889  51.836 0.0003634 ***
dap         1 126.580 126.580 952.441 7.685e-08 ***
posfis:dap  1   0.003   0.003   0.023 0.8843258
Residuals   6   0.797   0.133

ancova2<-update(ancova1,~.-posfis:dap)
summary(ancova2)

Call:
lm(formula = alco ~ posfis + dap)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36709 -0.23967 -0.07135  0.25732  0.50832

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.047002   0.312881   3.346   0.0123 *
posfisvegin 0.154374   0.218606   0.706   0.5029
dap         0.198109   0.005954  33.271 5.74e-09 ***

Residual standard error: 0.3382 on 7 degrees of freedom
Multiple R-squared:  0.994,     Adjusted R-squared:  0.9923
F-statistic: 583.6 on 2 and 7 DF,  p-value: 1.636e-08

anova(ancova2)
Analysis of Variance Table

Response: alco
          Df  Sum Sq Mean Sq  F value    Pr(>F)
posfis     1   6.889   6.889   60.244 0.0001105 ***
dap        1 126.580 126.580 1106.931 5.739e-09 ***
Residuals  7   0.800   0.114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

ancova3<-update(ancova2,~.-posfisvegin) #produce el mismo resultado anterior
ancova3<-lm(alco~dap)#Se deja para el modelo conjunto
summary(ancova3)

Call:
lm(formula = alco ~ dap)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34000 -0.22648 -0.05532  0.18755  0.57837

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08084    0.29935   3.611  0.00687 **
dap          0.19898    0.00564  35.280 4.56e-10 ***
```

El modelo conjunto pareciera expresar bien la relación independiente de las

posiciones fisiográficas.

2.17.1.2 Variables condicionales para comparación de curvas. Se llaman también variables binarias o indicadoras, porque son más cualitativas que cuantitativas. Toman generalmente valores discretos y permiten clasificar, jerarquizar o separar atributos. Se llaman también variables ficticias o fantasmas (*dummy*), por su papel a veces intermedio en algunos procesos y binarias por referencia al sistema de oposición. Para evitar problemas de singularidad se sigue como norma que una variable cualitativa con C clases será representada por (C-1) variables indicadoras ejemplo: P_1, P_2,…, P_{c-1}, que sólo pueden tomar valores de cero o uno, dependiendo de si entran o no en el modelo discrecionalmente. Útiles cuando se intenta presentar cierta información que está agrupada por parcelas o por cualquier proceso estratificador, así como las regresiones dentro de ellas.

Se acudirá a un ejemplo de datos tomados con anillos de crecimiento y edad de *Pinus patula*, en tres parcelas de diferentes sitios del país, Piedras Blancas, Cauca
y Santa Rosa, con el objeto de estimar la relación promedio edad-número de anillos con el modelo para todo el país Tabla

	nuan	edad	sitio
1	8	8.8	pb
2	9.2	11.2	pb
3	10.6	12	pb
4	12	14.4	pb
5	14.4	16.8	pb
6	20	21.2	pb
7	4	5	ca
8	6.3	8.2	ca
9	10	12.5	ca
10	13.8	15.2	ca
11	16	17.5	ca
12	10	7.5	sr
13	11	9.3	sr
14	12	12.2	sr
15	14.3	14.8	sr
16	17	15.3	sr

Reorganizamos los datos para el R:

```
nua<-read.table("clipboard")
attach(nua); names(nua); [1] "nuan"  "edad"  "sitio"
```

Lo primero es visualizar estos datos para ver si se aprecian comportamientos diferenciales.

```
plot(nuan,edad,,type="p",pch=16+as.numeric(sitio),col=c("red","blue","green")[as.numeric(sit
io)], cex.axis=1.5, cex.lab=1.5, main="Numero de anillos vs edad", cex.main=1.5)
lines(nuan[sitio=="pb"],edad[sitio=="pb"],col="blue")
lines(nuan[sitio=="ca"],edad[sitio=="ca"],col="red")
lines(nuan[sitio=="sr"],edad[sitio=="sr"],col="green")
text(12,8, "PB",col="blue")
text(12,7, "CA",col="red")
text(12,6, "SR",col="green")
```

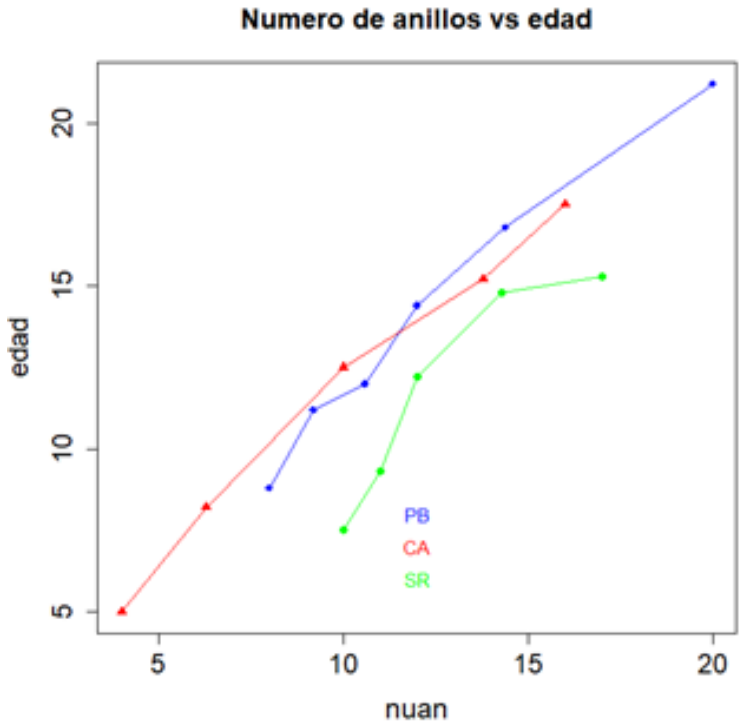


Figura 2.1. Datos de numero de anillos vs edad en 3 sitios del país

Visualmente aparecen comportamientos distintos en los tres sitios, aunque pareciera que entre PB Y CA podría no haber diferencias tajantes. Procederemos a la creación de las variables dummy en R, para lo cual se debe bajar el paquete "*dummies*" y con ello:

library(dummies) #Con la siguiente instrucción procedemos a crear las VD.

dummy(as.character(sitio))#transforma la variable categórica sitio en variables indicadoras

	as.character(sitio)ca	as.character(sitio)pb	as.character(sitio)sr
[1,]	0	1	0
[2,]	0	1	0
[3,]	0	1	0
. . .			
[14,]	0	0	1
[15,]	0	0	1
[16,]	0	0	1

Luego se crea el nuevo dataframe que contiene además de las variables originales las dummy creadas en vez de la categórica "sitio".

```
nuadu<-dummy.data.frame(nua)

nuadu
   nuan edad sitioca sitiopb sitiosr
1   8.0  8.8       0       1       0
2   9.2 11.2       0       1       0
3  10.6 12.0       0       1       0
4  12.0 14.4       0       1       0
5  14.4 16.8       0       1       0
6  20.0 21.2       0       1       0
7   4.0  5.0       1       0       0
8   6.3  8.2       1       0       0
9  10.0 12.5       1       0       0
10 13.8 15.2       1       0       0
11 16.0 17.5       1       0       0
```

```
12 10.0  7.5       0       0       1
13 11.0  9.3       0       0       1
14 12.0 12.2       0       0       1
15 14.3 14.8       0       0       1
16 17.0 15.3       0       0       1
```

Este nuevo dataframe (nuadu) eliminó la variable sitio y ubico los datos de acuerdo con la nueva estructura. Inicialmente se acudió al siguiente modelo cuadrático:

$$E = b_1(NUA) + b_2(NUA)^2 \quad (2.255)$$

```
mo255<-lm(edad~nuan+I(nuan^2))
summary(mo255)
Call:
lm(formula = edad ~ nuan + I(nuan^2))
Residuals:
    Min      1Q  Median      3Q     Max
-3.3466 -0.3527  0.5207  0.8910  1.6957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.728150   2.896793   0.597    0.561
nuan        0.873094   0.496892   1.757    0.102
I(nuan^2)   0.003876   0.020297   0.191    0.852
Residual standard error: 1.653 on 13 degrees of freedom
Multiple R-squared:  0.8679,    Adjusted R-squared:  0.8475
F-statistic: 42.69 on 2 and 13 DF,  p-value: 1.935e-06
```

cuya ecuación es:

$$Edad = 1,72815 + 0,873094(NA) + 0,00387552(NA)^2 \quad (2.256)$$

Que como se aprecia revela multicolinealidad, todos los t no significativos pero la F altamente significativa.

Para ajustar el nuevo modelo se hará uso de dos variables indicadoras p_1 y p_2 : cuando $p_1 = 0,\ p_2 = 0$, se tiene el modelo para Piedras Blancas, cuando $p_1 = 1$, $p_2 = 0$, se obtiene el modelo para el Cauca, si $p_2 = 1,\ p_1 = 0$, para Santa Rosa.

El modelo ajustado fue de la forma:

$$E = a_0 + a_1 p_1 + a_2 p_2 + b_1(NUA) + b_2(NUA)^2 \quad (2.257)$$

Para ajustar el nuevo modelo se hará uso de las dos variables indicadoras ya reseñadas. Los datos entonces se deberán agrupar como se observa en la Figura 2.1 mostrada. que además entrega dos estimados para un modelo general, cuadrático y el combinado con variables *dummy*. El primer modelo ajustado fue de la forma de la ecuación (2.225), estimado1. El siguiente modelo se corrió con variables dummy:

```
p1<-nuadu$sitiopb
p1
 [1] 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
p2<-nuadu$sitioca
p2
 [1] 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0

mo255du<-lm(edad~nuan+I(nuan^2)+p1+p2)
summary(mo255du)

Call:
lm(formula = edad ~ nuan + I(nuan^2) + p1 + p2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41466 -0.55810  0.08822  0.47901  1.46503
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.61936    1.87776  -1.927 0.080127 .
nuan         1.41104    0.29179   4.836 0.000522 ***
I(nuan^2)   -0.01576    0.01168  -1.349 0.204400
p1           2.89594    0.55319   5.235 0.000279 ***
p2           3.05911    0.62783   4.873 0.000493 ***

Residual standard error: 0.8953 on 11 degrees of freedom
Multiple R-squared:  0.9672,    Adjusted R-squared:  0.9553
F-statistic: 81.06 on 4 and 11 DF,  p-value: 4.354e-08
```

Se actualizó al modelo siguiente

```
mo255du1<-update(mo255du,~.-I(nuan^2))
mo255du1<-update(mo255du,~.-I(nuan^2))
summary(mo255du1)

Call:
lm(formula = edad ~ nuan + p1 + p2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3867 -0.6332 -0.2840  0.6633  1.5031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.36969    0.89248  -1.535 0.150795
nuan         1.02564    0.06149  16.680 1.15e-09 ***
p1           2.75265    0.56114   4.905 0.000363 ***
p2           2.77281    0.61074   4.540 0.000678 ***

Residual standard error: 0.9253 on 12 degrees of freedom
Multiple R-squared:  0.9618,    Adjusted R-squared:  0.9522
F-statistic: 100.6 on 3 and 12 DF,  p-value: 9.022e-09
```

que sin más análisis condujo al mejor modelo con todas sus variables significativas y con la menor suma de cuadrados entre estimados y observados de los modelos no multicolineales.

2.17.2 Regresión condicionada.

En la medición forestal es común su uso a causa de la definición de diámetro a la altura del pecho, que impone un intercepto conocido o deseado a los 1.3 m. del piso, o como sucede en inventarios forestales, en el muestreo bifásico bivariado, o cuando se espera que la línea ajustada pase por un punto conocido.

2.17.2.1 Regresión condicionada por el origen. Existe un tipo importante de regresión condicionada a pasar por el origen, por intuición o porque los valores de t del b_0 hacen innecesaria su consideración. Se hace un análisis para la regresión lineal simple, pero es similar para la múltiple:

$$Y_i = \beta_1 X_i + \mathcal{E}_i;\ \ E(Y) = \beta_1 X_i \quad (2.258)$$

Los estimadores mínimo cuadráticos se obtienen como se vio minimizando Q con respecto a $\boldsymbol{\beta}_1$:

$$Q=\sum_{i=1}^{n}\left(y_i-\beta_1 X_i\right)^2 \quad \therefore \quad \frac{\partial Q}{\partial \beta_1}=-2\sum_{i=1}^{n}X_i\left(Y_i-b_1X_i\right)=0 \quad \therefore \quad b_1=\frac{\sum_{i=1}^{n}\left(X_iY_i\right)}{\sum_{i=1}^{n}X_i^2} \quad (2.259)$$

b_1 es estimador de máxima verosimilitud para β_1, de donde $\hat{y}_i=b_1X_i$. El estimador de σ^2 será:

$$MSE=\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\left(n-1\right)}=\frac{\sum_{i=1}^{n}\left(Y_i-b_1X_i\right)^2}{\left(n-1\right)} \quad (2.260)$$

En este tipo de regresiones $\sum e_i=0$ y solamente $\sum X_ie_i=0$.

En R, con el ejemplo de anillos de crecimiento:

```
mode5<-lm(edad~nuan)

summary(mode5)
Call:
lm(formula = edad ~ nuan)
Residuals:
Min 1Q Median 3Q Max
-3.3923 -0.3401 0.5828 0.9311 1.6580

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.2338 1.2542 0.984 0.342
nuan 0.9658 0.1009 9.574 1.6e-07 ***

Residual standard error: 1.595 on 14 degrees of freedom
Multiple R-squared: 0.8675, Adjusted R-squared: 0.858
F-statistic: 91.65 on 1 and 14 DF, p-value: 1.598e-07

modesi<-lm(edad~nuan-1)#el anterior sin intercepto por no ser significativo

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
nuan  1.05994    0.03203   33.09 1.95e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.593 on 15 degrees of freedom
Multiple R-squared:  0.9865,    Adjusted R-squared:  0.9856
F-statistic:  1095 on 1 and 15 DF,  p-value: 1.952e-15
```

Debe además evaluarse su aptitud ya que podría ser no lineal o los términos de la varianza resultar no constantes.

```
anova(modesi)
Analysis of Variance Table
Response: edad
Df Sum Sq Mean Sq F value Pr(>F)
nuan 1 2778.35 2778.35 1094.9 1.952e-15 ***
Residuals 15 38.06 2.54
```

Es muy importante de considerar que el R^2 en este tipo de modelos puede resultar superestimado, cuando los datos de la variable independiente no cubren un amplio rango de valores, al estimarlo solo para el centro de gravedad de unos datos y el origen.

2.17.2.2 Otras regresiones condicionadas. En la misma tónica del anterior, se calculan modelos con paso obligado por una constante, ejemplo por 1.3 (altura del pecho), o como en el proceso de las regresiones con base en desviaciones alrededor de la media, ya

estudiadas, o en el proceso de reparametrización de variables, útil para el estudio de la multicolinealidad, a las cuales se llega por modelación. El tratamiento no difiere en nada del anterior por lo cual se remite a el para los casos específicos. Se acude de nuevo al ejemplo diamtros y alturas trabajado.
En R

```
dialt<-read.table("clipboard")
attach(dialt)
names(dialt)
[1] "d" "h"

mode1<-lm(h-1.3~d -1)
summary(mode1)

Call:
lm(formula = h - 1.3 ~ d - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4042 -0.2393 -0.1776  0.1331  0.5839

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
d   0.1951     0.0019   102.7 3.99e-15 ***

Residual standard error: 0.3188 on 9 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 1.055e+04 on 1 and 9 DF,  p-value: 3.991e-15

mode1$fitted.values
        1         2         3         4         5         6         7         8
 5.853142  3.902094 11.706283 15.608378  7.804189  9.950341 11.316074  5.853142
        9        10
13.852435 11.316074
mode1$fitted.values+1.3
        1         2         3         4         5         6         7         8
 7.153142  5.202094 13.006283 16.908378  9.104189 11.250341 12.616074  7.153142
        9        10
15.152435 12.616074
```

El modelo quedaría entonces como $alt = 1.3 + 0.1951 * d$

2.17.3 Variables indicadoras dependientes.

Existen circunstancias en que las variables dependientes tienen dos soluciones posibles únicamente y se les asignan los valores 0 y 1 de acuerdo con la existencia o no de un determinado suceso, por ejemplo: un árbol muerto o vivo, sano o enfermo, mayor o menor de t años. Como se ve la variable dependiente es dicotómica o de respuesta binaria, un experimento Bernoulli (Neter *et al.* 1983**).**

2.17.3.1 Significado de la función para una variable dependiente tipo Bernoulli. Sea un modelo lineal:

$$Y_i = \beta_0 + \beta_1 X_i + \mathcal{E}_i;\ \ Y_i = (0,1) \therefore E(Y_i) = \beta_0 + \beta_1 X_i \quad (2.261)$$

Se puede establecer la siguiente distribución de probabilidades para cada Y_i

Y_i	Probabilidad

1	$P(Y_i = 1) = p_i$
0	$P(Y_i = 0) = 1 - p_i = q_i$

Al usar la definición de esperanza matemática de una variable y siendo Y una variable que puede asumir los valores $Y_1, Y_2 \ldots, Y_k$, cuyas probabilidades se dan por la función de probabilidades $f(Y_i)$ tal que: $f(Y_i) = P(Y = Y_i)$; $i = 1, 2, \ldots, k$. El valor esperado de Y se define como $E(Y) = \sum Y_i f(Y_i)$, con lo cual: De (2.61):

$$E(Y_i) = p_i + 0(1 - p_i) = p_i = \beta_0 + \beta_1 X_i = p_i \quad (2.262)$$

La respuesta media de $Y_i = E(Y_i)$ es simplemente la probabilidad de que $Y_i = 1$ cuando el nivel de la variable independiente sea X_i. En otras palabras, la respuesta media cuando una $Y_i = 0; 1$, siempre representa una probabilidad de $Y = 1$ para los niveles dados de los X_i. Por ejemplo: sea Y una variable indicadora referida a si un árbol está sano o enfermo, y la variable independiente X_c edad del árbol, el modelo da la probabilidad de que un árbol de determinada edad esté en las condiciones anotadas.

2.17.3.2 Problemas de la función binaria.

1. Los términos del error no se distribuyen normalmente ya que pueden tomar uno de dos valores $\mathcal{E}_i = Y_i - (\beta_0 + \beta_1 X_i)$, así: Cuando $Y_i = 1 \rightarrow \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$, cuando $Y_i = 0 \rightarrow \mathcal{E}_i = 0 - \beta_0 - \beta_1 X_i$.
2. Se presenta heterocedasticidad.

$$\sigma^2(Y_i) = E\left\{\left[Y_i - E(Y_i)\right]^2\right\} = (1 - p_i)^2 p_i + (0 - p_i)^2 (1 - p_i) \quad (2.263)$$

ya que $Y_i = 1$ ó 0 entonces:

$$\sigma^2(Y_i) = p_i(1 - p_i) = p_i q_i = \left[E(Y_i)\right]\left[1 - E(Y_i)\right] \quad (2.264)$$

La varianza de los $\mathcal{E}_i$ es la varianza de Y_i ya que $\mathcal{E}_i = Y_i - p_i$; y p_i es constante, entonces

$$\sigma^2(\mathcal{E}_i) = \left[E(Y_i)\right]\left[1 - E(Y_i)\right] = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i) = p_i q_i \quad (2.265)$$

que muestra que $\sigma^2(\mathcal{E}_i)$ depende del valor tomado por X_i, por lo cual el método de los mínimos cuadrados no resulta ya el más procedente.

Puesto que $E(Y_i)$ representa las probabilidades asociadas para valores de $Y_i = 1$, 0 entonces aparece una fuerte restricción a muchas funciones:

$$0 \le E(Y) = p_i \le 1 \quad (2.266)$$

Por ejemplo: se hizo un estudio de la infección para plantaciones atacadas o no a diversas edades, con datos tomados a 12 árboles al azar, cuyos resultados muestra la Tabla2.3.

	ed	es	estm1	wi	estmpon
1	29	0	0.6412	4.3464	0.6509
2	25	1	0.5589	4.0563	0.5663
3	18	1	0.415	4.1192	0.4181
4	12	0	0.2916	4.8413	0.291
5	6	0	0.1682	7.1482	0.164

6	11	0	0.271	5.0618	0.2699
7	5	0	0.1476	7.9476	0.1428
8	24	0	0.5383	4.0237	0.5451
9	19	0	0.4355	4.0677	0.4392
10	28	1	0.6206	4.2471	0.6298
11	8	1	0.2093	6.0424	0.2064
12	32	1	0.7029	4.7881	0.7145

Por medio de mínimos cuadrados se ajustó la ecuación descrita, dejando la constancia de no ser una solución óptima ya que no se da la normalidad esencial para las pruebas de hipótesis, por ejemplo, ya no sería muy válido un intervalo de confianza para predicciones, problema que se resuelve con propuestas como el análisis *Probit* que se estudiará posteriormente.

```
essa<-read.table("clipboard")
attach(essa)
names(essa)
[1] "ed" "es"

mode1<- lm(es~ed)
summary(mode1)
Call:
lm(formula = es ~ ed)

Residuals:
Min 1Q Median 3Q Max
-0.6412 -0.3276 -0.1579 0.3948 0.7907

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.04479 0.31975 0.140 0.891
ed 0.02056 0.01578 1.303 0.222

Residual standard error: 0.4993 on 10 degrees of freedom
Multiple R-squared: 0.1451, Adjusted R-squared: 0.05964
F-statistic: 1.698 on 1 and 10 DF, p-value: 0.2218
Que se grafica
av<-ed
bv<-mode1$fitted.values
plot(ed,es, pch=20,col="red",main="Relacion estado sanitario vs edad", cex.main=1.5,
cex.axis=1.5,cex.lab=1.5)
lines(av,mode1$fitted.values,type="b",pch=20,col="black")
```

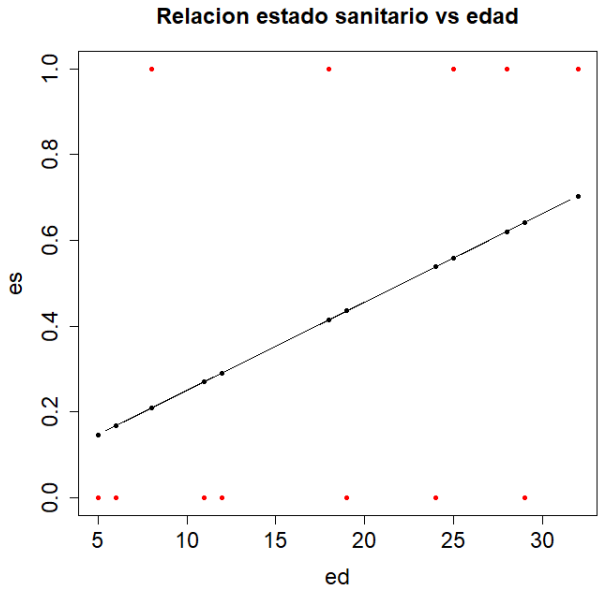
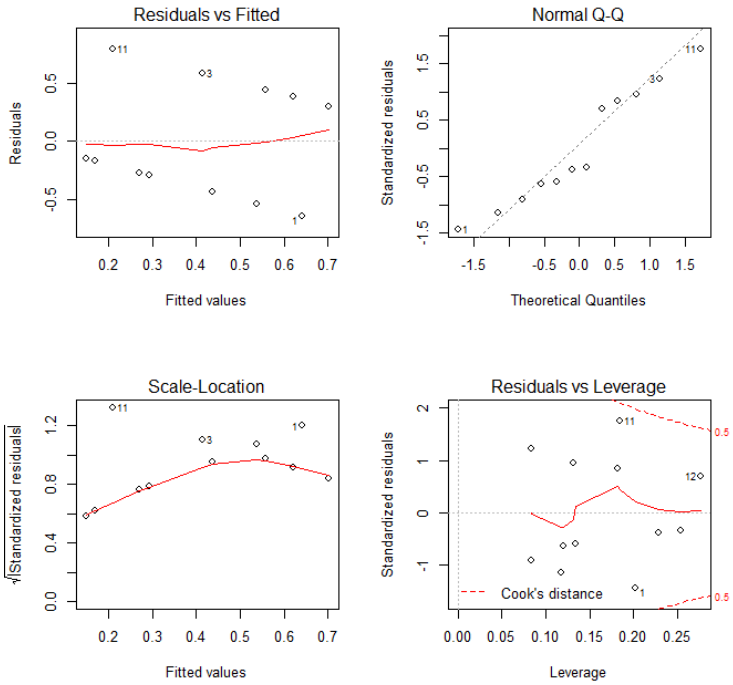


Figura 2.4 Relaciones para estados de infección contra edad.

No obstante, se presenta esta opción como un método susceptible de mejorar como se verá posteriormente, con el ajuste ponderado. Al revisarlo entonces, a pesar de las anotaciones anteriores se nota que la enfermedad avanza con el aumento de la edad. El coeficiente b_1 indicaría que por cada aumento de un año de edad crece la probabilidad de la enfermedad en 0,02, Figura 2.4 Los estadísticos encontrados no son alentadores como se aprecia luego del *plot model*.

```
plot(model)
```



2.17.3.3 Ajuste ponderado. Debido a la heterocedasticidad, se puede acudir a la regresión ponderada con el fin de corregir posiblemente los términos del error, en la misma forma propuesta de factores inversos a la variabilidad. De la ecuación (2.265), entonces:

$$w_i = \frac{1}{p_i q_i} = \frac{1}{p_i(1-p_i)} \quad (2.267)$$

sería una buena aproximación buscada. Como en $p_i q_i$ se involucra $E(Y_i)$ se propone ajustar el modelo como se hizo y estimar los respectivos

$$w_i = \left[\frac{1}{E(Y_i)(1-E(Y_i))} \right] \quad (2.268)$$

como ponderadores lo que se muestra también en la Tabla anterior. Al ajustarse la regresión ponderada sin diferir notablemente de la anterior sí mejora los estimados de $s(b_k)$

.

```
ajusm1<-mode1$fitted.values
> ajusm1
        1         2         3         4         5         6         7         8
0.6411623 0.5589043 0.4149530 0.2915661 0.1681792 0.2710016 0.1476147 0.5383399
        9        10        11        12
0.4355174 0.6205978 0.2093081 0.7028557
> wi=1/(ajusm1*(1-ajusm1))
> modpon <- lm(es ~ ed, weights = wi)
> summary(modpon)

Call:
lm(formula = es ~ ed, weights = wi)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-1.3571 -0.7017 -0.4206  0.7906  1.9509

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03698    0.27876   0.133    0.897
ed           0.02117    0.01474   1.436    0.182

Residual standard error: 1.09 on 10 degrees of freedom
Multiple R-squared:  0.1709,    Adjusted R-squared:  0.08804
F-statistic: 2.062 on 1 and 10 DF,  p-value: 0.1815
```

Esta refinación se podría reiterar ya que se van dando cada vez mínimas variaciones. Al respecto Neter *et al*. (1983) hacen algunos comentarios entre los cuales se resalta el hecho de que el proceso ponderado ajusta mejor cuando el rango del modelo incluye valores de $0.8 < Y_i < 0.2$ ya que de lo contrario las ganancias obtenidas serán escasas puesto que las varianzas no serán lo suficientemente desiguales para justificar el método. Si los Y_i se localizan sólo en el rango de 0.2 a 0.8 es poca la mejoría del modelo. También si la función de respuesta ajustada no cae por debajo de 0 ni por encima de 1 entonces se sugiere un ajuste curvilineal como la función logística. Con este criterio se podría justificar el esfuerzo realizado con este proceso.

2.17.3.4 Función de respuesta logística. Consideraciones tanto técnicas como empíricas sugieren que con variables dependientes binarias se dan muy a menudo respuestas curvilineales como la mostrada en la **Figura 2.2** que es una función con forma sigmoidal y

con asíntotas en 0 y 1 o sea que tiene automáticamente la restricción anotada para Y. Su función es

$$E(Y)=\frac{EXP(\beta_0+\beta_1X)}{1+EXP(\beta_0+\beta_1X)} \quad (2.269)$$

la cual puede ser linealizada fácilmente. Sea $E(Y)=\pi$, ya que la respuesta media de Y es una probabilidad, transformando (2.261) se tiene

$$\left.\begin{aligned} &E(Y)+E(Y)EXP(\beta_0+\beta_1X)=EXP(\beta_0+\beta_1X);\ \ \therefore \\ &E(Y)=e^{(\beta_0+\beta_1X)}-E(Y)e^{(\beta_0+\beta_1X)}=e^{(\beta_0+\beta_1X)}\left(1-E(Y)\right)\ y, \\ &e^{(\beta_0+\beta_1X)}=\frac{E(Y)}{1-E(Y)}\therefore\beta_0+\beta_1X=\ln\left[\frac{E(Y)}{1-E(Y)}\right]=\ln\left[\frac{\pi}{1-\pi}\right] \end{aligned}\right\} \quad (2.270)$$

llamando:

$$\pi'=\ln\left[\frac{\pi}{1-\pi}\right]\rightarrow\pi''=\beta_0+\beta_1X \quad (2.271)$$

llamada la transformación *logit* o logística de la probabilidad π. Por ejemplo para los datos de la tabla 2.4

```
piprim<-log(pporas/(1-pporas))#pi prima o logit
piprim
[1] -1.3862944 -1.1368455 -0.8472979 -0.6190392 -0.2006707 0.0236011
[7] 0.4054651 0.6632942 1.0986123 1.2926773

molog<-lm(piprim~dap)
summary(molog)
Call:
lm(formula = piprim ~ dap)
Residuals:
Min 1Q Median 3Q Max
-0.088301 -0.028887 -0.001246 0.021513 0.095723

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.438345 0.113322 -47.99 3.93e-11 ***
dap 0.153363 0.003195 48.00 3.92e-11 ***

Residual standard error: 0.05804 on 8 degrees of freedom
Multiple R-squared: 0.9965, Adjusted R-squared: 0.9961
F-statistic: 2304 on 1 and 8 DF, p-value: 3.925e-11

exp(piprim)
[1] 0.2500000 0.3208295 0.4285714 0.5384615 0.8181818 1.0238818 1.5000000
[8] 1.9411765 3.0000000 3.6425255
```

Entonces $\pi'=\ln\left[\frac{\pi}{1-\pi}\right]\rightarrow\pi'=-5.4383+.1533*d$

2.17.3.4.1 Ajuste de la función logística. Hace parte de funciones con forma sigmoidal bajo formas bi, tri o tetraparamétrica.

La biparamétrica de forma: $y=\frac{e^{a+bx}}{1+e^{a+bx}}$,

La triparametrica permite la variación de y en una escala $y=\frac{a}{1+be^{-cx}}$, con intercepto $\frac{a}{(1+b)}$, a un valor asintótico y la pendiente inicial medida por c.

La 4-parametrica con asíntotas a izquierda (a) y a derecha (b), escala c con respecto a su punto medio d y con punto de inflexión en $y=a+\frac{b-a}{1+e^{c(d-x)}}$

Su forma puede verse en la figura 2.3.0

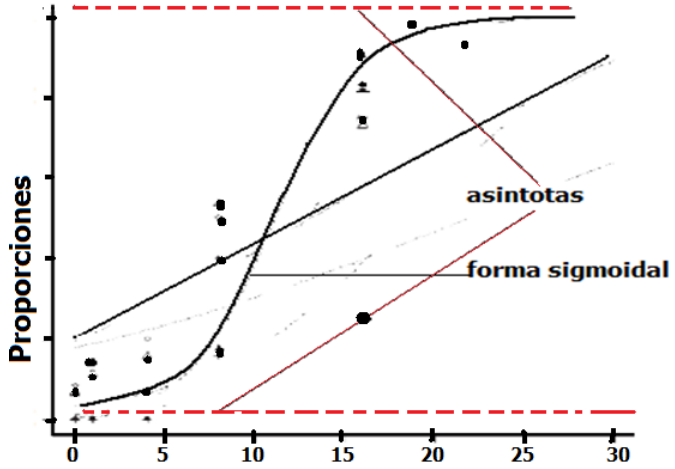


Figura 2.1.0 Forma y elementos de la regresión logística.

Trata de modelar una respuesta categórica, siendo la más común la binaria. El ajuste de (2.271) es relativamente simple cuando se dan observaciones repetidas para cada nivel de la variable X.

En la práctica es como sigue: 1) Se denotan los niveles de $X_1,\ X_2,\ldots,\ X_c$ con su respectiva frecuencia $n_j\ (j=1,\ldots,\ c)=$ número de observaciones por nivel j. Basta con considerar el número de "1" (unos) solamente para cada X en vez de los valores individuales de Y. Sea R_j el número de "1" del nivel X_j, con lo cual la proporción de "1" en ese nivel será $p_j = R_j/n_j$ con lo cual (2.271) se ajusta haciendo:

$$p'_j=\ln\left(\frac{p_j}{1-p_j}\right) \quad (2.272)$$

y usando p'_j como variable dependiente. Se debe anotar que la transformación logística a pesar de linealizar la función respuesta no elimina la heterocedasticidad de los errores por lo cual se puede acudir a mínimos cuadrados ponderados. Si n_j es suficientemente grande la

$$\sigma^2\left(p'_j\right)=\frac{1}{n_j\pi_j\left(1-\pi_j\right)}\rightarrow \textit{estimada por}\ \ s^2\left(p'_j\right)=\frac{1}{n_j p_j\left(1-\pi_j\right)} \quad (2.273)$$

con lo cual los ponderadores estimados serán:

$$\hat{w}_j = n_j p_j \left(1 - p_j\right) \quad (2.274)$$

Una vez obtenida

$$\pi' = b_0 + b_1 X \;\rightarrow\; \hat{\pi} = \frac{EXP\left(\beta_0 + \beta_1 X\right)}{1 + EXP\left(\beta_0 + \beta_1 X\right)} \quad (2.275)$$

se hace la transformación a las variables originales.

Como ejemplo se muestra un inventario en una plantación en la cual cada equipo calificó los árboles que a su juicio se catalogaran como de aserrío con base en el diámetro a la altura del pecho. La muestra seleccionada fue de árboles entre 24 y 44 cm de d y se asumió la función logística para probarla como modelo. Los datos se muestran en la Tabla 2.4

Tabla 2.4 Datos para el ajuste de la regresión logística.

	dap	nuar	aras	pporas	k	=LN(k)	wi=nj*pjqj
1	26	50	10	0.2000	0.2500	-1.3863	8.000
2	28	70	17	0.2429	0.3208	-1.1369	12.871
3	30	100	30	0.3000	0.4286	-0.8473	21.000
4	32	60	21	0.3500	0.5386	-0.6188	13.650
5	34	40	18	0.4500	0.8182	-0.2006	9.900
6	36	85	43	0.5059	1.0239	0.0236	24.247
7	38	90	54	0.6000	1.5000	0.4055	12.600
8	40	50	33	0.6600	1.9412	0.6633	11.220
9	42	80	60	0.7500	3.0000	1.0986	15.000
10	44	65	51	0.7846	3.0000	1.0986	15.000

aras=árboles para aserrío. $k = p_j \big/ \left(1 - p_j\right)$; $n_j =$tamaño muestral (nuar). Veamos los ponderadores

```
sigma2pprij<- 1/(nuar*pporas*(1-pporas))
```

con lo cual los ponderadores estimados serán:

```
wj<-nuar*pporas*(1-pporas)

modelogpon<-lm(piprim~dap,weights =wj)

summary(modelogpon)
Call:
lm(formula = piprim ~ dap, weights = wj)
Weighted Residuals:
Min 1Q Median 3Q Max
-0.29494 -0.10933 0.04102 0.08771 0.35065
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.49052 0.12030 -45.64 5.87e-11 ***
dap 0.15473 0.00338 45.77 5.74e-11 ***

Residual standard error: 0.2152 on 8 degrees of freedom
Multiple R-squared: 0.9962, Adjusted R-squared: 0.9957
F-statistic: 2095 on 1 and 8 DF, p-value: 5.735e-11
```

Una vez obtenida (2.275) se hace la transformación a las variables originales.

```
piest=exp(-5.49052+0.15473*dap)/(1+exp(-5.49052+0.15473*dap))
plot(dap,piest,type="b",pch=20,col="red",cex.main=1.5,cex.axis=1.5,cex.lab=1.5,main="Regresi
on logistica prob aserrío vs dap")
points(dap,pporas,pch=18,col="blue")
text(35,0.22,"valores observados",col="blue")
```

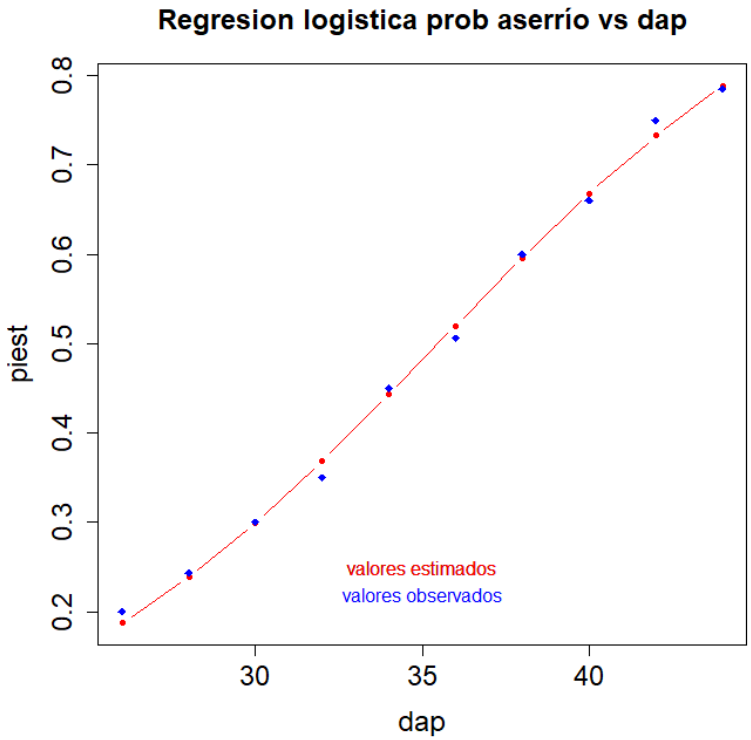


Figura 2.2. Datos y regresión logística para la ¡Error! No se encuentra el origen de la referencia.

Para completar el tema se propone, a pesar de lo anterior, estimar la probabilidad de encontrar árboles de aserrío con $d = 30\ cm$, la cual es: $\pi = -0,68$, la cual transformada a sus variables originales da:

$$\hat{\pi} = \frac{EXP(\beta_0 + \beta_1 X)}{1 + EXP(\beta_0 + \beta_1 X)} = \frac{e^{\pi'}}{1 - e^{\pi'}} = \frac{e^{-0,684067}}{1 + e^{-0,684067}} = 0,3354 \quad (2.276)$$

se espera entonces un 33.54% de árboles de aserrío de 30 cm de d.

2.17.3.2 Otros análisis de la regresión logística en R. Para estudiarla lo más completamente posible se acudirá al ejemplo presentado y al mismo tiempo a la explicación de las pruebas acompañantes. En primer lugar, se acude a la función glm, que se ampliará al estudiar modelos lineales generalizados, luego de calcular:

```
proase<-aras/nuar#Proporcion de árboles de aserrío.
```

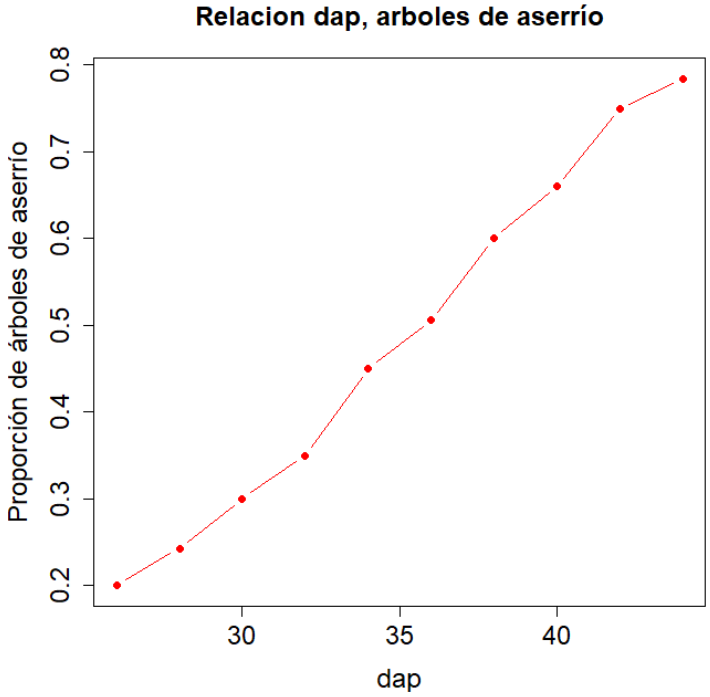
Enseguida graficamos los datos para mirar su comportamiento

```
plot(dap, proase,type="b",col=2,ylab="Proporción de árboles de aserrío", xlab="dap",
cex.main=1.5, cex.lab=1.5, cex.axis=1.5,main="Relacion dap, arboles de
aserrío",pch=16)#pagina siguiente
```

Se corre el modelo logístico:

```
Mod1=glm(aras/nuar~dap, weights=nuar, family=binomial)
summary(Mod1)
Call:
glm(formula = aras/nuar ~ dap, family = binomial, weights = nuar)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.29475  -0.11129   0.04162   0.08847   0.35016
```




```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.49455    0.56127  -9.790   <2e-16 ***
dap          0.15484    0.01575   9.829   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.83207  on 9  degrees of freedom
Residual deviance:   0.37192  on 8  degrees of freedom
AIC: 49.088

Number of Fisher Scoring iterations: 3

> names(Mod1)#Con esta orden ve que le mostraría Mod1$..
 [1] "coefficients"      "residuals"         "fitted.values"
 [4] "effects"           "R"                 "rank"
 [7] "qr"                "family"            "linear.predictors"
[10] "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"
[16] "df.residual"       "df.null"           "y"
[19] "converged"         "boundary"          "model"
[22] "call"              "formula"           "terms"
[25] "data"              "offset"            "control"
[28] "method"            "contrasts"         "xlevels"
```

Por ejemplo

```
Mod1$residuals
           1            2            3            4            5            6
 0.084460889  0.021986783  0.001956883 -0.078607772  0.029256646 -0.056324048
           7            8            9           10
 0.015947346 -0.036099205  0.087851495 -0.025976902
```

que muestra los parámetros, el *ANADE* y una prueba de razón de máxima verosimilitud. El modelo ajustado obtenido fue:

$$p_j = \frac{EXP\left(-5,49455 + 0,154843d\right)}{\left(1 + EXP\left(-5,49455 + 0,154843d\right)\right)}$$

el cual se presenta en la **¡Error! No se encuentra el origen de la referencia.** y analiza a partir del resto de pruebas mostradas.

2.17.3.3 Conversión del archivo en éxitos y fracasos.

Para análisis de predictibilidad es necesario convertir el dataframe original en un archivo de éxitos y fracasos, para ello generamos una variable, llamada fracasos y la incorporamos en un archivo que llamaremos arbas2.

```
frac<-nuar-aras
frac
 [1] 40 53 70 39 22 42 36 17 20 14

arbas2<-cbind(arbas,frac)
```

Creamos una variable y de 1 y 0 represetnado los éxitos y fracasos

```
y=c(rep(1,sum(arbas2$aras)) , rep(0,sum(arbas2$frac)) )

arbas2
   dap nuar aras frac
1   26   50   10   40
2   28   70   17   53
3   30  100   30   70
4   32   60   21   39
5   34   40   18   22
6   36   85   43   42
7   38   90   54   36
8   40   50   33   17
9   42   80   60   20
10  44   65   51   14
```

Creamos una variable y que solo tenga 1 y ceros:

```
y=c(rep(1,sum(arbas2$aras)) , rep(0,sum(arbas2$frac)) )

y#Solo se muestra parte de los valores, por ser 690
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
.
[630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[667] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Creamos otra variable dap2, con el valor correspondiente a cada y

```
dap2=c(rep(arbas2$dap,arbas2$aras),rep(arbas2$dap,arbas2$frac))
> dap2
  [1] 26 26 26 26 26 26 26 26 26 26 28 28 28 28 28 28 28 28 28 28 28 28 28 28
 [26] 28 28 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30
Etc.
```

Creamos un nuevo dataframe: arbas3

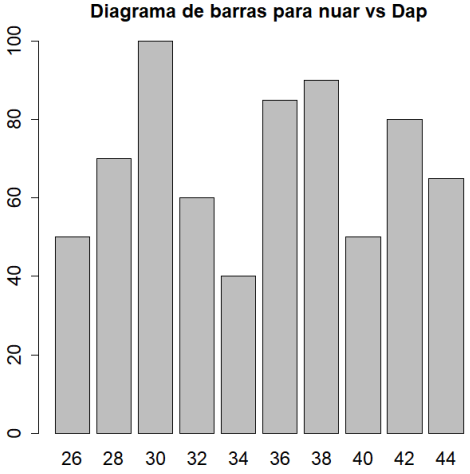
```
arbas3<-as.data.frame(cbind(y,dap2))
 str(arbas3)
'data.frame':   690 obs. of  2 variables:
 $ y   : num  1 1 1 1 1 1 1 1 1 ...
 $ dap2: num  26 26 26 26 26 26 26 26 26 ...
```

Y vemos las proporciones de 1 (aserrío) y 0 (no aseerío), por cada valor de dap(X)

```
with(arbas3, prop.table(table(dap2,y), margin = 1))
    y
dap2         0         1
  26 0.8000000 0.2000000
  28 0.7571429 0.2428571
  30 0.7000000 0.3000000
  32 0.6500000 0.3500000
  34 0.5500000 0.4500000
  36 0.4941176 0.5058824
  38 0.4000000 0.6000000
  40 0.3400000 0.6600000
  42 0.2500000 0.7500000
  44 0.2153846 0.7846154
```

Podemos construir un gráfico de barras para el número de arboles

barplot(table(arbas3$dap),cex.names=1.5,cex.axis=1.5,main="Diagrama de barras para nuar vs Dap",cex.main=1.5)

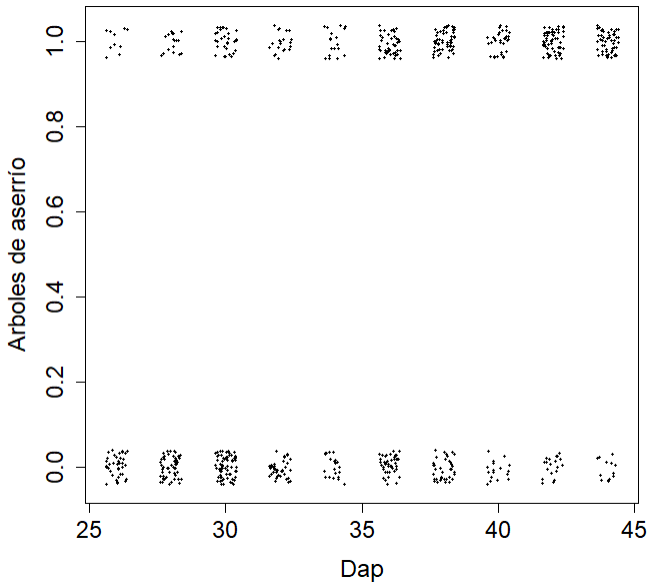


Podemos agrupar en otras categorías diamétricas, por ejmeplo en 6:

```
with(arbas3, prop.table(table(cut(arbas3$dap2, 6), arbas3$y), 1) )

                  0         1
  (26,29] 0.7750000 0.2250000
  (29,32] 0.6812500 0.3187500
  (32,35] 0.5500000 0.4500000
  (35,38] 0.4457143 0.5542857
  (38,41] 0.3400000 0.6600000
  (41,44] 0.2344828 0.7655172
```

Para mirar los datos sin que se traslapen, se usa la función *jitter* que, separa los puntos con un pequeño error aleatorio entre ellos.

```
with(arbas3, plot(jitter(dap2), jitter(y, 0.2), xlab = "Dap", ylab = "Arboles de aserrío",
cex = 0.7, cex.axis = 1.5, cex.lab = 1.5,pch=20))
```



Un modelo de regresión logística como estos, con valores 0,1 es llamado regresión logística binaria

```
modely <- glm(y ~ dap2, data = arbas3, family = binomial)
summary(modely)

Call:
glm(formula = y ~ dap2, family = binomial, data = arbas3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7639  -0.9585  -0.6438   1.0171   1.8308

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.49455    0.56127  -9.790   <2e-16 ***
dap2         0.15484    0.01575   9.829   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 956.17  on 689  degrees of freedom
Residual deviance: 843.71  on 688  degrees of freedom
AIC: 847.71

Number of Fisher Scoring iterations: 4
```

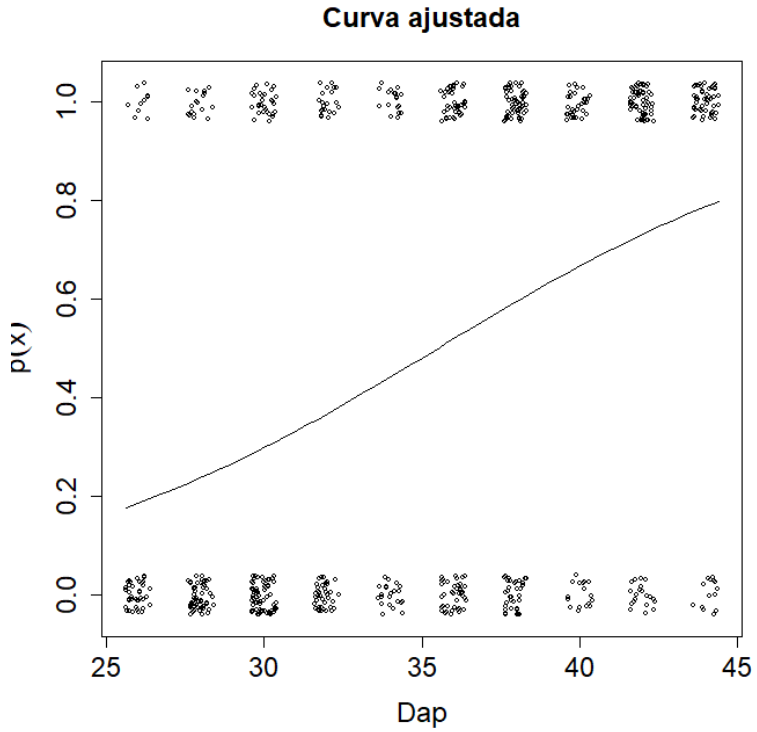
Se aprecia como cambian algunos estadísticos.

Grafica de los datos originales y el modelo ajustado.

```
with(arbas3, plot(jitter(dap2), jitter(y, 0.2), xlab = "Dap", ylab = quote(p(x)), main =
"Curva ajustada", cex = 0.6, cex.axis = 1.5, cex.lab = 1.5, cex.main = 1.5))
```

Le añadimos el modelo ajustado con:

```
curve(1/(1 + exp(-modely$coefficients[1] - modely$coefficients[2] *x)),add=T)
```



2.17.3.4 Análisis de desvianzas. No existe una traducción a la palabra deviance y por contexto se usará aca como desvianza. Cuando el p-value en el análisis de las desvianzas, similar al del ANAVA en los modelos corrientes, es menor a 0,05, la relación entre las variables resulta significativa al 95%, lo cual debe ser corroborado por el análisis de los residuales, el cual debe tener un p-value mayor o igual a 0.05, para indicar que el modelo no es significativamente peor que el mejor posible modelo para estos datos a un 95% de confianza.

Similar ocurre con el porcentaje de desvianza ajustado, de comportamiento similar a un R^2 ajustado, útil para comparar modelos con diferente número de variables independientes, *VI*.

En R:

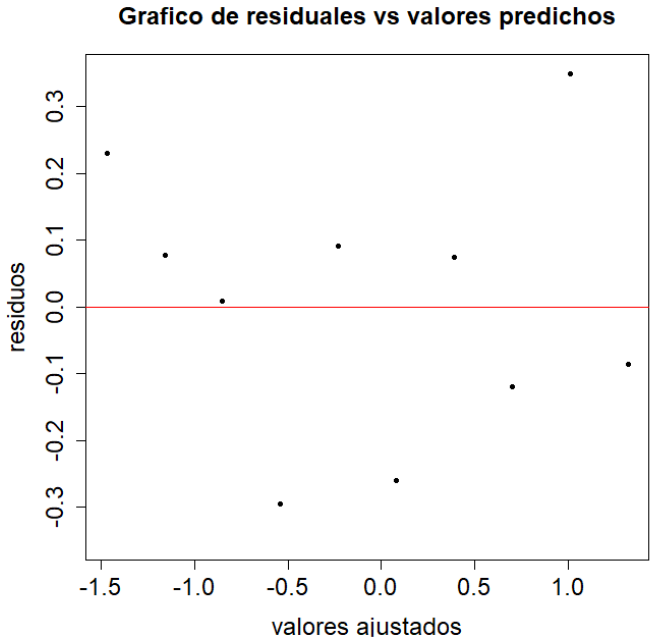
```
anova(Mod1,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: proase
Terms added sequentially (first to last)
Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL 9 112.832
dap 1 112.46 8 0.372 < 2.2e-16 ***para árboles de aserrio.
```

2.17.3.5 Prueba de razón de verosimilitud. Debe tener un p-valuer < 0.05, valores mayores indican que las variables son redundantes.

2.17.3.5 Analisis grafico de residuales. Se construye la siguiente gráfica a partir de:

```
p=predict(Mod1)#valores de los logits ajustados
r<-residuals(Mod1)
```

```
plot(p,r,xlab="valores  ajustados",ylab="residuos",ylim=c(-  1,1)*max(abs(r)),  cex.main=1.5,
cex.lab=1.5,cex.axis=1.5,pch=20,main="Grafico de residuales vs valores predichos")
abline(0,0, col="red")
```



Los residuos no caen fuera de la banda (-3, 3), un buen indicio de ajuste; sin embargo, parece existir cierta algo como un descenso cuando crecen los valores ajustados hasta un valor medio y da ahí nuevamente tienden a crecer.

Tambien podemos hacer

```
shapiro.test(residuals(mod3))#normalidad de residuales
        Shapiro-Wilk normality test

data:  residuals(mod3)
W = 0.97151, p-value = 0.9045

#Test de Leneve
library(car)

leveneTest(Y,as.factor(X))#para homocedasticidad

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  9
   0
```

2.17.3.6 Prueba de bondad de ajuste. Permite establecer si el modelo se ajusta bien a los datos, para lo cual a la manera mostrada para los modelos lineales corrientes, exigen que la $\chi^2 > 0,05$, a un 95% de confianza. **¡Error! No se encuentra el origen de la referencia.**. Esta Tabla así como la siguiente muestran los límites para el modelo, así como el ruido introducido en él.

Tabla 2.3. Pruebas de bondad de ajuste e intervalos de confianza. Prueba de bondad de ajuste por medio de la Chi cuadrada .

			Verdaderos		Falsos	
Clase	Intervalo logístico	n	Observados	Esperados	Observados	Esperados
1	< -0.849254	220	57.0	56.0367	163.0	163.963

2	-0.849254 a 0.0798062	185	82.0	84.0035	103.0	100.996
3	0.0798062 a 0.69918	140	87.0	87.0547	53.0	52.945
4	0.69918 o mayores	145	111.0	109.9050	34.0	35.095
Total	690	337	337.0	353.0000	353.0	

$\chi^2 =$ Chi cuadrada = 0,154916 con 2 gl. Valor P = 0,925466

Intervalos de confianza para los coeficientes estimados.

Parámetros	Estimados	Desviación estándar	Limite inferior	Limite superior
Intercepto	5.49455	0.561254	-6.78881	-4.2003
d	0.154843	0.0157534	0.118516	0.191171

2.17.3.7 Intervalos de confianza para los Antilog β_k. — Muestran el cambio en la proporción de la *VD* por cada cambio de un unidad en las independientes. Acá se observa que al antilog 0.154843 es 1.16748. Al 95% se dan los siguientes:

Parámetros	Estimados	Desviación estándar	Limite inferior
d	1.16748	1.12582	1.21067

2.17.3.8 Análisis de predictibilidad. Muestra el desempeño de las predicciones, en términos de verdaderos y falsos, Tabla 2.11 En ella se da un resumen de la capacidad predictiva del modelo logístico. En primer lugar, se ven las predicciones con base en los datos que lo originaron. Cuando el valor predicho es mayor que el corte o frontera, la respuesta es aceptada como verdadera, si es menor es predicha como falsa. El valor que maximiza los porcentajes totales, como lo hace 0.45, muestra que 71,5134% de todas las respuestas verdaderas fueron correctamente predichas, mientras que el 63.4561% de todas las falsas también lo fueron, para un total del 67.3913%. Luego se invierten esos porcentajes. Este corte entonces se vuelve útil para predecir valores individuales. Se muestran también los valores observados y predichos para los p_j para las variables o factores del modelo, así como sus intervalos de confianza al 95.0% para las proporciones verdaderas. La **¡Error! No se encuentra el origen de la referencia.** muestra el punto de maximización de los cortes cercanos a 0.5.

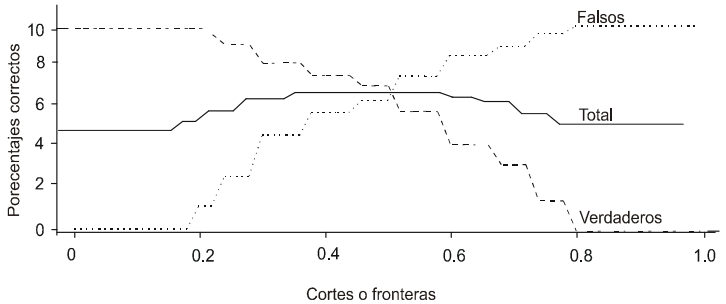


Figura 2.6. Capacidad de predictibilidad del modelo

Tabla 2.4 Desempeño de las predicciones. % correctos

Cortes	Verdadero	Falso	Total
0.00	100.00	0.00	48.84
0.05	100.00	0.00	48.84
0.10	100.00	0.00	48.84
0.15	100.00	0.00	48.84

0.20	97.03	11.33	53.19
0.25	91.99	26.35	58.41
0.30	83.09	46.18	64.20
0.35	83.09	46.18	64.20
0.40	76.85	57.22	66.81
0.45	71.51	63.46	**67.39**
0.50	71.51	63.46	**67.39**
0.55	58.75	75.35	67.25
0.60	42.73	85.55	64.64
0.65	42.73	85.55	64.64
0.70	32.94	90.37	62.32
0.75	15.13	96.03	56.52
0.80	0.00	100.00	51.16
0.85	0.00	100.00	51.16
0.90	0.00	100.00	51.16
0.95	0.00	100.00	51.16
1.00	0.00	100.00	51.16

Predicciones para p_j.

Fila	Observado	Val ajust	Err est	Lim inf de predic	Lim sup de predic
1	0.2	0.18715	0.025477	0.128401	0.245902
2	0.242857	0.23886	0.025626	0.179766	0.297954
3	0.3	0.29959	0.024538	0.243004	0.356175
4	0.35	0.36829	0.022638	0.316084	0.420493
5	0.45	0.44278	0.021004	0.394345	0.491218
6	0.505882	0.51994	0.020901	0.471742	0.56814
7	0.6	0.59616	0.022536	0.544192	0.648129
8	0.66	0.66801	0.024793	0.510832	0.725179
9	0.75	0.7328	0.02643	0.67185	0.793746
10	0.784615	0.78894	0.026812	0.727111	0.85077

2.18 Lecturas complementarias

2.18.1 Factores de inflación de varianzas (VIF)

Se muestran los elementos para llegar a los *VIF*, así como la transformación de correlaciones. Complementa 2.14.2. Se acudirá para efectos didácticos a un modelo con dos variables independientes dado por (2.128). En primer lugar, se deben reparametrizar las variables por normalidad, o sea trabajar con las desviaciones con respecto a la media de cada variable divididas por su error estándar,

$$X_{ij} = \frac{\left(X_{ij} - \bar{X}_j\right)}{\sqrt{\frac{s_j^2}{n}}} \quad (2.284) \qquad (0.1)$$

El uso de tales desviaciones $X_{ij} - \bar{X}_j$ permite modificar el modelo (2.128) por suma y resta de términos así:

$$Y_i = \left(\beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2\right) + \beta_1\left(X_{i1} - \bar{X}_1\right) + \beta_2\left(X_{i2} - \bar{X}_2\right) + \varepsilon_i \quad (2.285) \qquad (0.2)$$

que puede reescribirse como:

$$Y_i = \beta_0' + \beta_1\left(X_{i1} - \bar{X}_1\right) + \beta_2\left(X_{i2} - \bar{X}_2\right) + \varepsilon_i \quad (2.286) \qquad (0.3)$$

en la cual $\beta_0 = \beta_{0^*} + \beta_1\bar{X}_1 + \beta_2\bar{X}_2 = \bar{Y}$, con lo cual (2.128) puede reescribirse como:

$$Y_i - \bar{Y} = \beta_1\left(X_{i1} - \bar{X}_1\right) + \beta_2\left(X_{i2} - \bar{X}_2\right) \quad (2.287) \tag{0.4}$$

o sea, una regresión condicionada sin intercepto. Para normalizar basta entonces con expresar cada variable en términos de sus desviaciones estándar:

$$\frac{Y_i - \bar{Y}}{s_y} \;,\; \frac{X_{i1} - \bar{X}_1}{s_1} \;,\; \frac{X_{i2} - \bar{X}_2}{s_2} \quad (2.288) \tag{0.5}$$

en que s_y, s_1 y s_2 son las respectivas desviaciones estándar de los valores de las variables Y, X_1 y X_2, con lo cual se llega a las siguientes funciones de variables estandarizadas:

$$Y_i' = \frac{1}{\sqrt{(n-1)}}\left[\frac{Y_i - \bar{Y}}{s_Y}\right],\; X_{i1}' = \frac{1}{\sqrt{(n-1)}}\left[\frac{X_i - \bar{Y}}{s_Y}\right], \text{ etc.} \quad (2.289) \tag{0.6}$$

Lo que genera el modelo:

$$Y_i' = \beta_1' X_{i1}' + \beta_2' X_{i2}' + \varepsilon_i \quad (2.290) \tag{0.7}$$

que se relaciona con el modelo (0.4) así:

$$\beta_1 = \left(\frac{s_Y}{s_1}\right)\beta_1';\;\; \beta_2 = \left(\frac{s_Y}{s_2}\right)\beta_2' \;\;\therefore\;\; \beta_0 = \bar{Y} - \beta_1 X_1 - \beta_2 X_2 \quad (2.291) \tag{0.8}$$

Es decir, los nuevos coeficientes de regresión y los viejos se diferencian sólo por factores de escala que involucran razones de varianzas. La Matriz $\mathbf{X}$ para variables transformadas pierde entonces la columna de unos (1) quedando:

$$\mathbf{X} = \begin{bmatrix} X_{11}' & X_{11}' \\ X_{21}' & X_{22}' \\ \vdots & \vdots \\ X_{n1}' & X_{n2}' \end{bmatrix} \text{ y el producto } \mathbf{XX'} = \begin{bmatrix} \sum X_{i1}^{2\prime} & \sum X_{i1}{}' X_{i2}{}' \\ \sum X_{i2}{}' X_{i1}{}' & \sum X_{i2}^{2\prime} \end{bmatrix} \quad (2.292) \tag{0.9}$$

cuyos elementos son:

$$\left.\begin{aligned} &\sum X_{i1}'^2 = \sum \frac{1}{n-1}\frac{\left(X_{i1} - \bar{X}_1^2\right)^2}{s_1^2} = \frac{s_1^2}{s_1^2} = 1 \\ &\sum X_{i1}'^2 X_{i2}'^2 = \sum\left[\frac{1}{\sqrt{(n-1)}}\left(\frac{X_{i1} - \bar{X}_1}{s_1}\right)\frac{1}{\sqrt{(n-1)}}\left(\frac{X_{i2} - \bar{X}_2}{s_1}\right)\right] = \\ &\frac{1}{(n-1)}\frac{\sum\left(X_{i1} - \bar{X}_1^2\right)\left(X_{i2} - \bar{X}_2^2\right)}{\sqrt{\frac{\sum\left(X_{i1} - \bar{X}_1^2\right)^2}{(n-1)}\frac{\sum\left(X_{i2} - \bar{X}_2^2\right)^2}{(n-1)}}} = r_{12} \end{aligned}\right\} \quad (2.293),\ (0.10) \tag{0.10}$$

o sea el coeficiente de correlación entre X_1 y X_2, con lo cual:

$$\boldsymbol{X'X} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} = \mathbf{r}_{\mathbf{xx}} \quad (2.294) \tag{0.11}$$

llamada matriz de correlación de variables independientes que al extenderse al modelo general con $(p-1)$ de ellas toma la forma:

$$\mathbf{r}_{\mathbf{XX}_{(\boldsymbol{p-1})(\boldsymbol{p-1})}} = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p-1} \\ r_{21} & 1 & r_{23} & \dots & r_{2p-1} \\ r_{31} & r_{32} & 1 & \dots & r_{3p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p-11} & r_{p-12} & r_{p-13} & r_{12} & 1 \end{bmatrix} \quad (2.295) \qquad (0.12)$$

2.18.1.1 Estimadores mínimo cuadráticos de los β'_{ij}.

En el caso propuesto para el modelo (2.128) acudiendo al método conocido debemos calcular r_{XX}^{-1}. Para dos variables se tiene:

$$\mathbf{r}_{\mathbf{XX}}^{-1} = (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{(1-r_{12}^2)}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}; \ \mathbf{X}'\mathbf{Y} = \begin{bmatrix} r_{YX1} \\ r_{YX2} \end{bmatrix} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix} \quad (2.296) \qquad (0.13)$$

en que r_{Y1} y r_{Y2} son los coeficientes de correlación entre Y y X_1 y Y y X_2 con lo cual:

$$\mathbf{b}' = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} & r_{Y1} \\ -r_{12} & 1 & r_{Y2} \end{bmatrix} = \frac{1}{1-r_{12}^2}\begin{bmatrix} r_{Y1} & -r_{12} & r_{Y2} \\ r_{Y2} & 1 & r_{Y1} \end{bmatrix} \quad (2.297) \qquad (0.14)$$

Para volver a los coeficientes de regresión para el modelo original se emplean las relaciones:

$$b_1 = \left(\frac{s_Y}{s_1}\right) b'_1, \ \text{etc.} \quad (2.298) \qquad (0.15)$$

coeficientes expresados por la ecuación (2.291). La matriz de varianzas $\sigma^2(\mathbf{b}')$ se obtiene a partir de la ecuación $\sigma^2(\mathbf{b})$, entonces se encuentra que:

$$\boldsymbol{\sigma}^2(\mathbf{b}) = (\boldsymbol{\sigma}')^2 \mathbf{r}_{\mathbf{XX}}^{-1} = (\boldsymbol{\sigma}')^2 \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \quad (2.299) \qquad (0.16)$$

en que (σ')² es la varianza de los términos del error para el modelo transformado (2.290). De acuerdo con (2.299):

$$\sigma^2(b'_1) = \sigma^2(b'_2) = \frac{(\sigma')^2}{1-r_{12}^2} \quad (2.300) \qquad (0.17)$$

Aclaración importante, una alta multicolinealidad parece no ser un problema cuando el modelo de regresión se usa para hacer inferencias en la predicción de nuevas observaciones, siempre y cuando sean hechas dentro del rango de las observaciones, aunque con peligros inminentes en otras instancias e inferencias.

Los VIF son uno de los métodos más formales para detectar multicolinealidad. Para comprender el concepto se recuerda que la precisión, de los estimadores mínimo cuadráticos, depende de las varianzas de los coeficientes estimados de regresión, cuya matriz de varianzas covarianzas se da por (2.157). Asimismo la matriz de varianzas covarianzas de los coeficientes de regresión estandarizados estimados (b'_k) se obtiene como, como ya se vio :

$$\boldsymbol{\sigma}^2(\mathbf{b}) = (\boldsymbol{\sigma}')^2 \mathbf{r}_{\mathbf{XX}}^{-1} \quad (2.301) \qquad (0.18)$$

en que $\mathbf{r}_{XX}^{-1}$ es la matriz de correlaciones simples pareadas de las variables independientes y $(\boldsymbol{\sigma}')^2$ es la varianza de los términos del error para el modelo transformado. En (2.295) se observa que la varianza de b_k' $(k=1, 2,\ldots, p-1)$ es el producto de la varianza de los términos del error $(\boldsymbol{\sigma}')^2$ y el elemento k diagonal de la matriz r_{xx}^{-1}. Este segundo factor es el llamado factor de inflación de varianza (*VIF*). Puede entonces verse que el VIF_k, factor de inflación del coeficiente b_k' es:

$$VIF_k = \left(1 - R_k^2\right)^{-1} \quad (2.302) \tag{0.19}$$

en que R_k^2 es el coeficiente de determinación múltiple cuando se toma la variable independiente X_k en presencia de las otras X_{p-2} variables del modelo. Por lo tanto:

$$\sigma^2(b_k') = (\sigma')^2 (VIF_k) = \frac{(\sigma')^2}{1 - R_k^2} \quad (2.303) \tag{0.20}$$

El factor de inflación de varianza $(VIF)_k = 1$ si $R_k^2 = 0$, o sea cuando X_k no se relaciona linealmente con las otras X variables independientes pero, cuando $R_k^2 \neq 0$ $(VIF)_k > 1$, o sea que se infla la varianza de b_k'. Cuando $R_k^2 = 1$, tanto las varianzas como los VIF_k resultan indefinidos.

El mayor valor de VIF_k entre todas las $p-1$ variables es a menudo usado como indicador de multicolinealidad severa. VIF >10 se reportan como sí empezaran a afectar seriamente el proceso mínimo cuadrático. Si un $R_i^2 > 0{,}9$, la varianza de los correspondientes b_i se infla por un factor de al menos 10 (Daniel & Wood 1980d, 1980).
Otro criterio para medir la multicolinealidad es el valor promedio de los VIF definido como:

$$\overline{VIF} = \frac{\sum_{i=1}^{p-1} (VIF)_k}{p-1} \quad (2.304) \tag{0.21}$$

Valores de $\overline{VIF} \to 1$ son indicadores de seria multicolinealidad.

Por ejemplo: Se tomaron las siguientes variables con el deseo *exprofeso* de mostrar un efecto multicolineal, desde la total, la cual no requiere de análisis al resultar una variable como combinación lineal de otra, hasta la esperada en un modelo de tipo polinomial. Las variables son dap, dap^2, g=área basal y volumen.

d	d2	g	vol
9.00	81.000	0.01	0.029
21.00	441.000	0.03	0.179
14.00	196.000	0.02	0.063
11.00	121.000	0.01	0.075
9.00	81.000	0.01	0.038
15.00	225.000	0.02	0.071
16.00	256.000	0.02	0.161
10.00	100.000	0.01	0.035
19.00	361.000	0.03	0.186
23.00	529.000	0.04	0.170
10.00	100.000	0.01	0.027

14.00	196.000	0.02	0.083
12.00	144.000	0.01	0.049

Se presenta un primer modelo con multicolinealidad total, puesto que $g = \frac{\pi d^2}{40000}$

$$V = b_0 + b_1 d + b_2 d^2 + b_3 g \quad (2.305)(0.22)$$

```
d2n<-d^2
gn=3.1416*d2n/40000
mod1n<-lm(vol~d+d2n+gn)
 summary(mod1n)

Call:
lm(formula = vol ~ d + d2n + gn)

Residuals:
      Min        1Q    Median        3Q       Max
-0.034449 -0.017307 -0.003113  0.014614  0.043343

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1280408  0.0930167  -1.377    0.199
d            0.0187141  0.0128243   1.459    0.175
d2n         -0.0002099  0.0004098  -0.512    0.620
gn                  NA         NA      NA       NA

Residual standard error: 0.02629 on 10 degrees of freedom
Multiple R-squared:  0.8466,    Adjusted R-squared:  0.8159
F-statistic: 27.59 on 2 and 10 DF,  p-value: 8.504e-05
```

Se analiza el modelo polinomial de 2◦ grado que debe producir multicolinealidad intrínseca:

$$V = b_0 + b_1 d + b_2 d^2 \quad (2.306)(0.23)$$

```
mod2<-lm(vol~d+I(d^2))
summary(mod2)

Call:
lm(formula = vol ~ d + I(d^2))

Residuals:
      Min        1Q    Median        3Q       Max
-0.034449 -0.017307 -0.003113  0.014614  0.043343

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1280408  0.0930167  -1.377    0.199
d            0.0187141  0.0128243   1.459    0.175
I(d^2)      -0.0002099  0.0004098  -0.512    0.620

Residual standard error: 0.02629 on 10 degrees of freedom
Multiple R-squared:  0.8466,    Adjusted R-squared:  0.8159
F-statistic: 27.59 on 2 and 10 DF,  p-value: 8.504e-05
```

Se presentan además los modelos individuales con las variables separadas, para resaltar su efecto como tales; ejemplo con sus estadísticos:

$$V = b_0 + b_1 d \quad (2.307)(0.24)$$

```
mod3<-lm(vol~d)#
> summary(mod3)

Call:
lm(formula = vol ~ d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.02995 -0.01535 -0.00495  0.01025  0.04784
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08206    0.02347  -3.497    0.005 **
d            0.01220    0.00159   7.672  9.7e-06 ***

Residual standard error: 0.02539 on 11 degrees of freedom
Multiple R-squared:  0.8425,    Adjusted R-squared:  0.8282
F-statistic: 58.86 on 1 and 11 DF,  p-value: 9.703e-06
```

$$V = b_0 + b_1 d^2 \quad (2.308)$$

```
mod4<-lm(vol~I(d^2))#
summary(mod4)

Call:
lm(formula = vol ~ I(d^2))

Residuals:
      Min         1Q     Median         3Q        Max
-0.038959 -0.017562 -0.008281  0.003763  0.056657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.241e-03  1.426e-02   0.438     0.67
I(d^2)      3.832e-04  5.525e-05   6.936 2.47e-05 ***

Residual standard error: 0.0276 on 11 degrees of freedom
Multiple R-squared:  0.8139,    Adjusted R-squared:  0.797
F-statistic: 48.11 on 1 and 11 DF,  p-value: 2.468e-05
```

$$V = b_0 + b_1 gn \quad (2.308)$$

```
mod5<-lm(vol~gn)#
summary(mod5)

Call:
lm(formula = vol ~ gn)
Residuals:
      Min         1Q     Median         3Q        Max
-0.038959 -0.017562 -0.008281  0.003763  0.056657

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.006241   0.014261   0.438     0.67
gn          4.879182   0.703474   6.936 2.47e-05 ***

Residual standard error: 0.0276 on 11 degrees of freedom
Multiple R-squared:  0.8139,    Adjusted R-squared:  0.797
F-statistic: 48.11 on 1 and 11 DF,  p-value: 2.468e-05
```

A continuación, se muestra el estudio del caso polinomial, para lo cual se muestra el modelo reparametrizado de acuerdo con (2.288) y siguientes:

$$X_1 = \frac{d - 14{,}077}{\sqrt{12} * 4{,}61} \quad X_2 = \frac{d^2 - 217{,}77}{\sqrt{12} * 144{,}23} \quad X_3 = \frac{V - 0{,}0897}{\sqrt{12} * 0{,}06127} \quad (2.309) \qquad (0.22)$$

a la variable X_1 se le llamará d', a X_2 d'^2 y a X_3 V'. Sus elementos:

```
vpro<-mean(vol); vpro [1] 0.08969231; dpro<-mean(d); dpro [1] 14.07692; d2npro<-mean(d2n);
d2npro [1] 217.7692; gnpro<-mean(gn); gnpro [1] 0.0171036; var(vol) [1] 0.003753397; var(d)[1]
21.24359;  var(d2n);  [1]  20802.53;  var(gn);  [1]  0.000128321;  vpri<-(1/sqrt(length(vol)-
1))*(vol-vpro)/sd(vol);sd(d) [1] 4.609077; sd(d2);[1] 144.2308

vpri
 [1] -0.28597677  0.42080992 -0.12577178 -0.06922885 -0.24356957 -0.08807649
 [7]  0.33599552 -0.25770530  0.45379330  0.37840272 -0.29540059 -0.03153356
[13] -0.19173854

dpri<-(1/sqrt(length(d)-1))*(d-dpro)/sd(d)
```

```
dpri
 [1] -0.317977219  0.433605299 -0.004817837 -0.192713466 -0.317977219
 [6]  0.057814040  0.120445916 -0.255345343  0.308341546  0.558869052
[11] -0.255345343 -0.004817837 -0.130081590

d2npri<-((d2n-d2npro)/sd(d2n))/sqrt(12)
d2npri
 [1] -0.27374094  0.44679201 -0.04357069 -0.19368172 -0.27374094  0.01447224
 [7]  0.07651814 -0.23571281  0.28667358  0.62292229 -0.23571281 -0.04357069
[13] -0.14764767
```

$$V' = \beta_1 d' + \beta_2 d'^2 \quad (2.310)$$

```
mod6<- lm(vpri~dpri+d2npri-1)
summary(mod6)
Call:
lm(formula = vpri ~ dpri + d2npri - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16232 -0.08155 -0.01467  0.06886  0.20423

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
dpri     1.4079     0.9199   1.530    0.154
d2npri  -0.4941     0.9199  -0.537    0.602

Residual standard error: 0.1181 on 11 degrees of freedom
Multiple R-squared:  0.8466,    Adjusted R-squared:  0.8187
F-statistic: 30.35 on 2 and 11 DF,  p-value: 3.331e-05

anova(mod6)#Para mirar sus componentes
Analysis of Variance Table

Response: vpri
          Df  Sum Sq Mean Sq F value   Pr(>F)
dpri       1 0.84254 0.84254 60.4037 8.59e-06 ***
d2npri     1 0.00402 0.00402  0.2885   0.6019
Residuals 11 0.15343 0.01395
```

Se presenta también para comparar:

mod7<- lm(vpri~dpri-1) $V' = \beta_1 d'$

```
summary(mod7)

Call:
lm(formula = vpri ~ dpri - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14114 -0.07234 -0.02332  0.04830  0.22544

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
dpri   0.9179     0.1145   8.013  3.7e-06 ***

Residual standard error: 0.1145 on 12 degrees of freedom
Multiple R-squared:  0.8425,    Adjusted R-squared:  0.8294
F-statistic: 64.21 on 1 and 12 DF,  p-value: 3.697e-06
```

Los valores convertidos del mod relacionados con el modelo original de acuerdo con (2.291) son:

```
b1<-sd(vol)/sd(d)*1.4079
b1
[1] 0.01871415

b2<-sd(vol)/sd(d2n)*-0.4941
```

```
b2
[1] -0.0002098791
b0<-mean(vol)-b1*mean(d)-b2*mean(d2n)
b0
[1] -0.1280402
```

Para el cálculo de los *VIF* se tiene entonces:

```
matxtx<-matrix(c(1,cor(dpri,d2npri),cor(dpri,d2npri),1),nrow=2)
matxtx#matrxx
         [,1]     [,2]
[1,] 1.000000 0.991724
[2,] 0.991724 1.000000

matxtxin<-solve(matxtx)
matxtxin
          [,1]      [,2]
[1,]  60.66666 -60.16458
[2,] -60.16458  60.66666
```

$$VIF_1 = 60.666 \tag{0.23}$$

con lo cual se obtiene un resultado esperado y que muestra la alta correlación entre las dos variables independientes.

Se debe hacer notar que no siempre es destapable la multicolinealidad con la matriz de correlaciones $\mathbf{r_{xx}}$. También para el caso analizado se puede reanotar que una alta multicolinealidad no es usualmente un problema cuando se trata de inferencias o predicciones de nuevos resultados cuando estos no salen del rango de las observaciones que originaron el modelo, lo que ampara el uso de los modelos polinomiales en este campo. En R, la librería car, calcula los VIF así:

```
library(car)

vif(mod5)
    dpri   d2npri
60.66666 60.66666
```

Todos estos VIF superan el valor de 10, reflejando la alta multicolineallidad.

2.18.2 Modelos de correlación normal

2.18.2.1 Distinción entre regresión y correlación.

El análisis de regresión estima una función completa, la ecuación de regresión, un análisis de correlación produce únicamente un índice designado para dar una idea inmediata de la relación conjunta cercana de dos variables (Wonnacott & Wonnacott 1981), o en otro contexto para probar si dos variables aleatorias X_1 y X_2 son independientes o no bajo el supuesto de normalidad.

A veces ha existido mucha confusión entre ambos temas, unos problemas se tratan como de regresión siendo de correlación. Sokal & Rohlf (1969) presentan una buena discusión al respecto que escapa de lo elemental de nuestro capítulo.

Para aclarar un poco más los conceptos, Walpole & Miers (1984) usan el término regresión aplicable sólo a modelos en que la variable X es matemática, es decir, no aleatoria con un error insignificante. Los modelos estudiados en regresión asumieron que las variables

independientes X_1, X_2,…, X_k eran constantes fijas, pero que funcionaban bien aun cuando las X fueran variables aleatorias.

Los modelos de correlación son aquellos en que todas las variables son aleatorias, por lo cual su gran diferencia con los modelos de regresión es especificar la distribución conjunta de las variables completamente, lo que hace que ellas jueguen un papel simétrico sin que puedan designarse unas u otras como dependientes o independientes (Neter *et al*. 1983).

Se emplean los índices de correlación para estudiar la naturaleza de las relaciones entre las variables y para obtener inferencias de alguna de las variables con base en las otras. Los más ampliamente empleados son los de distribuciones normales bivariadas.

2.18.2.2 Distribución normal bivariada. Los modelos de correlación normal en el caso de dos variables se basan en la distribución normal bivariada (*DNB.*). Dos variables X_1, X_2 se dice que son distribuidas conjunta y normalmente si su distribución conjunta de probabilidades es la *DNB.*, la cual tiene la siguiente función de densidades:

$$f(X_1, X_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} EXP\left[-\frac{1}{2(1-\rho_{12}^2)} * \left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{X_1-\mu_1}{\sigma_1}\right) * \left(\frac{X_2-\mu_2}{\sigma_2}\right) + \left(\frac{X_2-\mu_2}{\sigma_2}\right)\right]\right] \quad (0.24)$$

2.18.2.2.1 Representación gráfica. La DNB es una superficie continua en un espacio euclidiano tridimensional, en que para cada par de puntos (X_1, X_2) hay una densidad $f(X_1, X_2)$ equivalente a la altura alcanzada por la superficie en el par cartesiano (X_1, X_2) y su probabilidad corresponde al volumen delimitado por el plano X_1 X_2 y la superficie S (**¡Error! No se encuentra el origen de la referencia.**).

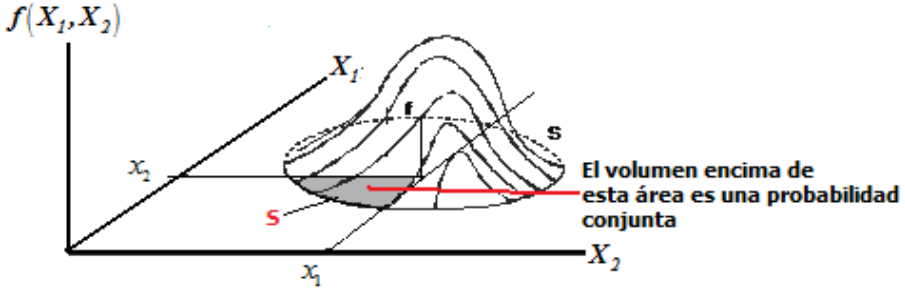


Figura 2.7. Representación de la distribución normal bivariada.

2.18.2.2.2 Distribuciones marginales. Si X_1 y X_2 son normal y conjuntamente distribuidas, sus distribuciones marginales tienen las siguientes características:
1-La distribución marginal de X_1 es normal con media μ_1, y desviación estándar σ_1

$$f_1(X_1) = \frac{1}{\sqrt{2\pi}\sigma_1} EXP\left[\frac{1}{2}\left(\frac{X_1-\mu_1}{\sigma_1}\right)\right] \quad (0.25)$$

2- La distribución de X_2 es similar a la anterior con media μ_2 y desviación estándar σ_2 y se llama $f_2(X_2)$.

3- No siempre, pero en la mayoría de los casos, sí una variable X_1 es normalmente distribuida, idem una variable X_2, entonces deberán ser conjunta y normalmente distribuidas.

2.18.2.2.3 Parámetros de la DNB. μ_1, σ_1; μ_2, σ_2 son respectivamente las medias y las desviaciones estándar de las distribuciones marginales de X_1 y X_2; $\boldsymbol{\rho}_{12}$ es el coeficiente de correlación entre las variables aleatorias X_1 y X_2 y se define como:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \tag{0.26}$$

en la que σ_{12} es la covarianza entre $X_1 X_2$, o $\sigma_{12} = E\left[\left(X_1 - \mu_1\right)\left(X_2 - \mu_2\right)\right]$.

Ya se sabe que si $\sigma_{12} = 0$ entonces $\rho_{12} = 0$. ρ_{12} es un número adimensional, toma valores entre -1 y +1 y su signo habla del tipo de relación inversa o directa entre las variables. Estas *DNB* frecuentemente se representan como unos diagramas de contorno compuestos de todos los puntos de la superficie equidistantes del plano $X_1\ X_2$, lo que equivale a decir que tienen una densidad constante $f\left(X_1, X_2\right)$. (**¡Error! No se encuentra el origen de la referencia.**). Una característica de estas superficies de contorno es que todas son elipses excepto en el caso $\rho_{12} = 0$ y $\sigma_1 = \sigma_2$, con centro $\left(\mu_1, \mu_2\right)$ y con ejes mayores y menores comunes; y que a medida que la distancia al plano escogido contra el de $X_1\ X_2$ sea mayor, menor será la correspondiente elipse de contorno.

2.18.2.2.4 Inferencias condicionales.

Un uso importante de la correlación bivariada es hacer inferencias condicionales relativas a una variable con respecto a otra, por ejemplo sobre probabilidades al estimar la altura de los árboles cuando su diámetro a la altura del pecho sea de 30 cm. Esto requiere trabajar las distribuciones de probabilidad condicional.

2.18.2.2.5 Distribuciones de probabilidades condicionales de X1 y X2. — La función de densidad condicional de X_1 para un valor dado de X_2 se denotará como:

$$f\left(X_1 \mid X_2\right) = \frac{f\left(X_1, X_2\right)}{f_2\left(X_2\right)} \tag{0.27}$$

función de X_1 dado un X_2, cuya distribución de probabilidad condicional será una normal con media $\alpha_{1.2} + \beta_{1.2} X_2$ y una desviación estándar $\sigma_{1.2}$:

$$f\left(X_1 \mid X_2\right) = \frac{1}{\sqrt{2\pi}\sigma_{1.2}} EXP\left[-\frac{1}{2}\left(\frac{X_1 - \alpha_{1.2} - \beta_{1.2} X_2}{\sigma_{1.2}}\right)\right] \tag{0.28}$$

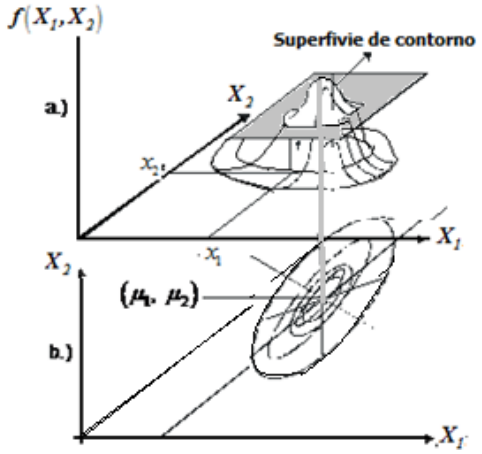


Figura 2.8. a. Superficie de contorno para la DNB; b. Diagrama de contorno para la DNB

Los parámetros $\alpha_{1.2}$, $\beta_{1.2}$ y $\sigma_{1.2}$ de esta distribución son funciones de los parámetros de la distribución conjunta de probabilidades así:

$$\alpha_{1.2}=\mu_1-\mu_2\rho_{12}\frac{\sigma_1}{\sigma_2};\ \ \beta_{1.2}=\rho_{12}\frac{\sigma_1}{\sigma_2};\ \ \alpha_{1.2}^2=\sigma_1^2\left(1-\rho_{12}^2\right) \tag{0.29}$$

similarmente:

$$f\left(X_2\mid X_1\right)=\frac{f\left(X_1,X_2\right)}{f\left(X_1\right)}=\frac{1}{\sqrt{2\pi}\sigma_{2.1}}EXP\left[-\frac{1}{2}\left(\frac{X_2-\alpha_{2.1}-\beta_{2.1}X_1}{\sigma_{2.1}}\right)^2\right] \tag{0.30}$$

con:

$$\alpha_{2.1}=\mu_2-\mu_1\rho_{12}\frac{\sigma_2}{\sigma_1};\ \ \beta_{2.1}=\rho_{12}\frac{\sigma_2}{\sigma_1};\ \ \alpha_{2.1}^2=\sigma_2^2\left(1-\rho_{12}^2\right) \tag{0.31}$$

2.18.2.2.6 Características de distribuciones conjuntas de probabilidades.

1. Las distribuciones condicionales de probabilidades de X_1, para un X_2 dado son normales, lo mismo sucede para X_2 dado un X_1. Si se cortara el volumen de la *DNB* por un punto X_{21} no importa su posición en el eje X_2 paralelo a X_1, se obtendrá una normal, que podría asignarle un valor del área bajo su curva cambiando escala. **Figura 2.3**.

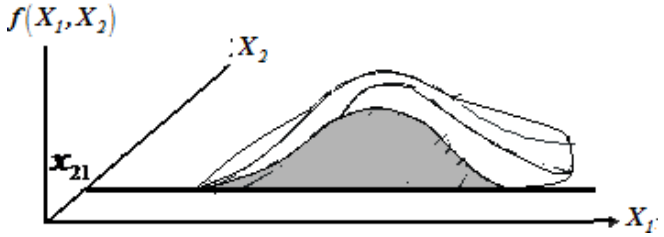


Figura 2.3. Distribución conjunta de probabilidades

2. Las medias de la distribución condicional de probabilidades de X_1 caen en una línea recta y son funciones lineales de X_2 (similarmente para X_2, funciones lineales de

$$E\left(X_1\mid X_2\right)=\alpha_{1.2}+\beta_{1.2}X_2;E\left(X_2\mid X_1\right)=\alpha_{2.1}+\beta_{2.1}X_1 \tag{0.32}$$

donde: $\alpha_{1.2}$ es el intercepto de la línea de regresión de X_1 en X_2 y $\beta_{1.2}$ la pendiente (lo mismo, $\alpha_{2.1}$ es el intercepto de la regresión de X_2 en X_1 y $\beta_{2.1}$ la pendiente). Entonces la relación entre las medias condicionales y X_1 (o X_2) se da a través de funciones de regresión. Todas las distribuciones condicionales de probabilidades de X_1 tienen la misma desviación estándar $\sigma_{1.2}$. Similarmente para las de X_2 con $\sigma_{2.1}$. (Poniendo escalas adecuadas además, el área bajo sus curvas será 1), o sea que las distribuciones de probabilidades condicionales de X_1 se caracterizan por tener homocedasticidad (similar para las de X_2). En vista de las similitudes y equivalencias de (2.316) y (2.318) con los modelos de regresión estudiados, las inferencias condicionales con estos modelos de correlación pueden hacerse a través de aquellos.

2.18.2.2.7 Usos de la correlación. El principal uso de la correlación es el estudio de las relaciones entre variables. En modelos *DNB* el parámetro ρ_{12} y su ρ_{12}^2 proporcionan el grado de esta relación entre X_1 y X_2, siendo ρ_{12}^2 más interesante estadísticamente, llamado coeficiente de determinación. De las ecuaciones (2.317) y (2.319) se obtiene que:

$$\rho_{12}^2 = \frac{\sigma_1^2 - \sigma_{1.2}^2}{\sigma_1^2} = \frac{\sigma_2^2 - \sigma_{2.1}^2}{\sigma_2^2} \tag{0.33}$$

Nuevamente se hace claro su significado. ρ_{12}^2 mide cuan más pequeña es la variabilidad relativa en una distribución condicional de X_1 para un valor dado de X_2, que es la variabilidad en la distribución marginal de X_1, o lo que es lo mismo ρ_{12}^2 mide la relativa reducción en la variabilidad de X_1 al usar otra variable X_2. (Similar para X_2). El coeficiente de determinación: $0 \le \rho_{12}^2 \le 1$. $\rho_{12}^2 = 0$ si X_1 y X_2 son independientes. Los estimadores puntuales de ρ_{12} se obtienen como se vio para la ecuación (2.291):

$$r_{12} = \frac{\sum\left(X_{i1} - \bar{X}_1\right)\left(X_{i2} - \bar{X}_2\right)}{\sqrt{\sum\left(X_{i1} - \bar{X}_1\right)^2 \sum\left(X_{i2} - \bar{X}_2\right)^2}} \tag{0.34}$$

r_{12} es sesgado a menos que $\rho_{12} = 0$ o $\rho_{12} = 1$; pero para valores de n grandes el sesgo es despreciable. Sokal & Rolph (1969) presentan una corrección al respecto, sugerida por Kendall & Stuart, citados por ellos

$$r_{ij}^* = r_{ij}\left[1 + \frac{1 - r_{ij}^2}{2(n-4)}\right] \tag{0.35}$$

$r_{ij}^* =$ coeficiente de correlación imparcial.

2.18.2.2.8 Prueba para analizar si $\rho_{12} = 0$. Cuando una población se supone *DNB* es deseable analizar:

$$H_0 : \rho_{12} = 0; \;\; H_a : \rho_{12} \neq 0 \tag{0.36}$$

porque en caso de ser X_1 y X_2 y distribuidas conjunta y normalmente $\rho_{12} = 0$ implicaría que X_1 y X_2 son independientes. Se hace énfasis de lo útil de estos conceptos para la ciencia forestal, para poder establecer alguna significación a los valores de R^2 obtenidos en regresión, como una manera de eliminar las subjetividades al analizar los r^2.

Se aclara nuevamente que es un error confundir estos conceptos y que lo único que se intenta es usar el uno como indicativo del otro. Como anotan Walpole & Miers (1984) "debe decirse entonces que las estimaciones muestrales de ρ, cuyo valor absoluto se acerque a la unidad, implican una buena correlación o asociación lineal entre X y Y, mientras que los valores cercanos a cero implican poca o nula correlación lineal". Este concepto debe revaluarse, puesto que estas cercanías son relativas. Existen tablas al respecto de Steell *et al.* (1985), incluso cuestionadas por no poderse ligar a una distribución dada. Utilizando procesos de regresión, de acuerdo con (2.323) y (2.324), las siguientes alternativas son equivalentes:

$$H_0 : \{\beta_{1.2} \text{ o } \beta_{2.1}\} = 0; \;\; H_a : \{\beta_{1.2} \text{ o } \beta_{2.1}\} \neq 0 \tag{0.37}$$

puede mostrarse que la prueba para analizar (2.323) y ((2.324) se puede expresar en términos de r_{12} como:

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} \tag{0.38}$$

y, si se cumple H_0, t^* sigue una distribución con $(n-2)$ grados de libertad. Las reglas de decisión son:

$$si \left|t^*\right| \leq t_{\left[(1-\alpha/2);\,(n-2)\right]} \rightarrow H_0 ; si \left|t^*\right| > t_{\left[(1-\alpha/2);\,(n-2)\right]} \rightarrow H_a \tag{0.39}$$

idéntico a lo obtenido para la regresión.

2.18.2.2.9 Estimación de intervalos de ρ_{12}. Si el coeficiente de correlación $\rho = 0$ la distribución de r no es simétrica, aumentando ésta sólo cuando r se aproxima a +1 ó -1. Además, hay dependencia del tamaño de la muestra (Gómez 1989). A causa de la complicación cuando $\rho_{12} = 0$ ó 1, Fisher citado por Neter *et al*. (1983) propone una transformación:

$$Z' = \frac{1}{2}\ln\left(\frac{1+r_{12}}{1-r_{12}}\right) \tag{0.40}$$

cuando n es grande $(n \geq 25)$ la Z' se vuelve aproximadamente normal con media y varianza:

$$E(Z') = \hat{Z}' = \frac{1}{2}\ln\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right) \text{ y } \sigma^2(Z') = \frac{1}{n-3} \tag{0.41}$$

Se nota que Z' y $E(Z')$ coinciden y que $\sigma^2(Z')$ es una constante que depende del tamaño de la muestra. Existen tablas para calcular los valores derechos e izquierdos de Z' que eliminan la necesidad de cálculos al respecto. Supongamos r_{12} o $\rho_{12} = 0.43$ entonces Z' o $E(Z') = 0.4599$ o viceversa. Si r_{12} o ρ_{12} fuera negativa se cambia el signo respectivo. Ej: Si $r_{12} = -0.43$, $Z' = E(Z') = -0.4599$.

Por ejemplo, se buscará estimar el grado de asociación y además un intervalo de estimación para el coeficiente de correlación entre las variables diámetro a la altura del pecho y volumen en un bosque degradado en la zona de Piedras Blancas, para lo cual se extrae una muestra de 30 árboles de 150 cubicados al respecto, (Puche 1988),

Tabla 2.5. Diámetro d y volumen V, de 30 árboles cubicados en un bosque degradado en la zona de Piedras Blancas (Colombia), tomada de Puche (1981).

Diámetro *d*	Volumen *V*	Diámetro *d*	Volumen *V*
16.50	0.1611	15.00	0.0704
9.00	0.0280	11.00	0.0407
10.00	0.0597	8.51	0.0273
11.50	0.0529	12.50	0.0851
6.50	0.0121	13.00	0.0650
14.00	0.0856	21.00	0.1249
17.00	0.0988	23.00	0.2689
24.00	0.2169	29.00	0.3067
32.00	0.4346	16.00	0.1280
19.00	0.1630	18.00	0.0993
22.00	0.2070	12.00	0.0482
26.00	0.3911	14.00	0.826
8.00	0.0175	15.00	0.1061
23.00	0.1696	27.00	0.5608
21.00	0.1792	15.00	0.1417

Para el cálculo del intervalo de confianza a un nivel $\alpha = 0,05$ se procede así:

$$Z' = \frac{1}{2}\ln\left(\frac{1+0.5834}{1-05834}\right) = 0.6676^{*};\ \ \sigma(Z') = \sqrt{\frac{1}{30-3}} = 0.1925 \tag{0.42}$$

Los intervalos de confianza para $E(Z')$ son:

$$L_i = Z' \pm Z_{(1-\alpha/2)}\sigma(Z) \therefore L = 0.6676 \pm 1.96 * 0.1945 = \begin{cases} L_1 = 0.2904 \\ L_2 = 0.7797 \end{cases} \tag{0.43}$$

Los límites en su valor real se obtienen transformando los anteriores por medio de la función inversa de Z' dada por:

$$f^{-1}(Z) = r = \frac{e^{2Z}-1}{e^{2Z}+1} \ \therefore \left\langle \begin{array}{l} r_1 = 0.2825 \\ r_2 = 0.7797 \end{array} \right. \tag{0.44}$$

$n = 30$ datos; $r = 0.5834$. Lo anterior quiere decir que si se muestreara en el mismo lugar tendríamos un r sigificativo entre estos dos valores. En R:

```
cordv<-(cor(d,v))
cordv
[1] 0.5834327

zpri<-0.5*log((1+cordv)/(1-cordv))

zpri
[1] 0.6676511

varZ<-1/(length(d)-3)
varZ
[1] 0.03703704
LIz<-zpri-1.96*varZ^.5

LIz
[1] 0.2904489

LSz<-zpri+1.96*varZ^.5
LSz
[1] 1.044853

finZinf<-(exp(2*LI)-1)/(exp(2*LI)+1)
finZinf
[1] 0.282548
```

```
finZsup<-(exp(2*LS)-1)/(exp(2*LS)+1)
finZsup
[1] 0.7797974
>
```

Podría decirse que con una probabilidad del 95% el coeficiente de correlación ρ entre diámetro y volumen en ese bosque se encuentra bien calculado, lo cual corroboramos con R donde se encuentra una prueba de t para los coeficientes de correlacion:

```
cor.test(d,v)
Pearson's product-moment correlation
data: d and v
t = 3.8029, df = 28, p-value = 0.0007109
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2827910 0.7798954
sample estimates:
cor
0.5836023
```

Se puede anotar también que el intervalo de confianza para ρ_{12} puede emplearse para verificar si o no ρ_{12} tiene un valor especificado, como en el ejemplo 0.53, es decir, para mostrar que es altamente posible la relación lineal escogida pero no para asegurar que cae dentro de esos límites.

Debe puntualizarse entonces que si un modelo de regresión es adecuado es posible usar con cautela las inferencias salidas del estudio de correlación, cuando además es posible suponer densidad normal bivariada para las variables X y Y (Walpole 1984).

BIBLIOGRAFIA

ABRAHAM, B. & JOHANNES LEDOLTER. 1983. Statiscal methods for forecasting. N. Y. John Wiley and Sons. 445 p.

BARRENA, V. 1988. La regresión ponderada en la elaboración de ecuaciones de Volumen. Rev. For. del Perú. 15 (2): 21-28.

CAILLEZ. 1980. Estimación del volumen forestal y predicción del rendimiento, con referencia especial a los trópicos. Vol. 1 (Estimación de volumen). Roma, FAO. 92 p.

CHATTERJEE, S. y PRICE, B. 1977. Regression analysis by example. N. Y. John Wiley and Sons. 407 p.

DANIEL, C. y WOOD, F. 1980. Fitting Equations to data. 2ª ed. N. Y. John Wiley and Sons.

DRAPER, N. & SMITH, H. 1966. Applied regression analysis. N. Y. John Wiley and Sons. 407 p.

FINNEY. 1978. Statistical Method in Biological Assay. London. Charles Griffin & Co. 3ª edición. 508 p.

FURNIVAL, G. 1961. An Index for comparing equations used in constructing volumen tables. Forest Scie: 7(4); p. 337-341.

GOMEZ, HERNÁN. 1989. Estadística experimental con aplicaciones a las ciencias agrícolas. Medellín. Centro de Publicaciones U. N. 615 p.

GREEN, EDWIN J. 1983. Evaluating the predictive Abilities of Regressions with PRESS. Forest Scie. 29 (4); p. 712-714.

JOHNSTON, J. 1963. Econometrics Methods. N. Y. McGraw Hill. 300 p.

JOHNSTON, R. y D. WICHERN. 1988. Applied Multivariate Statiscal Analysis. London. Prentice - Hall International. 607 p.

KVÅLSETH, TARALDO. 1985. Cautionary Note About R□. The American Statistician. Parte 1. 39 (4); p. 279-285.

MATEO, L. y EMILIO MIGUEL. 1979. Evaluación de la falta de ajuste en modelos de regresión obtenidosmediante diferentes procedimientos de selección. Chapingo, México. Tesis de Posgrado.65p.

MONTGOMERY, D. y E. PECK. 1982. Introduction to linear regession analysis. N. Y. John Wiley and Sons. 504 pp.

NETER, J. WASSERMAN, W. & M., KUTNER. 1985. Applied linear statistical Models. 2ª ed. Illinois. Richard Iroing. 1127 p.

OROZCO, G y MAGIDIN, M. 1976. Métodos para estimación no lineal. Chapingo, México. Agrociencia # 26. 103-112

PUCHE, IVÁN. 1989. Datos de campo. (Proyecto de Trabajo de Grado). Sin publicar.

ROUSSEEUW, P, y LEROY A. 1987. Robust Reression and Outlier Detection. N.Y. John Wiley & Sons. 329

SCHREUDER *et al*. 1987. PPS and Randon Sampling Estimation Using some Regression and Ratio Estimators for Underlying Linear and Curvilinear Models. For. Sci. 33 (4): 997-1009.

SCLAGEL, B. 1985. Confidence Bounds for the sum of Volumes Predicted by Weighted Regressions. For. Sci., 31: 65-71.

SOKAL, RR. y F. S. RHOLF. 1979. Biometría. Principios y métodos estadísticos en investigación biológica. Ed. Blume. Madrid. 832 p.

STEEL, R. G. y J. H. TORRIE. 1985. Bioestadística. Principios y procedimientos. Mc Graw Hill. Bogotá. 622 p.

SNEDECOR, G. and W. COCHRAN. 1971. Métodos estadísticos. México C.E.C.S.A. 703 p.
URIBE, A. 1985. Cómputo de la regresión lineal múltiple y de la regresión no lineal. Medellín. Crónica Forestal y del Medio Ambiente. Vol. III.
WALPOLE, R. & MYERS, R. 1984. Probabilidad y estadística para ingenieros. 2ª ed. México. Interamericana S.A. 578 p.
WONNACOTT, T. and R. WONNACOTT, R. 1981. Regression: A second course in statistics. N. Y. John Wiley and sons. 489 p.