ANÁLISIS DE REGRESIÓN

Edgar Acuña Fernandez

Departamento de Matemáticas

Universidad de Puerto Rico

Reciento Universitario de Mayaguez

Enero 14, 2008

PREFACIO

La razón principal de escribir este libro es la carencia de un texto completo de regresión que cubra las diversas técnicas de regresión, especialmente aquellas que han tomado auge en la última década. Un par de buenos libros de regresión son el "Classical and Modern Regression with applications" de Myers y el "Applied linear Regression" de Weisberg, pero ambos cubren muy poco material de selección de variables, regresión robusta y la muy importante área de regresión no paramétrica, la cual es prácticamente ignorada en ambos textos. Existen por otro lado buenos textos cubriendo solamente Regresión Robusta como el "Robust regression and outlier detection" de Rousseeuw y Leroy y otros que tratan exclusivamente Regresión noparamétrica como el "Applied nonparametric regression" de Haerdle. El objetivo de este texto es cubrir la parte más transcendental de los libros antes mencionados.

En el transcurso de los ocho años que he venido desarrollando el texto he usado varios programa estadísticos tales como: MINITAB, SAS, MATLAB, S-PLUS y últimamente R. La meta final es desarrollar todo el texto usando el programa gratuito R, el cual está disponible en www.r-project.org. Aún quedan en el texto algunas salidas de MINITAB. Las salidas de SAS, MATLAB y S-Plus están siendo eliminadas poco a poco.

Aunque el texto es en regresión aplicada también se ha tratado de probar varias identidades y propiedades de estimadores que aparecen en regresión. Sin embargo no es nuestra intención llenar el texto con demostraciones teóricas. Dos buenos textos donde se ve el lado teórico de Regresión son "Linear Regression Analysis" de Seber y "Linear statistical inference and its applications" de Rao.

El texto está organizado en 9 capítulos. El primer capítulo se enfoca en regresión lineal simple y el segundo en regresión lineal múltiple. En el tercer capítulo se discute los diversos métodos de diagnosticar si las suposiciones del modelo de regresión se están cumpliendo o no. En el capítulo 4 se estudian diferentes transformaciones que se pueden hacer de las variables predictoras y de la variable de respuesta con la finalidad de mejorar el modelo de regresión para que haga un mejor ajuste de los datos. En el capítulo 5 se discute modelos de regresión considerando la presencia de variables categóricas. Aquí se incluye el estudio de la regresión logística. El capítulo 6 está dedicada al importante problema de selección de variables en regresión y en el problema 7 se discute la forma de detectar y resolver el problema de multicolinealidad entre las variables predictoras. Los capítulos 8 y 9 están dedicados a regresión robusta y regresión noparamétrica respectivamente.

Los conjuntos de datos que aparecen en este texto pueden ser obtenidos en el siguiente sitio de la internet en www.math.uprm.edu/~edgar/class6205.htm.

Finalmente, deseo agradecer la ayuda de mi asistentes de investigación Srtas. Milena Restrepo y Frida Coaquira por colaborar conmigo en la depuración de errores presentes en el texto, así como en la edición de algunos capítulos y en la preparación de las transparencias del texto.

Por favor para reportar cualquier sugerencia o error mandarme un e-mail a edgar@cs.uprm.edu.

Mayagüez, Enero 14, 2008

CONTENIDO

CAPÍTULO 1

REGRESIÓN LINEAL SIMPLE

1.1. Introducción

Regresión es un conjunto de técnicas que son usadas para establecer una relación entre una variable cuantitativa llamada *variable dependiente* y una o más variables independientes llamadas *variables predictoras*. Las variables independientes también deberían ser cuantitativas, sin embargo es permitido que algunas de ellas sean cualitativas. La ecuación que representa la relación es llamada el **modelo de regresión.** Si todas las variables independientes fueran cualitativas entonces el modelo de regresión se convierte en un modelo de diseños experimentales.
Ejemplos de modelos de regresión:

a) La variable de respuesta puede ser la tasa de divorcio en tanto que una variable predictora puede ser el nivel de ingreso familiar.
b) El precio de una casa puede ser la variable dependiente mientras que el área, el número de cuartos, el número de baños, y los años de antiguedad de la casa pueden ser usadas como variables predictoras.

Para estimar la ecuación del modelo se debe tener una muestra de entrenamiento. En el caso de una sola variable independiente, esta muestra consiste de n pares ordenados (x_i,y_i) para i=1,..,n. En el caso de varias variables independientes se deben tener n nuplas $(\mathbf{x}_i,y_i)$, para i=1,..,n, domde $\mathbf{x}_i$ es el vector de mediciones de las variables predictoras para la i-ésima observación.

Ejemplo 1. En la siguiente tabla se muestra la tasa de mortalidad infantil (muertes de niños de 5 años o menos por cada 1,000 nacidos vivos) y el porcentaje de vacunación en veinte países del mundo.

	NACION	%INMUNIZACION	TASA_mor
1	"Bolivia"	77	118
2	"Brazil"	69	65
3	"Cambodia"	32	184
4	"Canada"	85	8
5	"China"	94	43
6	"Czech_Republic"	99	12
7	"Egypt"	89	55
8	"Ethiopia"	13	208
9	"Finland"	95	7
10	"France"	95	9
11	"Greece"	54	9
12	"India"	89	124
13	"Italy"	95	10
14	"Japan"	87	6
15	"Mexico"	91	33
16	"Poland"	98	16
17	"Russian_Federation"	73	32
18	"Senegal"	47	145
19	"Turkey"	76	87
20	"United_Kingdom"	90	9

El objetivo es hallar una ecuación que represente lo más preciso posible la relación entre la variable independiente: el porcentaje de inmunización, y la variable dependiente:la tasa de mortalidad. El siguiente es un plot de los datos.

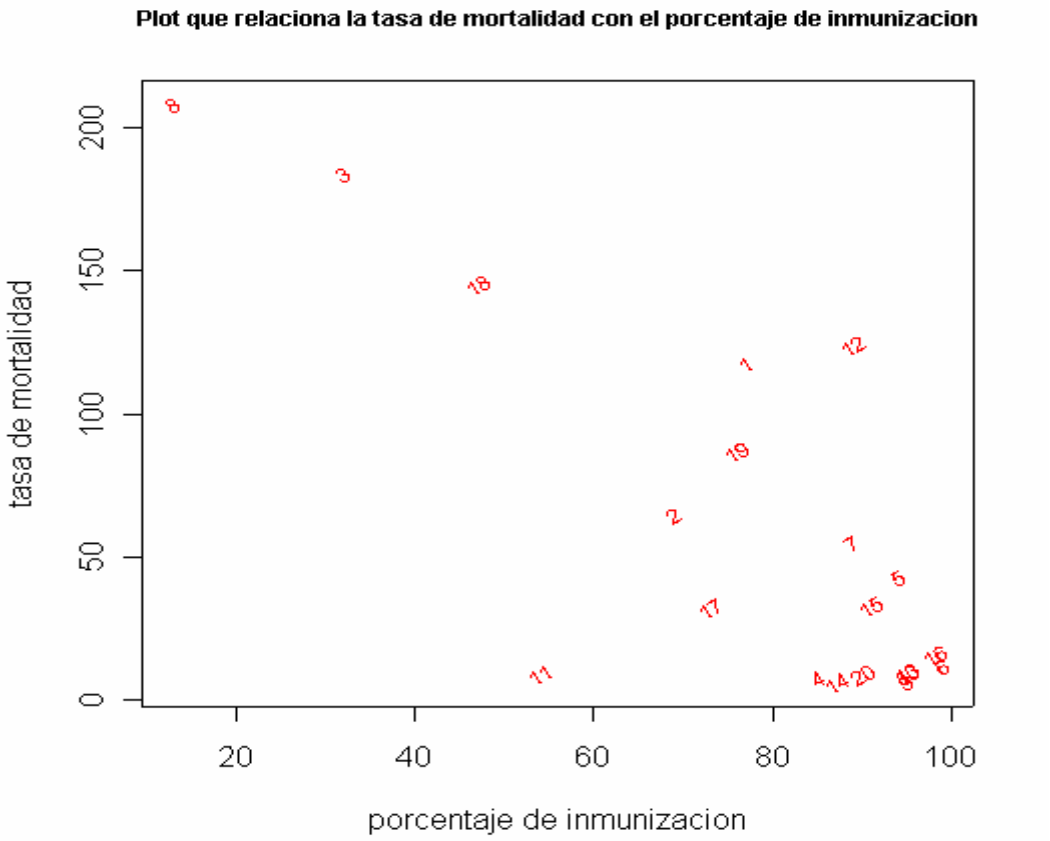


Figura 1. Plot que relaciona la tasa de mortalidad con el porcentaje de inmunización en cada país

De la figura 1 se puede ver que los países 8, 3 y 18 parecen estar algo alejados de la mayoría de los datos. Igualmente 11 y 12 aparecen algo fuera de la tendencia. No es muy obvio concluir que hay una relación lineal entre las variables. La figura 2 muestra la línea de regresión obtenida usando el programa R. Los comandos aparecen en el laboratorio 1 de la página de internet del texto.

Observando la siguiente salida obtenida en R para la regresión lineal correspondiente.

```
> l1<-lsfit(x,y)
> ls.print(l1)
Residual Standard Error=40.1393
R-Square=0.6258
F-statistic (df=1, 18)=30.1006
p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept 224.3163 31.4403  7.1347        0
X          -2.1359  0.3893 -5.4864        0
```

se tiene que la medida de confiabilidad del modelo, llamada **coeficiente de determinación (R^2)** , es sólo 62.58%, lo cual no es muy alto. Sin tomar en cuenta que esta medida se vee afectada por la presencia de los valores anormales, nos indica que la relación lineal entre las variables no es muy fuerte.

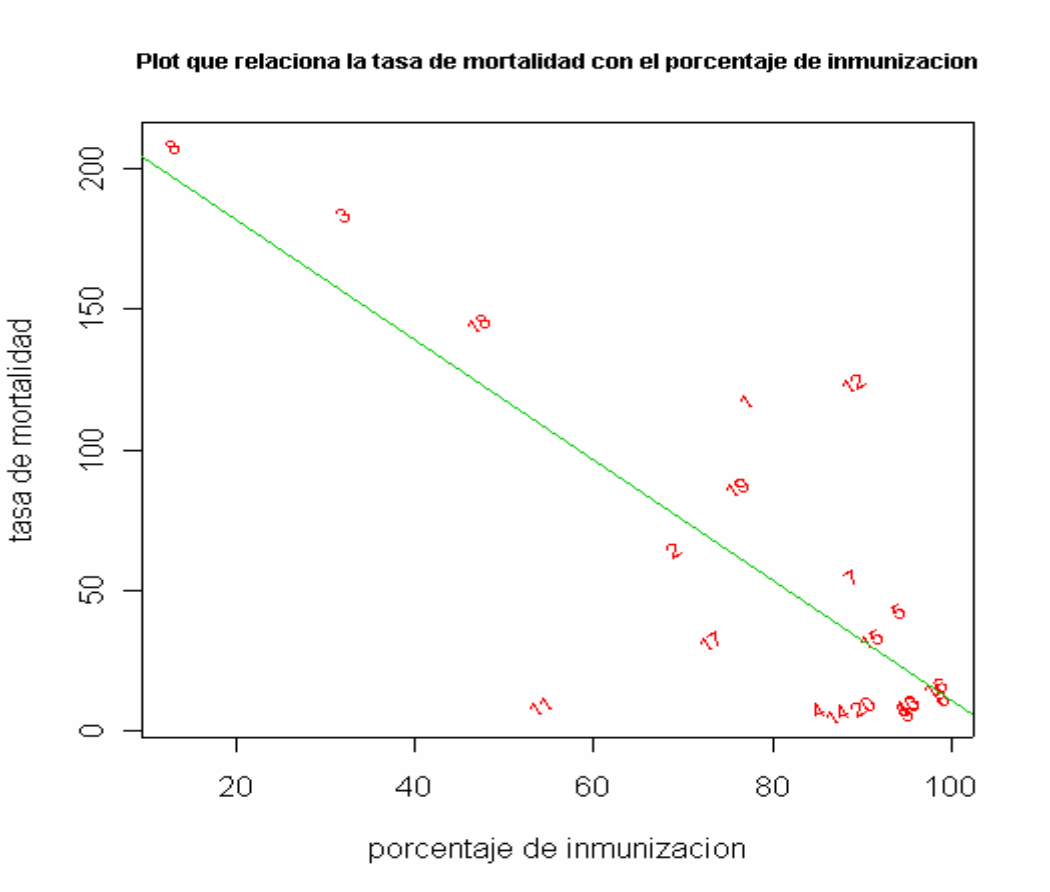


Figura 2. Linea de Regresión para los datos del ejemplo 1.

Si eliminamos las observaciones 11 y 12 la relación mejora notablemente, lo cual se puede ver en la siguiente salida de R

```
> l2<-lsfit(x1,y1)
> ls.print(l2)
Residual Standard Error=24.73
R-Square=0.8617
F-statistic (df=1, 16)=99.7027
p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept 251.4824 20.2188 12.4380       0
X          -2.4766  0.2480 -9.9851       0
```

Donde ahora el R^2 subió a un 86.2%, que es bastante aceptable. Asimismo en la figura 3 muestra la nueva linea de regresión que ajusta a mejor a los datos.

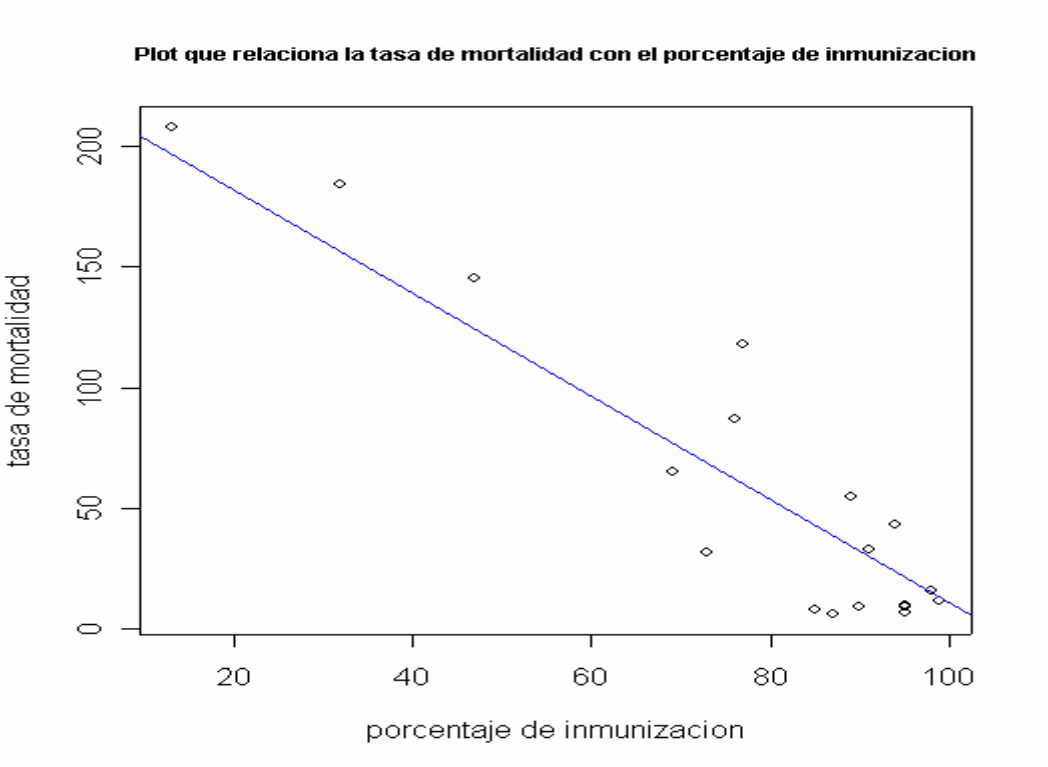


Figura 3. Linea de regresión despues de eliminar las observaciones atipícas 11 y 12

El análisis de regresión es un proceso interactivo y el desarrollo de las computadoras en la última década ha facilitado e incentivado el uso de regresión en el análisis estadístico.

Regresión también es conocido como **Ajuste por cuadrados mínimos**, debido al método que se usa para estimar el modelo de regresión. Cuadrados Mínimos es acreditado a Karl Gauss y data desde los inicios de 1800. El nombre regresión fue introducido por F. Galton a finales de 1800 cuando trató de relacionar las alturas de hijos y padres.

1.1.1 Usos del análisis de regresión:

Los siguientes son los usos de un modelo de regresión, aunque muchas veces ellos se dan al mismo tiempo:

a) **Predicción:** El objetivo aquí es pronosticar valores de la variable de respuesta para valores futuros de la variables predictoras, es decir para valores más allá de rango de valores de la variable predictora en la muestra de entrenamiento. Tal vez ésta sea la razón principal para usar regresión.
b) **Descripción:** La idea es establecer una ecuación que describa la relación entre la variable dependiente y las variables predictoras.
c) **Control:** Controlar el comportamiento o variación de la variable de respuesta de acuerdo a los valores de las variables predictoras.
d) **Selección de variables:** Inicialmente se pueden haber considerado muchas variables para explicar el comportamiento de la variable de respuesta, pero la presencia de muchas variables puede afectar el rendimiento del modelo además de que la computación del mismo se vuelve lenta. Por lo tanto hay que usar técnicas para escoger solo las variables predictoras que sean más relevantes y las que no sean redundantes en explicar la variación de la variable de respuesta.

1.2 El modelo de Regresión Lineal simple

En este caso se tiene una variable de respuesta o dependiente, denotada por Y y una sola variable predictora representada por X. El modelo de regresión lineal simple es de la forma

$$Y = \alpha + \beta X + \varepsilon \qquad (1.1)$$

Aqui α y β son el intercepto y la pendiente del modelo de regression respectivamente y ε es un error aleatorio. Considerando que la muestra es representada por los n pares ordenados (X_i,Y_i) entonces el modelo se puede escribir como

$$Y_i = \alpha + \beta X_i + e_i \quad \text{para i=1,…n} \qquad (1.2)$$

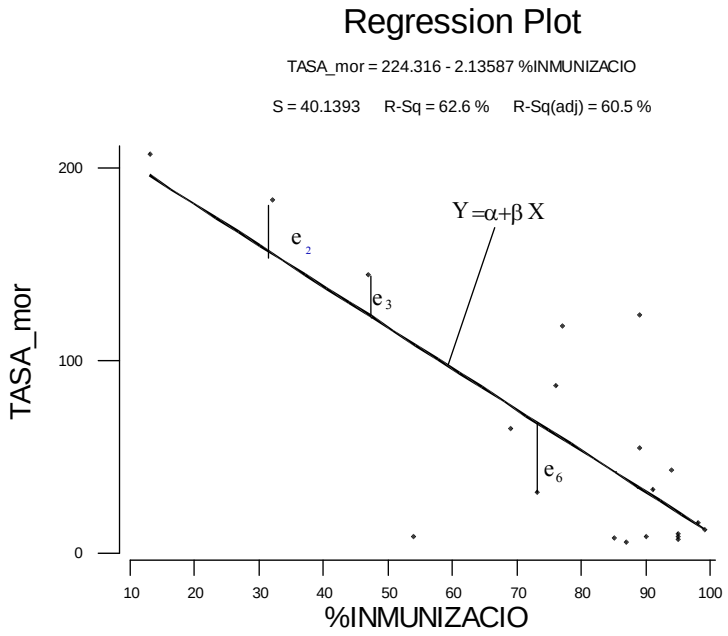


Figura 4. Errores con repecto a la linea de regresión para algunas de las observaciones del ejemplo 1

En la figura anterior se muestra la línea de regresión y los errores para algunas de las observaciones.

Suposiciones del modelo:

a) La variable predictora X es no aleatoria y se supone que ha sido medida con la mejor precisión posible. Sin embargo hay algunas situaciones donde también se supone que X es aleatoria .

b) Los errores e_i son variables aleatorias con media 0 y varianza constante σ^2. Por ahora no se requerirá normalidad de los errores.

c) Los errores e_i y e_j (i≠j=1…,n) son independientes entre si. Es decir, $Cov(e_i, e_j) = 0$

Como en la ecuación del modelo solamente los e_i's son aleatorios entonces las y_i's deben tener también varianza constante σ^2 y deben ser independientes por parejas.

1.2.1 Estimación de la línea de regresión usando Mínimos Cuadrados

Si se toma el valor esperado de y_i para el valor x_i de x entonces de (1.2) se obtiene

$$E(y_i) = E(\alpha + \beta x_i + e_i) = \alpha + \beta x_i \qquad (1.3)$$

O más formalmente que

$$E(y / x) = \alpha + \beta x \qquad (1.4)$$

Es decir, la esperanza (o media) condicional de y dado x es una ecuación lineal en x. Los parámetros α y β deben ser estimados en base a la muestra tomada. El método usual para estimarlos es el de los cuadrados mínimos. La idea es minimizar la suma de los cuadrados de los errores e_i, con respecto a α y β. Es decir,

$$Q(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \qquad (1.5)$$

Derivando parcialmente $Q(\alpha, \beta)$ con respecto a α y β e igualando a cero se obtienen las siguientes ecuaciones

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \qquad (1.6)$$

$$\frac{\partial Q}{\partial \beta} = -2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0 \qquad (1.7)$$

simplificando ambas ecuaciones se obtiene

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (1.8)$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (1.9)$$

este par de ecuaciones es conocido como las **ecuaciones normales del modelo**. Resolviendo este par de ecuaciones se obtiene que

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad (1.10)$$

lo cual es equivalente a $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$

donde: $S_{xy} = \sum_{i=1}^{n} x_i y_i - \dfrac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$ es llamada la suma de productos corregida y

$S_{xx} = \sum_{i=1}^{n} x_i^2 - \dfrac{(\sum_{i=1}^{n} x_i)^2}{n}$ es la llamada suma de cuadrados corregidos de X.

De la primera ecuación normal es fácil ver que:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (1.11)$$

Por la forma de Q, es natural pensar que en el punto $(\hat{\alpha}, \hat{\beta})$ hay un mínimo. Más formalmente, se podría aplicar el criterio de la segunda derivada para máximos y mínimos de la función bivariada. En este caso habría que cotejar que:

$Q_{\alpha\alpha}(\alpha,\beta)>0$, y que $D = Q_{\alpha\alpha}(\alpha,\beta)Q_{\beta\beta}(\alpha,\beta) - (Q_{\alpha\beta}(\alpha,\beta))^2 > 0$

como $Q_{\alpha\alpha}(\alpha,\beta) = 2n > 0$ y

$D = 4n\sum_{i=1}^{n} x_i^2 - 4(\sum_{i-1}^{n} x_i)^2 = 4n\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0$, las condiciones requeridas se cumplen .

Finalmente la línea de regresión estimada o la línea ajustada por cuadrados mínimos será:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \qquad (1.12)$$

Una vez que se ajusta la línea de regresión, el error aleatorio se vuelve un valor observado y es llamado ***residual*** , el cual es representado por r_i o por $\hat{e}_i$.

Sustituyendo el valor de $\hat{\alpha}$ en la ecuación anterior se tiene:

$$\hat{y} = \bar{y} + \hat{\beta}(x_i - \bar{x}) \qquad (1.13)$$

Esta ecuación puede ser considerada como la estimación de un modelo de regresión donde la variable predictora ha sido centrada.

El método de Máxima verosimilitud también puede ser usado pra estimar los coeficientes de la línea de regresión pero se necesita considerar la suposición de que los errores aleatorios e_i se distribuyen normalmente con media cero y varianza σ^2. En este caso las estimaciones se obtienen maximizando la función $L(\alpha,\beta,\sigma^2) = \prod_{i=1}^{n} f(y_i - \alpha - \beta x_i)$, donde f representa la función de densidad de una normal $N(0,\sigma^2)$. Notar que en este caso además de las estimaciones de los coeficientes α y β, también se hace la estimación de la varianza σ^2.

Las propiedades de los estimadores mínimo cuadráticos de los coeficientes de regresión se discuten en las secciones 1.2.3 y 1.2.5.

1.2.2 Interpretación de los coeficientes de regresión estimados

La pendiente $\hat{\beta}$ indica el cambio promedio en la variable de respuesta cuando la variable predictora aumenta en una unidad adicional. El intercepto $\hat{\alpha}$ indica el valor promedio de la variable de respuesta cuando la variable predictora vale 0. Sin embargo carece de interpretación práctica si es irrazonable pensar que el rango de valores de x incluye a cero.

En el ejemplo 1, la ecuación de la línea de regresión estimada es

Tasa_mort = 224.316 – 2.13587%inmuniz,

lo que significa que en promedio la tasa de mortalidad de niños menores de 5 años disminuirá en promedio en 2.13 cuando el % de inmunizacion aumenta en uno porciento.
Por otro lado la tasa de mortalidad promedio de los paises donde no hay inmunización será de 224.316. Aunque es difícil pensar que exista un país donde no se vacunen a los niños, ya que muchas veces la UNICEF dona las vacunas.

1.2.3 Propiedades de los estimadores mínimos cuadráticos de regresión

a) $\hat{\beta}$ es un estimador insegado de β. Es decir, E($\hat{\beta}$)=β

Recordar que:

$\hat{\beta} = \dfrac{Sxy}{Sxx} = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. Luego, como x no es variable aleatoria y $E(y_i) = E(\alpha+\beta x_i+e_i) = \alpha+\beta x_i$, por suposición b) del modelo, se obtiene que:

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})E(y_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\alpha\sum_{i=1}^{n}(x_i-\bar{x}) + \beta\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (1.14)$$

como las suma de las desviaciones con respecto a la media es cero, se sigue que:

$$E(\hat{\beta}) = \beta\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta\frac{Sxx}{Sxx} = \beta$$

b) $\hat{\alpha}$ es un estimador insegado de α. Es decir, E($\hat{\alpha}$)=α

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (1.15)$$

Luego,

$$E(\hat{\alpha}) = E(\bar{y}) - E(\hat{\beta})\bar{x} = E(\bar{y}) - \beta\bar{x} =$$

$$E(\frac{\sum_{i=1}^{n} y_i}{n}) - \beta\bar{x} = \frac{1}{n}\sum_{i=1}^{n} E(y_i) - \beta\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \beta\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha \quad (1.16)$$

c) La varianza de $\hat{\beta}$ es $\frac{\sigma^2}{Sxx}$ y la de $\hat{\alpha}$ es $\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{Sxx})$

Usando la propiedad que Var(cy)=c^{2}Var(y) y el hecho que la suposición de que Cov(e$_i$,e$_j$)=0 es equivalente a Cov(y$_i$,y$_j$)=0, se tiene que Var($\sum_{i=1}^{n} c_i y_i$)= $\sum_{i=1}^{n} c_i^2 Var(y_i)$. En consecuencia,

$$Var(\hat{\beta}) = Var(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(y_i)}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} = \sigma^2 \frac{Sxx}{(Sxx)^2} = \frac{\sigma^2}{Sxx} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1.17)$$

Por otro lado notar que $\hat{\alpha}$ puede ser reescrita de la siguiente manera

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \sum_{i=1}^{n}[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{Sxx}]y_i \,. \quad (1.18)$$

Luego,

$$Var(\hat{\alpha}) = \sigma^2\sum_{i=1}^{n}[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{Sxx}]^2 = \sigma^2\sum_{i=1}^{n}[\frac{1}{n^2} - \frac{2\bar{x}(x_i - \bar{x})}{nSxx} + \frac{\bar{x}^2(x_i - \bar{x})^2}{(Sxx)^2}] \quad (1.19)$$

el segundo término de la suma se cancela y finalmente se obtiene que

$$Var(\hat{\alpha}) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2 Sxx}{(Sxx)^2}] = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}] \quad (1.20)$$

Hay otra forma de calcular la varianza de $\hat{\alpha}$, usando el hecho que $Cov(\bar{y}, \hat{\beta}) = 0$. Las propiedades discutidas en esta sección serán usadas cuando se haga inferencia estadística para el modelo de regresión.

1.2.4 Distribución de los estimadores mínimos cuadráticos

Para efecto de hacer inferencia en regresión, se requiere asumir que los errores e_i, se distribuyen en forma normal e independientemente con media 0 y varianza constante σ^2. En consecuencia, también las y_i's se distribuyen normalmente con media $\alpha + \beta x_i$ y varianza σ^2.

En el cálculo de los valores esperados de $\hat{\alpha}$ y $\hat{\beta}$ se estableció que estos son una combinación lineal de las y_i's. Esto es que $\hat{\alpha} = \sum_{i=1}^{n} a_i y_i$ y $\hat{\beta} = \sum_{i=1}^{n} b_i y_i$ .Por lo tanto, usando el hecho que una combinación lineal de variables aleatorias normales e independientes también se distribuye normalmente, y los resultados de la sección 1.2.3 se puede establecer que:

i) $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$ (1.21)

ii) $\hat{\alpha} \sim N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2)$

1.2.5 Propiedades de los residuales

Los residuales $r_i = y_i - \widehat{y}_i$ son las desviaciones de los valores observados de la variable de respuesta con respecto a la línea de regresión. Los residuales representan los errores aleatorios observados, y satisfacen las siguientes propiedades:

a) La suma de los residuales es 0. Es decir, $\sum_{i=1}^{n} r_i = 0$

En efecto, $\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i) = \sum_{i=1}^{n} y_i - n\widehat{\alpha} - \widehat{\beta} \sum_{i=1}^{n} x_i = 0$. La última igualdad se justifica por la primera ecuación normal.

b) $\sum_{i=1}^{n} r_i x_i = 0$. Similarmente, a la propiedad a) se tiene

$\sum_{i=1}^{n} r_i x_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) x_i = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i) x_i = \sum_{i=1}^{n} x_i y_i - \widehat{\alpha} \sum_{i=1}^{n} x_i - \widehat{\beta} \sum_{i=1}^{n} x_i^2 = 0$. La última igualdad se justifica por la segunda ecuación normal.

c) $\sum_{i=1}^{n} r_i \widehat{y}_i = 0$. Claramente, $\sum_{i=1}^{n} r_i \widehat{y}_i = \sum_{i=1}^{n} r_i (\widehat{\alpha} + \widehat{\beta} x_i) = \widehat{\alpha} \sum_{i=1}^{n} r_i + \widehat{\beta} \sum_{i=1}^{n} r_i x_i = 0$. La última igualdad se justifica por a) y b).

1.2.6 Estimación de la varianza del error

La varianza del error, representada por σ^2 es desconocida y debe ser estimada usando los residuales. Un estimador insesgado de σ^2 es

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2} \qquad (1.22)$$

s^2 es llamado también **el cuadrado medio del error.** Existe una fórmula alterna para calcular s^2, pero esta será discutida más adelante cuando se haga el análisis de varianza para regresión simple.

Verificación de que $E(s^2)= \sigma^2$

Notar que

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \hat{y}_i)y_i - \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i \qquad (1.23)$$

Usando la propiedad c) de los residuales, la segunda de las sumas anteriores se cancela y usando las propiedades a) y b) se tiene que

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)y_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)(\alpha + \beta x_i + e_i) = \sum_{i=1}^{n}(y_i - \hat{y}_i)e_i \qquad (1.24)$$

Por otro lado,

$$(y_i - \hat{y}_i) = (\alpha + \beta x_i + e_i) - (\hat{\alpha} + \hat{\beta} x_i) = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_i + e_i \qquad (1.25)$$

Asímismo,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = (\alpha + \beta\bar{x} + \bar{e}) - \hat{\beta}\bar{x} = \alpha + (\beta - \hat{\beta})\bar{x} + \bar{e} \qquad (1.26)$$

Sustituyendo (1.26) en (1.25) se obtiene que

$$(y_i - \hat{y}_i) = (\beta - \hat{\beta})(x_i - \bar{x}) + e_i - \bar{e} \qquad (1.27)$$

Reemplazando (1.27) en (1.24) se llega a

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)e_i = \sum_{i=1}^{n}[(\beta - \hat{\beta})(x_i - \bar{x})e_i + e_i^2 - e_i\bar{e}]$$

Tomado valores esperados en la última expresión y sustituyendo en la ecuación (1.23) se consigue

$$E[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2]=\sum_{i=1}^{n}[(x_i-\bar{x})E((\beta-\hat{\beta})e_i)+E(e_i^2)-E(e_i\bar{e})] \quad (1.28)$$

Usando la suposiciones del modelo de regresión lineal es fácil ver que $E(e_i^2)=\sigma^2$ y que

$$E(e_i\bar{e})=E(e_i\frac{\sum_{j=}^{n}e_j}{n})=E(\frac{e_i^2}{n})=\frac{\sigma^2}{n}. \quad (1.29)$$

Por otro lado, de la fórmula para $\hat{\beta}$ se obtiene lo siguiente

$$\hat{\beta}=\frac{Sxy}{Sxx}=\frac{\sum_{j=1}^{n}(x_j-\bar{x})y_j}{Sxx}=\frac{\sum_{j=1}^{n}(x_j-\bar{x})(\alpha+\beta x_j+e_j)}{Sxx}=\beta+\frac{\sum_{j=1}^{n}(x_j-\bar{x})e_j}{Sxx}$$

Por lo tanto

$$E[(\beta-\hat{\beta})e_i]=-\frac{E\sum_{j=1}^{n}(x_j-\bar{x})e_je_i}{Sxx}=-\frac{(x_i-\bar{x})\sigma^2}{Sxx} \quad (1.30)$$

Finalmente, sustituyendo (1.29) y (1.30) en (1.28) se obtiene,

$E[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2]=-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2}{Sxx}+n\sigma^2-\sigma^2=(n-2)\sigma^2$, con lo cual concluye la prueba.

1.2.7 Descomposición de la suma de cuadrados total

Lo que se va hacer aquí es tratar de descomponer la variación total de Y en dos partes, una que se deba a la relación lineal de Y con X y otra a causas no controlables. Lo ideal es que gran parte de la variación de Y se explique por su relación lineal con X.

De la siguiente gráfica se puede ver que la desviacion de un valor observado y_i con respecto a la media $\bar{y}$ se puede escribir como

$$(y_i-\bar{y})=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y}) \quad (1.31)$$

Elevando al cuadrado en ambos lados de 1.31 y sumando sobre todas las observaciones se obtiene

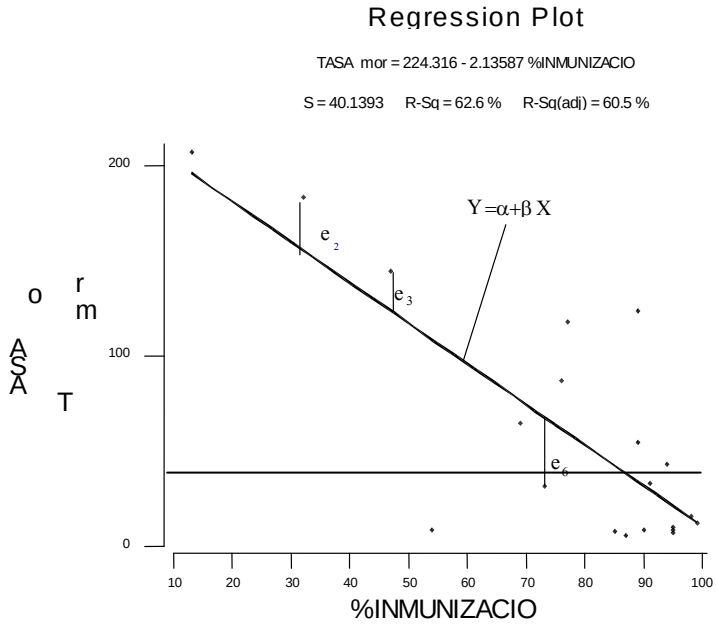


Figure 5. Diagrama para descomponer la deviación total en desviacion debido a la regresion mas desviacion debido al error.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) \qquad (1.32)$$

La suma de productos del lado derecho se puede escribir como,

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i-\bar{y}\sum_{i=1}^{n}(y_i-\hat{y}_i) \qquad (1.33)$$

la primera de las sumas es 0 por la propiedad c) de los residuales y la segunda es 0 por la propiedad a) de los residuales. En consecuencia (1.32) se reduce a

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad (1.34)$$

donde

$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$ es llamada la suma de cudrados del total y representa la variación total de las

y's.

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ es llamada la Suma de Cuadrados del Error o Residual y representa la variación de las y's que se debe a causas no controlables. Notar que el estimado de la varianza poblacional s^2, puede ser calculado por $s^2 = \frac{SSE}{n-2}$

$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ es llamada la Suma de Cuadrados debido a la Regresión y representa la variación de la y's que es explicada por su relación lineal con X. Sustituyendo $\hat{y}_i$ por $\hat{\alpha} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$ se tiene que

$$SSR = \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1.35)$$

Las sumas de cuadrados definidas anteriormente son variables aleatorias, pues dependen de y, la cual es aleatoria. Ellas van a jugar un papel muy importante cuando se haga inferencia en regresión, por eso es importante saber como es su distribución. Por teoría de modelos lineales se puede establecer que las sumas de cuadrados son formas cuadráticas de las variables y_i y por lo tanto se distribuyen como una Ji-cuadrado. Más especificamente, se pueden establecer los siguientes resultados:

i). $\frac{SST}{\sigma^2} \sim \chi'^2_{(n-1)}$ (Ji-cuadrado no central con n-1 grados de libertad).Los grados de libertad se pueden establecer de la fórmula de cálculo de SST, pues en ella se usan *n* datos, pero en ella aparece un valor estimado ($\bar{y}$) por lo tanto se pierde un grado de libertad.

ii). $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$ (Ji-cuadrado con n-2 grados de libertad). Para calcular *SSE* se usan *n* datos pero hay presente un estimado $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, cuyo cálculo depende a su vez de dos estimaciones. Por lo tanto se pierden dos grados de libertad. También se puede escribir que $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{(n-2)}$

iii). $\frac{SSR}{\sigma^2} \sim \chi'^2_{(1)}$ (Ji-cuadrado no central con 1 grado de libertad y parámetro de nocentralidad $\frac{\beta^2 S_{xx}}{\sigma^2}$). De la ecuación (1.35) se puede notar que el cálculo de *SSR* envuelve el cuadrado de una variable distribuida normalmente. Por un resultado de Estadística Matemática se sabe que el cuadrado de una normal estándar es una Ji-cuadrado con un grado de libertad.

Por otro lado, tomando valor esperado en cada lado de relación (1.35) se tiene

$$E(SSR) = S_{xx}E(\hat{\beta}^2) = S_{xx}[Var(\hat{\beta}) + (E(\hat{\beta}))^2] = S_{xx}(\frac{\sigma^2}{S_{xx}} + \beta^2)$$

Luego,

$$E(SSR)=\sigma^2+\beta^2 S_{xx} \qquad (1.36)$$

1.2.8 El Coeficiente de Determinación R^2

Es una medida de la bondad de ajuste del modelo y se define como

$$R^2 = \frac{SSR}{SST} * 100\%$$

Un modelo de regresión con R^2 mayor o igual a 75% se puede considerar bastante aceptable. Aunque se puede ser un poco flexible dependiendo del tipo de datos y de la cantidad de datos disponible. En el ejemplo 1 sólo un 62.6% de la variacion de la mortalidad infantil de niños menores de 5 años es explicada por su relación lineal con el porcentaje de inmunización, lo cual no es muy alto y hace poco confiable las predicciones. Lamentablemente, el valor de R^2 es afectado por la presencia de valores anormales. Asi, un valor de R^2 bien cercano al 100% no necesariamente garantiza una buena predicción del modelo. Pero si se puede decir que un modelo con R^2 bajo es inadecuado para hacer predicciones.

1.3 Inferencia en Regresión Lineal Simple

En esta sección discutirá pruebas de hipótesis e intervalos de confianza acerca de los coeficientes de regresión del modelo de regresión poblacional. También se construirán intervalos de confianza de las predicciones y del valor medio de la variable de respuesta.

1.3.1 Inferencia acerca de la pendiente y el intercepto usando la prueba t.

Inferencia acerca de la pendiente de la línea de regresión se discutirá detalladamente, en lo que respecta al intercepto será tratado brevemente. Como se ha visto en la sección 1.2.8 si se asume que las y_i's tienen una distribución normal para cada valor de la variable predictora x entonces el estimado $\widehat{\beta}$ de la pendiente de regresión se distribuye como una normal con media β y varianza $\frac{\sigma^2}{Sxx}$. Esto es equivalente a decir, que el estadístico $z = \frac{\hat{\beta} - \beta}{\frac{\sigma}{\sqrt{Sxx}}}$ se distribuye como una normal estándar, N(0,1).

Desafortunadamente, este estadístico no se puede usar en la práctica, pues por lo general σ es desconocida. Por otro lado, también sabemos que el estadístico $\chi^2_{(n-2)} = \frac{(n-2)s^2}{\sigma^2}$ se distribuye como una Ji-cuadrado con n-2 grados de libertad. Por un resultado de Estadística Matemática y probando previamente que hay independencia entre $\widehat{\beta}$ y s^2, se tiene que

$$t = \frac{z}{\sqrt{\frac{\chi^2_{(n-2)}}{n-2}}} = \frac{\hat{\beta} - \beta}{\frac{s}{\sqrt{Sxx}}} \qquad (1.37)$$

se distribuye como una t de Student con n-2 grados de libertad. El estadístico t es usado para hacer prueba de hipótesis y calcular intervalos de confianza acerca de β.

Un intervalo de confianza del 100(1-α)% para la pendiente poblacional β es de la forma

$$(\widehat{\beta} - t_{(n-2,\alpha/2)} \frac{s}{\sqrt{Sxx}}, \widehat{\beta} + t_{(n-2,\alpha/2)} \frac{s}{\sqrt{Sxx}})$$

donde α, que varía entre 0 y 1, es llamado el nivel de significación, $t_{(n-2,\alpha/2)}$ es un valor t tal que el área debajo de la curva y a la derecha de dicho valor es igual a α/2. La expresión $\frac{s}{\sqrt{Sxx}}$ es llamada el error estándar (propiamente es un estimado) de $\widehat{\beta}$. Muy raros son los programas estadísticos que muestran, en sus salidas de análisis de regresión, intervalos de confianza para la pendiente, solamente dan el $\widehat{\beta}$ y su error estándar . Hay que calcular $t_{(n-2,\alpha/2)}$ usando cálculos de percentiles (por computadora o en tablas) y luego se calcula la fórmula del intervalo.

Ejemplo 2: Para los datos del ejemplo 1. Calcular un intervalo de confianza del 95% para la pendiente poblacional.
Solución. Usando el laboratorio 2 en R que aparece en la página de internet del texto se obtienen los siguientes resultados

```
> summary(l2)

Call:
lm(formula = tasa.mort ~ porc.inmuniz, data = muertes)

Residuals:
     Min       1Q   Median       3Q      Max
-99.97934 -16.57854   0.06684  20.84946  89.77608

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  224.3163    31.4403   7.135 1.20e-06 ***
porc.inmuniz  -2.1359     0.3893  -5.486 3.28e-05 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 40.14 on 18 degrees of freedom
Multiple R-Squared: 0.6258,     Adjusted R-squared: 0.605
F-statistic:  30.1 on 1 and 18 DF,  p-value: 3.281e-05
```

Notar que $\widehat{\beta}$ =-2.1359 y que su error estándar es 0.3893. Los grados de libertad del la t son 20-2=18 y el α=0.05 , luego hay que buscar el percentil $t_{(.025,18)}$. Este percentil, o su simetrico correspondiente, puede ser obtenido usando el comando de R, qt(.975,18), el cual da un valor de 2.1009. Usando nuevamente el laboratorio 2 resulta

```
> # Hallando el intervalo de confianza del 95% para la pendiente Beta
> bint<-c(beta-qt(.975,18)*eebeta,beta+qt(.975,18)*eebeta)
> bint
```

[1] -2.95290 -1.31890
Luego, el Intervalo de confianza del 95% para β será

(-2.95290, -1.31890)

Por lo tanto, hay un 95% de confianza de que la pendiente de regresión poblacional caiga enter –2.95 y –1.32.

Haciendo una discusión análoga al caso de la pendiente se puede llegar a establecer que un intervalo de confianza del 100(1-α)% para el intercepto α de la linea de regresión poblacional es de la forma

$$(\widehat{\alpha} - t_{(n-2,\alpha/2)} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}}, \widehat{\alpha} + t_{(n-2,\alpha/2)} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}})$$

Ahora se considerará prueba de hipótesis en regresión. Desde el punto de vista clásico, la siguiente tabla muestra la manera de hacer pruebas de hipótesis para la pendiente β, asumiendo que su valor es β*

Caso I	Caso II	Caso III
Ho: β=β* Ha: β<β*	Ho: β=β* Ha: β≠β*	Ho: β=β* Ha: β>β*
Prueba Estadística		
$t = \frac{\widehat{\beta} - \beta^*}{\frac{s}{\sqrt{Sxx}}} \sim t_{(n-2)}$		
Regla de Decisión		
Rechazar Ho, si $t_{cal} < -t_{(\alpha,n-2)}$	Rechazar Ho, si $\lvert t_{cal} \rvert > t_{(\alpha/2,n-2)}$	Rechazar Ho, si $t_{cal} > t_{(\alpha,n-2)}$

Obviamente el caso más importante es el caso II cuando β*=0. Porque de rechazarse la hipótesis nula sugeriría de que hay relación lineal entre las variables X y Y. En la manera clásica uno rechaza o acepta la hipótesis nula comparando el valor de la prueba estadística con un valor obtenido de la tabla de t para un nivel de significación dado, usualmente de 0.05 ó 0.01.
A inicios de los años 80's y con la ayuda de los programas de computadoras se comenzó a probar hipótesis usando la técnica del "P-value", que es el nivel de significación observado, es decir, el valor de α al cual se rechazaría la hipótesis nula si se usaría el resultado que dá la prueba estadística. Un "P-value" cercano a cero, sugiriría rechazar la hipótesis nula. Sin embargo, existe un consenso en la mayoría de los autores a rechazar la hipótesis nula si el "P-value" es menor de 0.05.

Ejemplo 3: Para los datos del ejemplo 1, probar la hipótesis de que la pendiente poblacional es cero.

Solución: Usando los resultados del laboratorio 2 de R.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	224.3163	31.4403	7.135	1.20e-06	***

```
porc.inmuniz  -2.1359     0.3893  -5.486 3.28e-05 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 40.14 on 18 degrees of freedom
Multiple R-Squared: 0.6258,     Adjusted R-squared: 0.605
```

Las hipótesis serían:
Ho: $\beta=0$ (es decir no hay relación lineal entre las variables)
Ha: $\beta\neq 0$ (hay relación lineal entre las variables)

Se observa que el "P-value" correspondiente a porcentaje de inmunización es 0.0000328<0.05 . Por lo tanto se concluye que hay relación lineal entre las variables, aunque no se puede decir aún que tan fuerte es esta relación.

Similarmente se pueden hacer pruebas de hipótesis para el intercepto.

Las hipótesis serían:
Ho: $\alpha=0$ (La línea de regresión poblacional pasa por el origen)
Ha: $\alpha\neq 0$ (La línea de regresión poblacional no pasa por el origen)

Como el "P-value" es 0.000012<0.05 se concluye que hay suficiente evidencia de que la línea de regresión poblacional NO pasa por el origen.

1.3.2 El análisis de varianza para regresión lineal simple

El análisis de varianza para regresión consiste en descomponer la variación total de la variable de respuesta en varias partes llamadas fuentes de variación. Como se vió en la sección 1.2.7, para el caso de regresión lineal solo hay dos fuentes: Una variación debido a la Regresión y otra variación debido al error. Cada variación es cuantificada por una suma de cuadrados, las cuales como se mencionó anteriormente tienen una distribución Ji-cuadrado.
La división de la suma de cuadrados por sus grados de libertad es llamada **cuadrado medio**.
Así se tienen tres cuadrados medios

Cuadrado Medio de Regresión=MSR=SSR/1
Cuadrado Medio del Error= MSE=SSE/(n-2)
Cuadrado Medio del Total=MST=SST/(n-1),

Pero este último no es usado. Notar también que $MSE=s^2$.

Por otro lado, en la sección 1.2.6, se ha demostrado que $E[MSE]=\sigma^2$ y en la ecuación 1.36 de la sección 1.2.7 se tiene que $E[MSR]= \sigma^2 +\beta^2 S_{xx}$. Si estuviesemos probando la hipótesis Ho:$\beta=0$ y ésta fuera cierta entonces $E[MSR]= \sigma^2$, y su distribución pasa a ser una Ji-Cuadrado (central) con 1 grado de libertad. Luego, tanto MSE como MSR estimarían a la varianza poblacional.
Por resultados de Estadística Matemática se puede mostrar que la división de dos Cuadrados medios independientes se distribuye como una F. Más precisamente,

$$F = \frac{MSR}{MSE} \sim F_{(1,n-2)}$$

siempre que la hipótesis nula Ho:β=0 es cierta. Aquí el numerador tiene 1 grado de libertad y el denominador tiene n-2. La independencia descansa en el hecho que $COV(\hat{Y}_i - \overline{Y}, \hat{Y}_i - Y_i) = 0$

Todos los cálculos se resumen en la siguiente tabla llamada **tabla de Análisis de Varianza**

Fuente de Variacion	**g.l.**	**Sumas de Cuadrados**	**Cuadrados Medios**	**F**
Debido a la Regresion	1	SSR	MSR=SSR/1	$\frac{MSR}{MSE}$
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	n-1	SST		

Desde el punto de vista clásico la hipótesis Ho:β=0 se rechazaría en favor de Ho:β≠0 si el valor de la prueba de F es mayor que $F_{(\alpha,1,n-2)}$. En la manera moderna de probar hipótesis se rechazaría la hipótesis nula si el "P-value" de la prueba de F es menor de 0.05.

Para los datos del ejemplo 1, la tabla de análisis de varianza obtenida al correr el programa del laboratorio 2 será como sigue:

```
> anova(l2)
Analysis of Variance Table

Response: tasa.mort
             Df Sum Sq Mean Sq F value    Pr(>F)
porc.inmuniz  1  48497   48497  30.101 3.281e-05 ***
Residuals    18  29001    1611
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Claramente se rechaza la hipótesis nula pues el p-value da 0.0000328. Notar que, $t^2_{(n-2)} = F_{(1,n-2)}$.

1.3.3 Intervalo de confianza para el valor medio de la variable de respuesta e Intervalo de Predicción

Talvez el uso más frecuente que se le da a una línea de regresión es para hacer predicciones acerca de la variable de respuesta *Y* para un valor dado de x. Supongamos que queremos predecir el valor medio de las *Y* para un valor x_0 de la variable predictora x. Es decir, $E(Y / x = x_0) = \alpha + \beta x_0$.

Es natural pensar que el estimado puntual sera $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$. Sin embargo, es muy riesgoso predecir basado en un solo valor y es más conveniente usar un intervalo donde se espera que caiga el valor de *Y* con un cierto grado de confianza.Como $\hat{\alpha}$ y $\hat{\beta}$ se distribuyen normalmente, entonces $\hat{y}_0$ también se distribuye normalmente con media $\alpha + \beta x_0$ y varianza igual a:

$$Var(\hat{Y}_0) = Var(\hat{\alpha} + \hat{\beta} x_0) = Var(\hat{\alpha}) + x_0^2 Var(\hat{\beta}) + 2x_0 Cov(\hat{\alpha}, \hat{\beta})$$

Sustituyendo expresiones halladas en la sección 1.2.3 y el hecho que $Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\overline{x}\sigma^2}{Sxx}$, se tiene:

$$Var(\hat{Y}_0) = \sigma^2 (\frac{1}{n} + \frac{\overline{x}^2}{Sxx}) + x_0^2 \frac{\sigma^2}{Sxx} + 2X_O (\frac{-\overline{x}\sigma^2}{Sxx})$$

de donde resulta

$$Var(\hat{Y}_0) = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx})$$

En consecuencia, estandarizando y sustituyendo la σ por s se tendrá que:

$$\frac{\hat{y}_0 - E(Y/x_0)}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}} \sim t_{(n-2)}$$

Usando el resultado previo se puede establecer que un intervalo de confianza del 100(1-α)% para el valor medio de las y's dado que x=x_0 es de la forma

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2,n-2)}s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}} \qquad (1.38)$$

Pero frecuentemente uno está interesado en estimar un valor individual de Y dado x=x_0 y no un promedio de valores. Evidentemente, que hay un mayor riesgo de hacer de esto. La predicción del valor individual $Y_0 = \alpha + \beta x_0 + e_0$, es también $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0$. Trabajando con la diferencia $Y_0 - \hat{Y}_0$, se puede ver fácilmente que E($Y_0 - \hat{Y}_0$)=0 y que

$$Var(Y_0 - \hat{Y}_0) = Var(Y_0) + Var(\hat{Y}_0) - 2Cov(Y_0, \hat{Y}_0)$$

Luego,

$$Var(Y_0 - \hat{Y}_0) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}) - 2Cov(Y_0, \hat{Y}_0)$$

como Y_0 y $\hat{Y}_0$ son nocorrelacionados, $Cov(Y_o, \hat{Y}_o) = 0$. Entonces,

$$Var(Y_0 - \hat{Y}_0) = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx})$$

Haciendo cálculos similares a cuando se obtuvo el intervalo de confianza para el valor medio, se puede estableser que un intervalo de confianza de 100(1-α)% (mas conocido como intervalo de predicción) para un valor individual de Y dado x=x_0 es de la forma

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2,n-2)}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}} \qquad (1.39)$$

Notar que este intervalo es más amplio que el intervalo de confianza, pues la varianza estimada incluye un termino adicional. Muchos programas estadísticos calculan unas curvas que se obtienen uniendo los limites superiores e inferiores de los intervalos de confianza (o de predicción) para varios valores de la variable predictora, y estas son llamadas **Bandas de confianza (o Bandas de predicción).**

Ejemplo 4: a) Hallar un intervalo de confianza del 99% para la tasa de mortalidad promedio de niños menores de 5 años en los paises cuyo porcentaje de inmunizacion es 80%. Hallar un intervalo de predicción del 95% para la tasas de mortalidad de niños menores de 5 años en los paises cuyo porcentaje de inmunización sea del 80%.

Solución: Usando nuevamente los resultados producidos por el programa del laboratorio 2 se obtiene los siguiemtes resultados.

```
> predict(l2,porc.inmuniz,se.fit=T,interval=c("confidence"),level=.99)
$fit
       fit     lwr      upr
[1,] 53.44674 27.44776 79.44572

$se.fit
[1] 9.032315

$df
[1] 18

$residual.scale
[1] 40.13931

> predict(l2,porc.inmuniz,se.fit=T,interval=c("prediction"),level=.95)
$fit
       fit     lwr      upr
[1,] 53.44674 -32.9915 139.8850

$se.fit
[1] 9.032315

$df
[1] 18

$residual.scale
[1] 40.13931
```

Interpretación: Hay un 99% de confianza de que la tasa de mortalidad media de todos los paises con porcentaje de inmunización del 80% caiga entre 27.45 y 79.45 y la tasa de mortalidad de un país, cuyo porcentaje de inmunización es 80% caerá entre –32.99 y 139.88 con un 95% de confianza.

La siguiente figura muestra las bandas de confianza y predicción del 95 porciento para los datos del ejemplo 1.

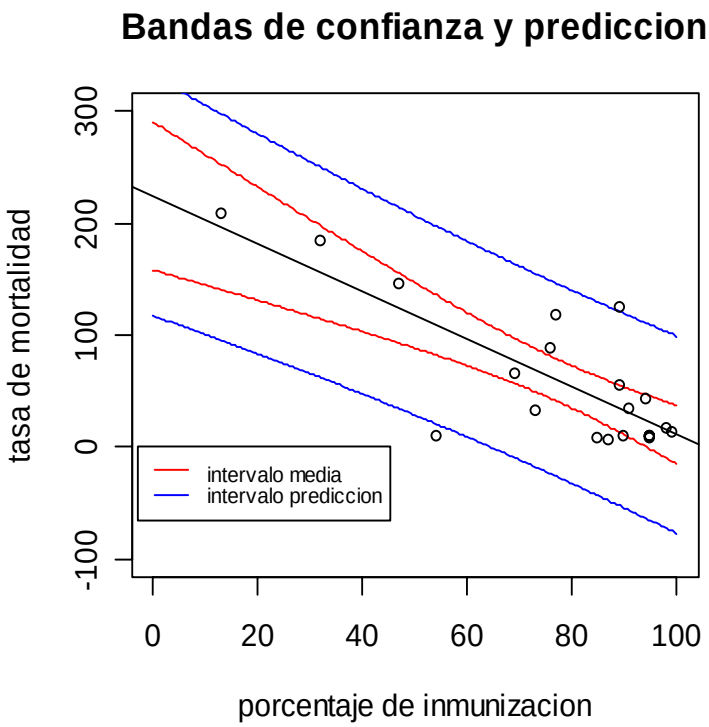


Figure 6: Bandas de confianza y predicción para los datos del ejemplo 1.

1.4 Análisis de residuales

Los residuales, que son estimaciones de los errores del modelo, son importantes para establecer si las suposiciones del modelo se cumplen y para explorar el porqué de un mal ajuste del modelo. La manera más fácil de examinar los residuales es mediante plots los cuales permiten cotejar:

- Si la distribución de los errores es normal y sin "outliers".
- Si la varianza de los errores es constante y si se requieren transformaciones de las variables.
- Si la relación entre las variables es efectivamente lineal o presenta algún tipo de curvatura
- Si hay dependencia de los errores, especialmente en el caso de que la variable predictora sea tiempo.

Existen varios tipos de residuales, por ahora solo introduciremos dos:

i) Residual Estandarizado: En este caso se divide el residual entre la desviación estándar del error. Es decir,

$$\text{Residual estandarizado}=\frac{y_i-\widehat{y}_i}{s}$$

ii) Residual Estudentizado: En este caso se divide el residual entre su desviación estándar estimada. Notar que,

$$Var(y_i - \hat{y}_i) = Var(y_i) + Var(\hat{y}_i) - 2Cov(y_i, \hat{y}_i)$$

Usando resultados de la sección 1.3.3, lo anterior se puede escribir como

$$Var(y_i - \hat{y}_i) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{Sxx}) - 2Cov(y_i, \bar{y} + \hat{\beta}(x_i - \bar{x}))$$

calculando la covarianza, se obtiene

$$Var(y_i - \hat{y}_i) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{Sxx}) - 2\sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{Sxx})$$

En consecuencia,

$$Var(y_i - \hat{y}_i) = \sigma^2(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{Sxx})$$

Por lo tanto,

$$r_i^* \frac{y_i - \hat{y}_i}{s\sqrt{(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{Sxx})}}$$

En algunos programas estadísticos, los r_i^* son llamados residuales estandarizados. También son llamados residuales estudentizados internamente (ver más adelante, la sección 3.1).

1.4.1 Cotejando normalidad de los errores y detectando outliers

Normalidad de los errores es un requisito indispensable para que tengan validez las pruebas estadísticas de t y F que se usan en regresión. Existen varios métodos gráficos y pruebas estadísticas tanto paramétricas cono noparamétricas para cotejar la normalidad de un conjunto de datos. La manera más fácil es usando gráficas tales como histogramas, "stem-and-leaf" o "Boxplots".
Una gráfica más especializada es el plot de Normalidad. Aquí se plotea los residuales versus los valores que se esperarían si existiera normalidad, estos valores son llamados los scores normales. Dado el i-ésimo residual, su score normal se encontraría determinando primero a que percentil le corresponde en la distribución de los datos, se han propuesto varias maneras de hacer esto. Luego de determinar la percentil se halla el valor que le corresponde a dicha percentil en la distribución normal estándar. Habrá normalidad si los puntos del plot se alinean cerca de una linea que pasa por el origen. Si se usan los residuales estudentizados la línea además de pasar por el origen deberia tener pendiente cercana a 1.

Ejemplo 5. Cotejar si existe normalidad para los datos del ejemplo1.

Considerando los residuales estudentizados y la funciones **hist** y **boxplot** de R se obtiene el histograma y "boxplot" correspondientes.

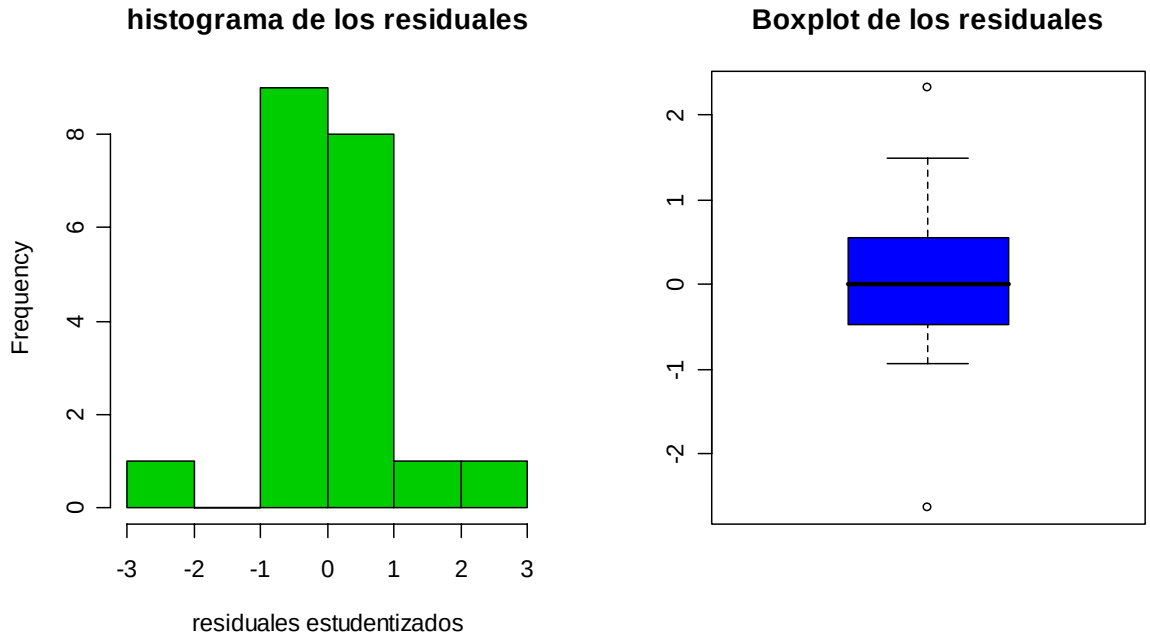


Figura 7. Histograma y boxplot de residuales de la regresion del ejemplo 1.

El histograma no parece ser de forma acampanada, es decir no hay normalidad, además parece haber un "outlier" inferior. El boxplot indica bastante simetría en el centro pero no en los extremos de la distribución. Además se identifican dos outliers, uno superior y el otro inferior.

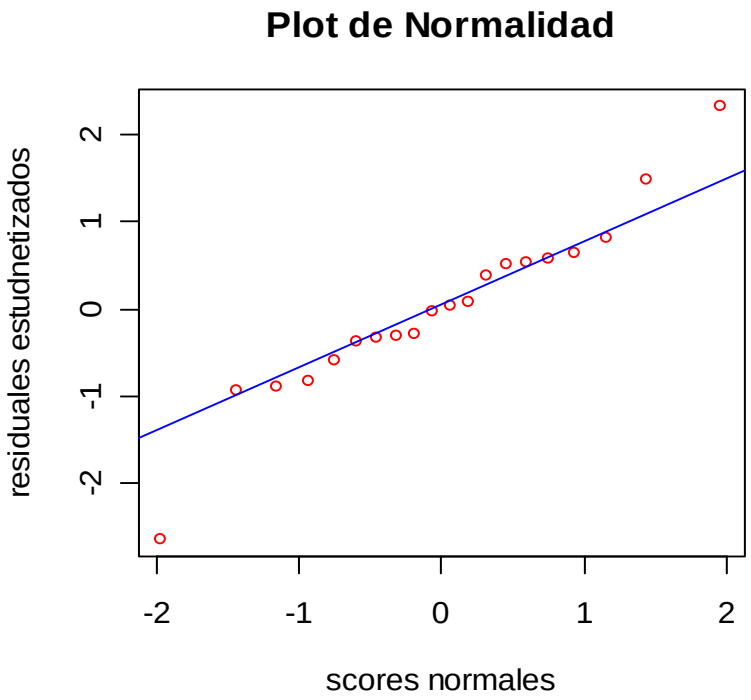


Figura 8. Plot de Normalidad para los residuales de la regresion del ejemplo 1.

En el plot de normalidad los puntos siguen una tendencia bastante lineal, especialmente en el centro. Pero lo que es más notorio es la presencia de un "outlier" inferior y dos probables "outliers" superiores.
Otra manera de detectar si hay "outliers" es cotejando si los residuales estudentizados son mayores que 2 en valor absoluto. En el capítulo dedicado a regresión múltiple se hará una discusión más detallada de los criterios para detectar "outliers".

1.4.2 Cotejando que la varianza sea constante

En este caso se plotea los residuales estandarizados versus los valores ajustados o versus la variable predictora X. No se plotea versus las y_i observadas porque los residuales y las y_i 's se espera que estén correlacionadas.
Si los puntos del plot caen en una franja horizontal alrededor de 0 entonces la varianza de los errores es constante. Si los puntos siguen algún patrón entonces se dice que la varianza de los errores no es constante.

Ejemplo 6: Hacer un plot de residuales para cotejar si hay varianza constate de los errores para los datos del ejemplo 1.

Siguiendo la misma secuencia de comandos del ejemplo anterior, y eligiendo las gráficas adecuadas obtenemos los siguientes plots

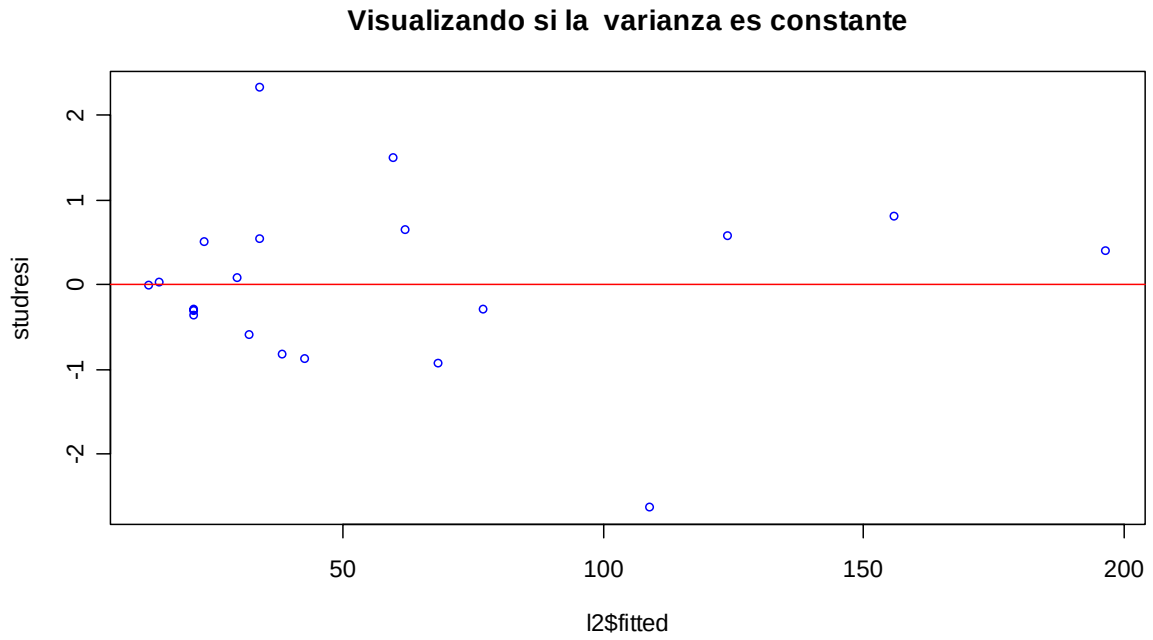


Figura 9. Plot de residuales para detectar si la varianza constante con respecto a los valores ajustados

Los puntos se reparten equitativamente alrededor de la línea horizontal. Nuevamente lo que llama más la atención es la presencia del "outlier". Por lo tanto, la varianza parece ser constante. Si ploteamos los residuales versus la variable predictora en lugar de los valores ajustados, se obtiene la siguiente gráfica

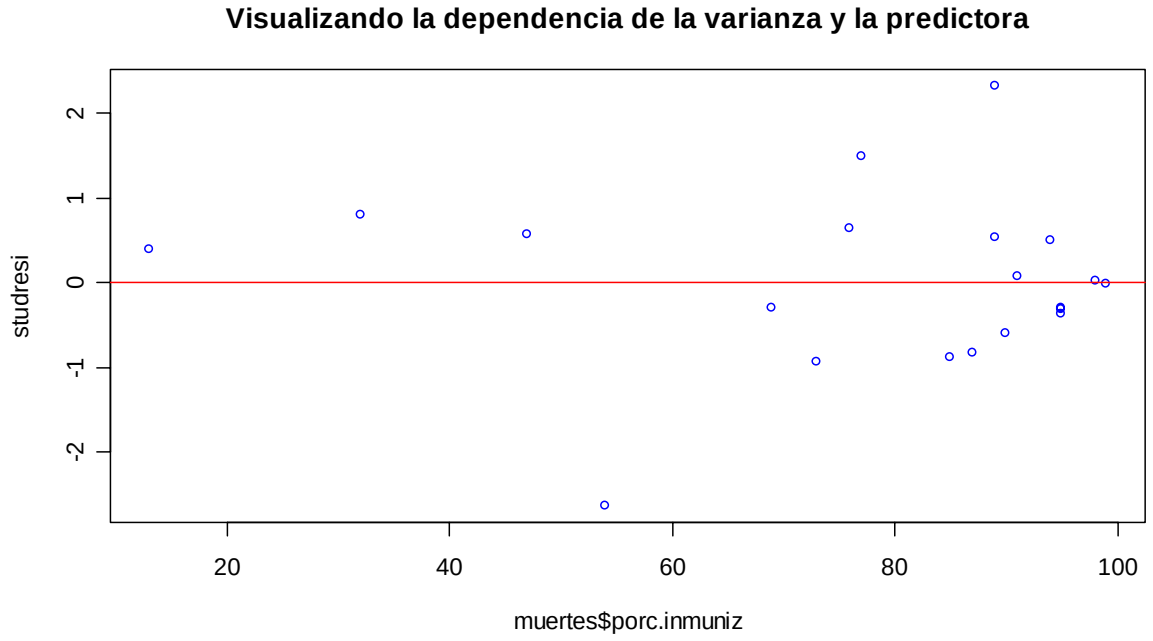


Figura 10. Plot de residuales para detectar si la varianza constante con respecto a la variable predictora

Nuevamente, excluyendo los dos "outliers", los puntos parecen estar en una franja horizontal, por lo tanto se podría considerar que la varianza es constante con respecto a la predictora. Notar que también hay cuatro puntos alejados en la dirección horizontal. Estas observaciones también pueden tener influencia en los cálculos de la línea de regresión.
Si se observa algún patrón en el plot se puede hacer transformaciones en una o en ambas variables para estabilizar la varianza. Otra alternativa es usar ***mínimos cuadrados ponderados***. Nuevamente esto será discutido más detalladamente en el capítulo 3 del texto cuando se haga análisis de residuales en regresión múltiple.

En R se puede hacer un plot simultáneo de los residuales. Usando el laboratorio 3 de R para el ejemplo 1 se obtiene la siguiente gráfica.

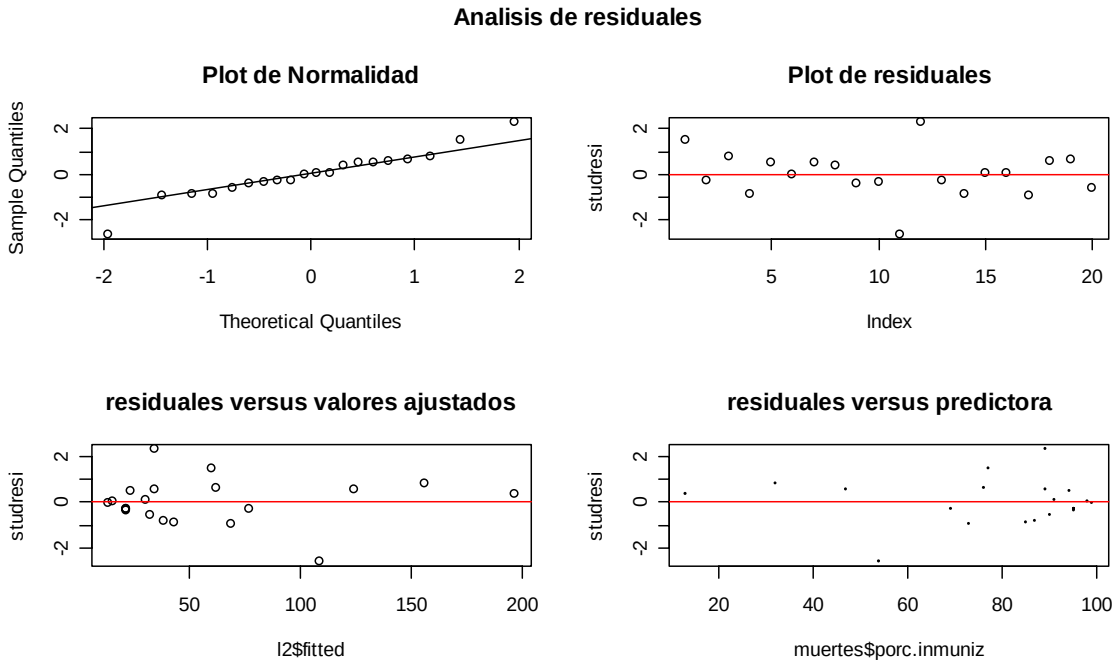


Figura 11: Plots para hacer análisis de residuales

1.4.3 Cotejando si los errores estan correlacionados.

Cuando la variable predictora es tiempo, puede ocurrir que los errores esten correlacionados secuecialmente entre si. Si en el plot de residuales versus valores ajustados se observa un patrón cíclico entonces hay correlación entre los errores.
Existe también la prueba de Durbin-Watson que mide el grado de correlación de un error con el que anterior y el posterior a él . El estadístico es

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Notar que D es aproximadamente igual a 2(1-r), donde r representa la correlación lineal entre los errores e_i's y e'_{i-1}'s. Usando ese hecho se puede mostrar que D varía entre 0 y 4. Si D es cerca de 0 los errores están correlacionados positivamente. Si D está cerca de 4 entonces la correlación es negativa. Además la distribución de D es simétrica con respecto a 2. Así que un valor de D cercano a 2 indica que no hay correlación de los errores. Más formalmente hay que comparar el valor de D con dos valores criticos D_L y D_U de una tabla.

Aplicando la function **dw** del laboratorio 4 de R a los datos del ejemplo 1 resulta

```
el estadistico Durbin Watson de la regresion lineal es= 2.678912
```

Como el valor está cerca de 2, se concluirá que no hay correlación entre los errores. También se puede ver en el plot de residuales, que no hay un patrón cíclico de los puntos.

1.5 El Coeficiente de Correlación

Algunas veces se considera que tanto la variable de respuesta como la predictora son aleatorias. Por ejemplo si se quiere relacionar horas de estudio (X) y nota en un examen (Y). La manera estándar sería establecer de antemano los posibles números de horas que se va a considerar y luego para cada una de las horas elegir al azar por lo menos un estudiante y registrarle su nota. Sin embargo, también se puede elegir al azar un estudiante y hacerle las dos preguntas: Cuántas horas estudió? y qué nota obtuvo en el examen?. En este caso (X,Y) se comporta como una variable aleatoria bivariada, que generalmente se distribuye como una normal bivariada. Una Normal Bivariada tiene cinco parámetros: las medias μ_x, μ_y, las desviaciones estándares σ_x y σ_y y el coeficiente de correlación ρ. El coeficente de correlación, es un valor que mide el grado de asociacón lineal entre las variables X y Y y se define como

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{1.40}$$

Se puede mostrar que

a) $-1 \le \rho \le 1$

b) La media condicional de Y dado X es $E(Y/X) = \alpha + \beta x$. Donde $\beta = \rho \dfrac{\sigma_y}{\sigma_x}$, y $\alpha = \mu_y - \beta \mu_x$. Notar que si la pendente de la línea de regresión es cero entonces la correlación es 0, y que β y ρ varian en la misma dirección.

c) La varianza condicional de las Y dado X, está dado por $\sigma^2_{y/x} = \sigma^2_y (1-\rho^2)$. Luego, si $\rho = \pm 1$, entonces $\sigma^2_{y/x} = 0$, implicando que hay una perfecta relación lineal entre Y y X. Más especificamente, si $\rho = 1$, entonces X y Y crecen en la misma dirección y si $\rho = -1$, Y decrece cuando X crece.

Todo lo anterior ocurre en la población, así que ρ es un parámetro que debe ser estimado. Suponiendo que se ha tomado una muestra de n pares (x_i,y_i), entonces, el **coeficiente de correlación muestral** se calcula por

$$r = \frac{Sxy}{\sqrt{SxxSyy}} \tag{1.41}$$

Notar que $r = \widehat{\beta}\sqrt{\frac{Sxx}{Syy}}$ y que $r^2 = \frac{\widehat{\beta}^2 Sxx}{Syy} = \frac{SSR}{SST}$. Es decir, que el cuadrado del coeficiente de correlación es igual al coeficiente de determinación. Al igual que el parámetro poblacional, la correlación muestral varía entre –1 y 1. Por lo general, un r mayor de 0.75 en valor absoluto es considerado acceptable, aunque algunas veces debido a la naturaleza de los datos hay que exigir un valor más alto, digamos mayor de 0.90.

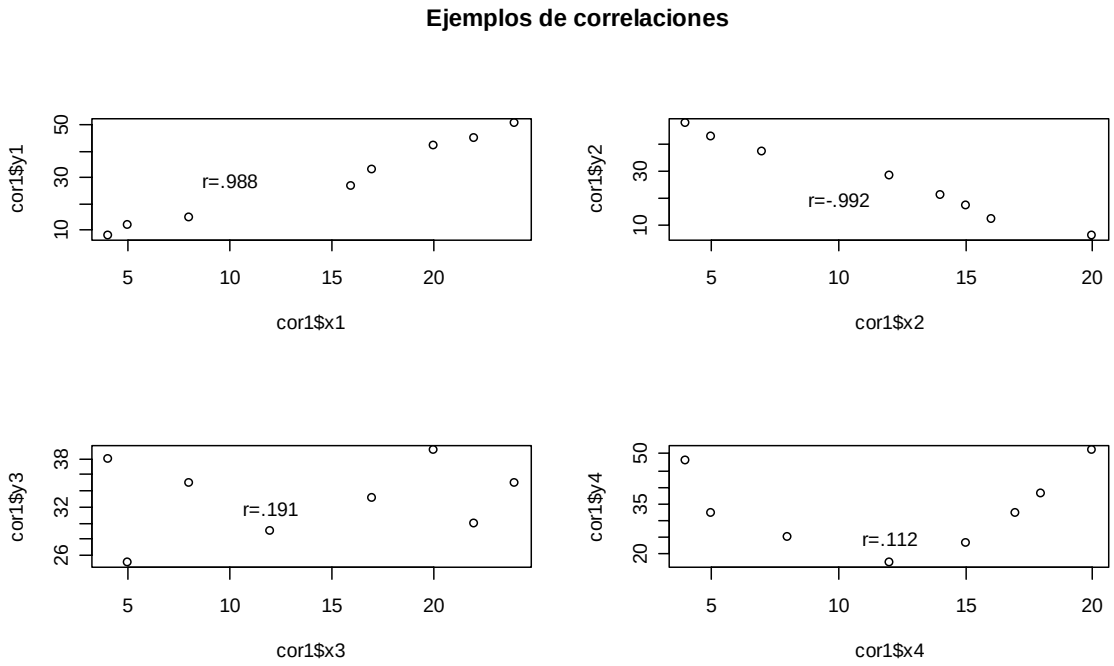
En R, el commando cor permite calcular la correlación entre dos o mas variables. Para el ejemplo 1, los resultados son:

```
> cor(muertes$tasa.mort,muertes$porc.inmuniz)
[1] -0.7910654
>
```

El valor de la correlación en valor absoluto es algo mayor de 0.75, lo que implicaría una aceptable relación lineal entre las variables, además cuando el porcentaje de inmunización aumenta la tasa de mortalidad disminuye.

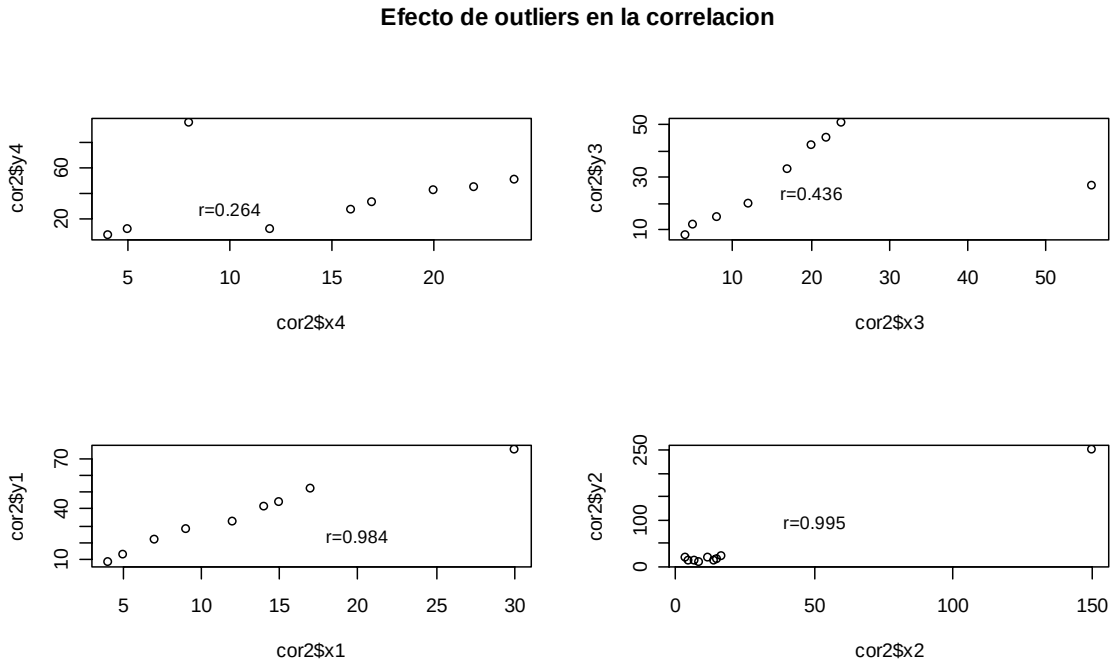
Advertencia: ***Correlación alta no implica necesariamente una relación causa efecto entre las variables. Usando la fórmula de correlación entre dos variables que en la vida real no tiene ninguna relación entre si, (por ejemplo X: peso de los profesores y Y=salario del profesor) se puede obtener un r bastante alto cercano a 1 o –1 y eso no implica necesariamente que X explique el comportamiento de Y (podria darse el caso que mientras menos pesa un profesor gana menos).***

La siguiente gráfica muestra varios diagramas de puntos y sus respectivas correlaciones.



Notar que en los dos últimos plots la correlación es cercana a cero, pero en el primer caso no parece haber ningún tipo de relación entre las variables y en el otro no hay relación lineal pero si existe una relación cuadrática.

El valor de correlación es afectado por la presencia de valores anormales, en la siguiente gráfica se puede ver el efecto de los valores anormales en el valor de la correlación para 4 diferentes relaciones



Interpretación de la figura: *En el primer caso existe un valor bastante anormal en la dirección vertical que hace que la correlación sea bastante bajo a pesar de que los otros valores parecen estar bastante alineados.*

En el segundo caso existe un valor bastante alejado horizontalmente de la mayor parte de los datos y que hace que la correlación sea relativamente baja a pesar de que los otros valores muestran una alta asociación lineal.

En el tercer caso hay una observación bastante alejado en ambas direcciones sin embargo no tiene ningún efecto en la correlación, cuyo valor de por sí es alto.

En el cuarto caso hay un valor bastante alejado en ambas direcciones y las restantes observaciones están poco asociadas, pero el valor anormal hace que el valor de la correlación sea bastante alto.

Debido a la relación entre la pendiente de la línea de regresión y el coefiente de correlación, la prueba estadística para probar Ho:ρ=0 (la correlación poblacional es cero) versus Ha: ρ≠0 (hay correlación entre las poblaciones X e Y) es similar a la prueba de la pendiente de la linea de Regresión: Es decir,

$$t = \frac{\hat{\beta}}{\frac{s}{\sqrt{Sxx}}} = \frac{r\sqrt{\frac{Syy}{Sxx}}}{\sqrt{\frac{Syy(1-r^2)}{n-2}}{Sxx}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)} \qquad (1.42)$$

La prueba estadística para probar Ho:ρ=ρ_o (la correlación poblacional es de magintud ρ_o) versus Ha:ρ≠ρ_o involucra el uso de una transformación del coeficiente de correlación muestral, llamada la ***transformación z de Fisher***, ya que la distribución de r no es normal y tiende ser asimétrica para valores grandes de ρ. La transformación está definida por

$$z = \frac{1}{2}\log(\frac{1+r}{1-r}) = \tanh^{-1}(r) \qquad (1.43)$$

la cual tiene una distribución aproximadamente normal con media

$$E(z) = \frac{1}{2}\log(\frac{1+\rho}{1-\rho}) + \frac{\rho}{2(n-1)}$$

y varianza $Var(z) = \frac{1}{n-3}$. La aproximación es bastante buena si n>50. En consecuencia, la prueba estadística será:

$$Z = \sqrt{n-3}\left[\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\log\left(\frac{1+\rho_o}{1-\rho_o}\right) - \frac{\rho_o}{2(n-1)}\right]$$

EJERCICIOS

1. Considerando un modelo de regresión lineal simple, calcular $Cov(\overline{Y}, \hat{\beta})$

2. Probar que la línea de regresión estimada pasa por $(\overline{X}, \overline{Y})$

3. En un modelo de regresión lineal simple calcular E[SST]=E[$\sum_{i=1}^{n}(y_i - \bar{y})^2$]

4. **Regresión que pasa por el origen.** Algunas veces se conoce de anteamano que la linea de regresión pasa por el origen. Es decir el modelo es de la forma $y_i=\beta x_i+e_i$.
   a) Hallar el estimador por cuadrados mínimos de β. Cuál es su varianza?
   b) Hallar el estimador de la varianza poblacional σ^2
   c) Establecer la formula para un intervalo de confianza del 100(1-α)% de confianza para β

5. Probar que $\text{Cov}(\hat{\alpha}, \hat{\beta}) = \dfrac{-\overline{x}\sigma^2}{Sxx}$

6. En un estudio del desarrollo del conocimiento se registra la edad (X) en meses) a la que 21 niños dicen su primera palabra y el puntaje en la prueba de Gessell (Y), un test de habilidad que toma posteriormente el niño. Los resultados son como siguen

Edad	Puntaje	Edad	Puntaje
15	95	9	96
26	71	10	83
10	83	11	84
9	91	11	102
15	102	10	100
20	87	12	105
18	93	42	57
10	100	17	121
8	104	11	86
20	94	10	100
7	113		

a) Hallar la linea de regresión. e interpretar los coeficientes de la linea de regresión
b) Trazar la linea de regresión encima del diagrama de puntos.
c) Probar la hipótesis de que la pendiente es cero. Comentar su resultado
d) Interpretar el coeficiente de determinación R^2
e) Hallar un intervalo de confianza del 99% para la pendiente de la linea de regresión poblacional
f) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 95% para el valor individual y valor medio de la variable de respuesta e intrepretar el resultado.

7. En un pueblo se eligen 15 personas al azar y se anota su salario mensual (X) y la cantidad que ahorran mensualmente (Y).

Salario	Ahorro
800	150
850	100
900	280
1200	400

1500	350
1700	500
1900	635
2000	600
2300	750
2500	680
2700	900
3000	800
3200	300
3500	1200
5000	1000

a) Hallar la linea de regresión. e interpretar los coeficientes de la linea de regresión
b) Trazar la linea de regresión encima del diagrama de puntos.
c) Interpretar el coeficiente de determinación
d)Probar la hipótesis de que la pendiente es cero. Comentar su resultado
e) Hallar un intervalo de confianza del 95% para la pendiente de regresión poblacional.
f) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 para el valor individual y el valor medio de la variable de respuesta e intrepretar el resultado.

8. Leer el conjunto de datos **brain** que aparece en la página de internet del texto y considerar las variables:

MRI (X), conteo en pixels del 18 scans de resonancia magnetica del cerebro de una persona
Score_IQ, (Y) score en un test de inteligencia.

Mientras más alto sea el conteo de pixels mas grande es el cerebro de las personas.

a) Hallar la linea de regresión ajustada. e interpretar los coeficientes de la linea de regresión
b) Trazar la linea de regresión encima del diagrama de puntos.
c) Probar la hipótesis de que la pendiente es cero (usando las pruebas t y F). Comentar su resultado
d) Interpretar el Coeficiente de Determinación.
e) Hallar un intervalo de confianza del 99% para la pendente de la regresion poblacional e interpretar su resultado
f) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 porciento para el valor individual y el valor medio de la variable de respuesta e intrepretar el resultado.

9.
a) Si Y=3.5-1.5X , SST=219 y SSE=59, hallar e interpretar el valor de la correlación entre X y Y
b) Considerando los datos dados en a) y que la muestra de entrenamiento consiste de 36 datos, hallar el valor de la prueba estadística para probar que la pendiente de regresión es cero.

10. Considerando un modelo de regresión lineal simple, calcular

$$Cov\,(Y_i - \hat{Y}_i, \hat{Y}_i - \overline{Y})$$

11. Probar que el coeficiente de correlación muestral r cae entre –1 y 1.

12. Suponga que en el modelo de regresión lineal simple los valores x_i y y_i son reemplazados por ax_i+b y cy_i+d respectivamente donde a,b,c y d son constantes tales que $a\neq 0$ y $c\neq 0$. Cuál es el efecto de estas transformaciones en $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$, R^2 y la prueba estadística para probar la hipótesis nula Ho:β=0.?

13. Considere el modelo de regresión lineal simple $Y = \alpha + \beta X + \varepsilon$, donde tanto X como Y y ε son variables aleatorias con varianzas σ_x^2, σ_y^2 y σ_ε^2 respectivamente y σ_{xy} representa la covarianza entre X y Y. En la estimación mínimo cuadrática de α y β se minimiza la suma de cuadrados de las **distancias verticales** de las observaciones a la linea ajustada. En **Regresión Ortogonal** la estimación de α y β se hace considerando que la línea es ajustada de tal manera que se minimiza la **distancia** más corta de las observaciones a la la linea ajustada. Hallar los estimadores de los coeficientes de la regresión ortogonal.

CAPÍTULO 2

REGRESIÓN LINEAL MULTIPLE

2.1 Introducción

Es evidente que lo más económico y rápido para modelar el comportamiento de una variable Y es usar una sola variable predictora y usar un modelo lineal. Pero algunas veces es bastante obvio de que el comportamiento de Y es imposible que sea explicada en gran medida por solo una variable. Por ejemplo, es imposible tratar de explicar el rendimiento de un estudiante en un examen, teniendo en cuenta solamente el número de horas que se preparó para ella. Claramente, el promedio académico del estudiante, la carga académica que lleva, el año de estudios, son tres de las muchas otras variables que pueden explicar su rendimiento. Tratar de explicar el comportamiento de Y con más de una variable predictora usando una funcional lineal es el objetivo de regresión lineal multiple.

Frecuentemente, uno no es muy familiar con las variables que están en juego y basa sus conclusiones solamente en cálculos obtenidos con los datos tomados. Es decir, si ocurre que el coeficiente de determinación R^2 sale bajo (digamos menor de un 30%) , considerando además que su valor no se ha visto afectado por datos anormales, entonces el modelo es pobre y para mejorarlo hay tres alternativas que frecuentemente se usan:

a) Transformar la variable predictora, o la variable de respuesta Y, o ambas y usar luego un modelo lineal.
b) Usar regresión polinómica con una variable predictora.
c) Conseguir más variables predictoras y usar una regresión lineal múltiple.

En el primer caso, se puede perder el tiempo tratando de encontrar la transformación más adecuada y se podria caer en sobre-ajuste (***"overfitting"***), es decir, encontrar un modelo demasiado optimista, que satisface demasiado la tendencia de los datos tomados pero que es pobre para hacer predicciones debido a que tiene una varianza grande.
En el segundo caso el ajuste es más rápido, pero es bien fácil caer en ***"overfitting"*** y, además se pueden crear muchos problemas de cálculo ya que pueden surgir problemas de colinealidad, es decir relación lineal entre los terminos del modelo polinómico.
El tercer caso es tal vez la alternativa más usada y conveniente. Tiene bastante analogía con el caso simple, pero requiere el uso de vectores y matrices.
En el siguiente ejemplo se mostrará el uso interactivo de las tres alternativas a través de seis modelos de regresión y servirá como un ejemplo de motivación para introducirnos en regresión lineal multiple.

Ejemplo 1: Considerar el conjunto de datos **millaje** donde la variable de respuesta esY= (MPG) millas promedio por galón de un auto, y las variables predictotras son; X_1=(VOL): Capacidad en volumen del carro, X_2=(HP): Potencia del Motor, X_3= (SP) :Velocidad Máxima y X_4=(WT): Peso del auto. Los datos fueron adaptados de la "Data and Story Library" y están dsponibles en la página de internet del texto.

Primero, se explorará las relaciones entre todas las parejas de variables, en particular la relación de Y con cada una de las variables predictoras. Esto se logra con una gráfica lllamada **plot matricial**, la cual está disponible en la mayoría de programas estadísticos. La function **pairs** de **R**

produce el plot matricial para las variables del ejemplo 1 tal como se muestra en la siguiente figura:

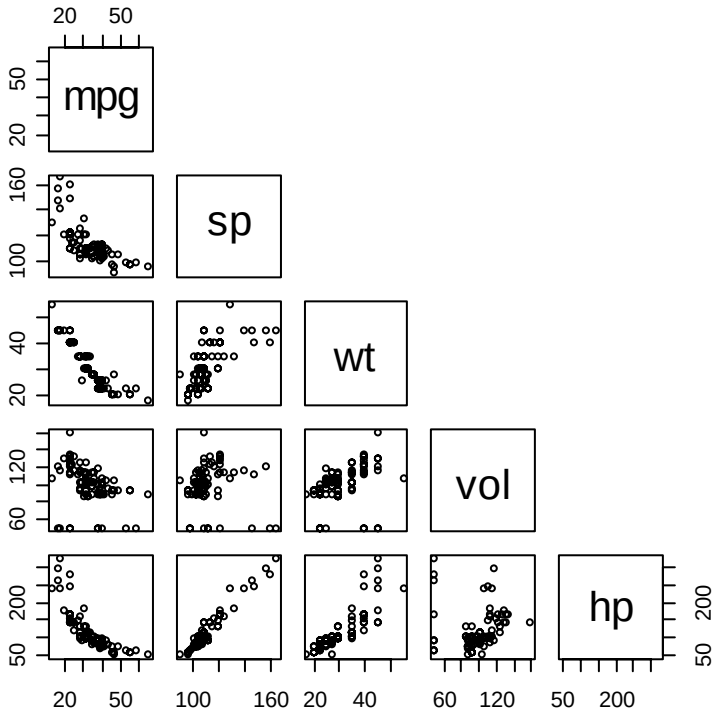


Figura 2.1. Plot matricial de las variables del conjutnode datos **millaje.**

Claramente se puede ver que la variable predictora (WT) es la que tiene mejor relación lineal con MPG y que VOL tiene una pobre relación lineal con MPG. En tanto que para HP y SP la relación lineal no es muy marcada.

Ahora , analizemos la relación entre HP y MPG. Un plot de estas variables se muestra en la figura 2.2. Si hacemos la regresión lineal entre las dos variables se obtiene

```
>  l1<-lm(mpg~hp)
> l1

Call:
lm(formula = mpg ~ hp)

Coefficients:
(Intercept)           hp
    50.0661      -0.1390

> summary(l1)$r.squared
[1] 0.6239
```

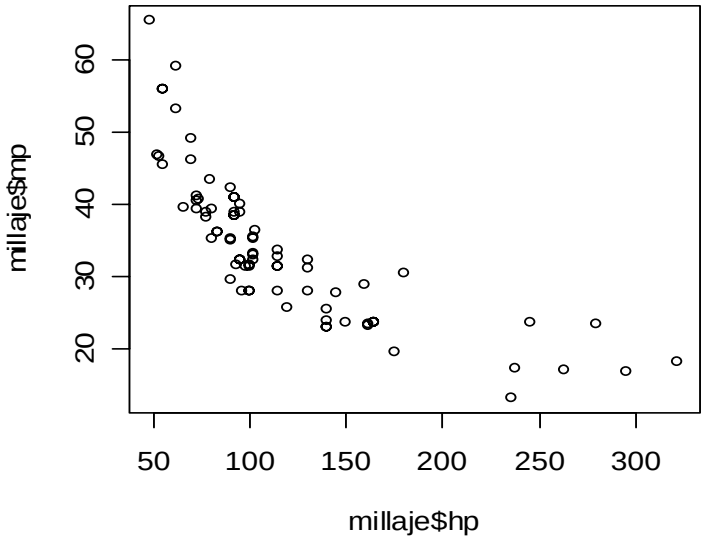


Figura 2.2. Plot de MPG versus HP.

El R^2 =62.4% no está bajo, pero hay que tratar de mejorarlo, usando las alternativas a y b. En la gráfica se observa una curvatura, así que se podría ajustar una regresión cuadrática. Los resultados usando hp y hp2=hp^2 como variables predictoras son los siguientes:

```
> l2=lm ( mpg ~ hp + hp2)
> l2

Call:
lm(formula = mpg ~ hp + hp2)

Coefficients:
(Intercept)          hp         hp2
 71.2313424   -0.4598708    0.0009707

> summary(l2)$r.squared
[1] 0.8067
```

El R^2 resulta ser 80.7% lo que representa una gran mejora, pero hay un peligro de hacer predicciones porque al final la cuadrática tiende a subir, y se podría concluir que un auto con 400 HP podría tener un rendimiento de 42.59 millas por galón, similar al de un carro de 50 HP. Este es un ejemplo de un modelo sobreajustado ("overfitted"). Notar también el valor bien pequeño del coeficiente del término cuadrático, el cual podría causar problema en el cálculo de las predicciones.

Observando más detenidamente la gráfica de la figura 2.2 se puede ver que hay un comportamiento asintótico en la parte inferior, es decir, que después de cierto nivel de HP, el millaje tiende a estabilizarse. Esto sugiere que podriamos tratar un modelo hiperbólico de la

forma MPG=$\alpha+\beta\frac{1}{HP}$ para ajustar los datos. Considerando la predictora hp1=1/hp, se obtienen los siguientes resultados en R.

```
> l3=lm(mpg~hp1)
> l3

Call:
lm(formula = mpg ~ hp1)

Coefficients:
(Intercept)         hp1
      9.73     2373.11

> summary(l3)$r.squared
[1] 0.8429
```

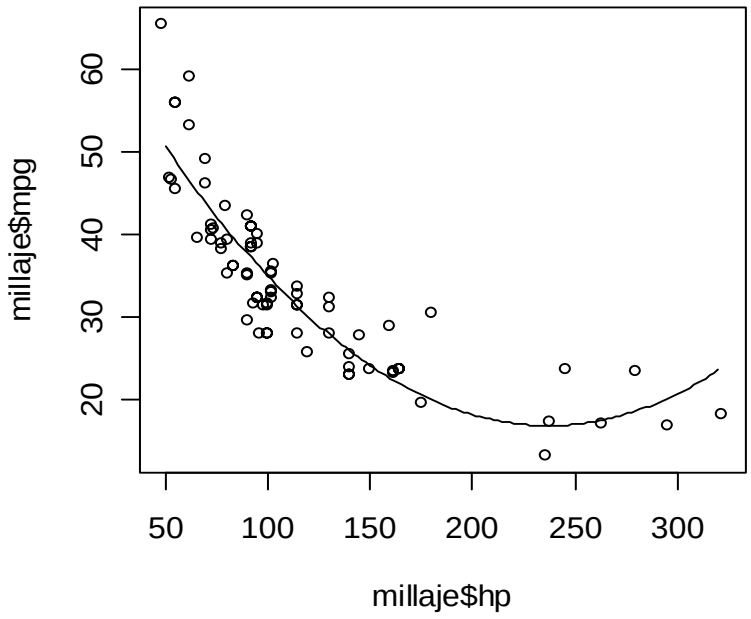


Figura 2.3. Regresión cuadrática de MPG versus HP

Notar que el R^2 =84.3% está bastante aceptable lo cual indica un buen ajuste del modelo. Así para un carro con 400 de HP su MPG será 15.66.

Alguién que no quiere perder el tiempo explorando relaciones polinómicas o haciendo transformación de variables, tratará de conseguir información acerca de otras variables

predictoras, con la esperanza de subir sustancialmente su R^2 pero usando solamente modelos lineales.
Del plot matricial que aparece en la figura 2.1 no hay ninguna duda de que la variable a considerar conjuntamente con HP sería WT. Haciendo uso de R se obtiene

```
> l4<-lm(mpg~hp+wt)
> l4

Call:
lm(formula = mpg ~ hp + wt)

Coefficients:
(Intercept)          hp          wt
   66.85500     -0.02097     -0.99037

> summary(l4)$r.squared
[1] 0.8235
```

El cual sería el segundo mejor modelo usando el criterio de R^2 ya que produce un valor de 82.4%. Si usamos el hecho de que la relación de Y con HP1 resulta ser bastante buena, podemos intentar ajustar un modelo lineal con HP1 y WT como las variables predictoras. Los resultados aparecen a continuación:

```
> l5<-lm(mpg~hp1+wt)
> l5

Call:
lm(formula = mpg ~ hp1 + wt)

Coefficients:
(Intercept)          hp1          wt
    36.5361    1387.1768      -0.5439

> summary(l5)$r.squared
[1] 0.8933
```

Este último sería el mejor modelo hasta ahora ya que su R^2=89.9 es el mayor de todos. Así se puede seguir explorando más modelos, pero teniendo cuidado de no caer en ***"overfitting"***.

Si ajustamos un modelo de regresión lineal multiple con las 4 variables predictoras disponibles se obtiene

```
> l6<-lm(mpg~vol+hp+sp+wt)
> l6

Call:
lm(formula = mpg ~ vol + hp + sp + wt)

Coefficients:
(Intercept)          vol          hp          sp          wt
  192.43775     -0.01565      0.39221     -1.29482     -1.85980

> summary(l6)$r.squared
```

[1] 0.8733

Si habría que decidir entre este último modelo y el anterior, habría que escoger el anterior porqué con solo dos variables predictoras se obtiene un mejor R^2.

2.2 El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones tomadas es de la forma

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + e_i \quad (2.1)$$

para i=1,2,….n. Escribiendo el modelo para cada una de las observaciones, éste puede ser considerado como un sistema de ecuaciones lineales de la forma

$$y_1 = \beta_o + \beta_1 x_{11} + \beta_2 x_{12} + ...\beta_p x_{1p} + e_1$$
$$y_2 = \beta_o + \beta_1 x_{21} + \beta_2 x_{22} + ...\beta_p x_{2p} + e_2$$
……..
$$y_n = \beta_o + \beta_1 x_{n1} + \beta_2 x_{n2} + ...\beta_p x_{np} + e_n$$

que puede ser escrita en forma matricial como

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ .. & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

O sea,

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e} \quad (2.2)$$

donde **Y** es un vector columna n dimensional**, X** es una matriz n x p', con p'=p+1, **β** es el vector de coeficientes de regresión a ser estimados, su dimensión es p' y **e** es un vector columna aleatorio de dimensión n

Por ahora, las únicas suposiciones que se requieren son que E(**e**)=**0** y que la matriz de varianza-covarianzas de los errores está dada por Var(**e**)=$\sigma^2\mathbf{I_n}$, donde $\mathbf{I_n}$ es la matriz identidad de orden n.

2.2.1 Estimación del vector de parámetros β por Cuadrados Mínimos

Al igual que en regresión lineal simple hay que minimizar la suma de cuadrados de los errores. La suma de cuadrados de los errores puede ser expresada vectorialmente de la siguiente manera

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} = (\mathbf{Y} - \mathbf{X\beta})'(\mathbf{Y} - \mathbf{X\beta}) \quad (2.3)$$

donde el símbolo ' indica transpuesta del vector o matriz (es decir, la matriz que se obtiene intercambiando las fila por columnas en la matriz original). Haciendo operaciones con los vectores y matrices se obtiene

$$Q(\boldsymbol{\beta}) = \mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y} - \mathbf{Y'X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} = \mathbf{Y'Y} - 2\boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} \quad (2.4)$$

En la igualdad anterior se ha usado la propiedad **(AB)'=B'A'.** Derivando Q con respecto a β e igualando a cero se obtiene el sistema de ecuaciones normales;

$$\mathbf{X'X}\boldsymbol{\beta} = \mathbf{X'Y} \quad (2.5)$$

de donde resolviendo para β se obtiene

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \quad (2.6)$$

aqui $(\mathbf{X'X})^{-1}$ representa la matriz inversa de **(X'X).** Notar que **X'X** es simétrica, pues su transpuesta da la misma matriz.

En la regresión lineal simple, p=1 y el modelo puede ser escrito en forma matricial como

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Manipulando las matrices se obtiene que

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_{11} & x_{21} & . & . & . & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_{11} & x_{21} & . & . & . & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \end{bmatrix}$$

Luego las ecuaciones normales se reducen a:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \end{bmatrix}$$

Por comodidad podemos eliminar el segundo subindice de las x's ya que no afecta en nada. Como

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

se concluye que

$$\begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{nSxx} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_i y_i \end{bmatrix} \quad (2.7)$$

y haciendo manipuleo algebraíco se llega a las formulas para los estimadores del intercepto y de la pendiente que se vieron en la sección 1.2 del capítulo 1.

2.2.2 Propiedades del estimador $\hat{\boldsymbol{\beta}}$

En forma similar al caso simple, el estimador minimo cuadrático tiene las siguientes propiedades:

a) $\hat{\boldsymbol{\beta}}$ es insesgado, o sea $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. En efecto,

$$E(\hat{\boldsymbol{\beta}}) = E[(\mathbf{X'X})^{-1}\mathbf{X'Y}] = E[(\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})] = E[(\mathbf{X'X})^{-1}(\mathbf{X'X})\,\boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'e}]$$
$$= \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'}E(\mathbf{e})$$

como $E(\mathbf{e})=0$, se llega a $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

b) Var($\hat{\boldsymbol{\beta}}$)=$\sigma^2(\mathbf{X'X})^{-1}$

Para probar esto debemos usar una propiedad de la matriz de varianza-covarianza de **Az** donde **A** es matriz y **z** vector columna. La propiedad dice que Var(**Az**)==**A**Var(**z**)**A' .**

Luego, $Var(\hat{\boldsymbol{\beta}}) = Var[(\mathbf{X'X})^{-1}\mathbf{X'Y}] = (\mathbf{X'X})^{-1}\mathbf{X'}Var(\mathbf{Y})((\mathbf{X'X})^{-1}\mathbf{X'})'$

Usando los hechos que **(X')'=X** y que **[(X'X)$^{-1}$]'=(X'X)$^{-1}$**, por simetría de la matriz inversa de **X'X,** se obtiene que

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X'X})^{-1}\mathbf{X'}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1}$$

c) Si no se asume normalidad el estimador minimo-cuadrático $\hat{\boldsymbol{\beta}}$ es el mejor estimador dentro de los estimadores lineales insesgados de **β,** en el sentido que es el que tiene la varianza más pequeña. Este es llamado el **teorema de Gauss-Markov.**

d) Si se asume normalidad de los errores entonces $\hat{\boldsymbol{\beta}}$ es el mejor estimador entre todos los estimadores insesgados de **β**

2.2.3 Estimación de la varianza σ^2

En un modelo de regresión lineal múltiple con p variables predictoras (con el intercepto habrían en total p+1 parámetros a estimar), se tiene que un estimado de la varianza de los errores es

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} = \frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-p-1} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p-1} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1} \qquad (2.8)$$

El numerador de la expresión representa la suma de cuadrados de los residuales y puede ser escrito como:

$$SSE = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{Y}-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y})'(\mathbf{Y}-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y}) = \mathbf{Y'(I-H)'(I-H)Y}$$

donde **H=X(X'X)$^{-1}$X'** es conocida en regresión como la ***"Hat Matrix"*** (la matriz sombrero).
Notar que **H'=H** y que **H^2=HH= X(X'X)$^{-1}$X' X(X'X)$^{-1}$X'=H.** En álgebra matricial se dice que H es idempotente. **H** tiene muy buenas propiedades una de ellas es que Traza(**H**)=rango(**H**)=p'. Por otro lado, **(I-H)'(I-H)=(I-H)(I-H)=I-H-H+H^2=I-2H+H=I-H.** Así que también **I-H** es también simétrica e idempotente.

En consecuencia, la varianza estimada de los errores puede ser escrita como:

$$\hat{\sigma}^2 = \frac{\mathbf{Y'(I - X(X'X)^{-1}X')Y}}{n-p-1} = \frac{\mathbf{Y'(I-H)Y}}{n-p-1} \qquad (2.9)$$

Más común es usar el símbolo s para la desviación estándar estimada de los errores y $s = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{MSE}$

Propiedad: Sea **Y** un vector aleatorio n-dimensional tal que E(**Y**)= $\boldsymbol{\mu}$ y VAR(**Y**)=**V** entonces

E(**Y'AY**)=Traza(**AV**)+$\boldsymbol{\mu}$**'A**$\boldsymbol{\mu}$ **(2.10)**

Usando la propiedad anterior con $\boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}$ y $\mathbf{V}=\sigma^2\mathbf{I_n}$ se puede mostrar que $E[s^2]=\sigma^2$.

Ejemplo 2: Hallar el modelo de regresión lineal multiple para explicar el rendimiento en millaje promedio por galón (MPG) de los vehículos de acuerdo a cuatro variables predictoras: VOL, HP, SP y WT e interpretar los valores estimados.

Las variables predictoras que fueron usadas antes en el ejemplo 1 están definidas como sigue:

VOL: Capacidad de la cabina en pies cúbicos
HP: Potencia del motor
SP: Velocidad máxima (mph)
WT: Peso del vehículo (100 lb)

El modelo de regresión que ya fue obtenido en el ejemplo 1 es el siguiente:

```
MPG = 192 - 0.0156 VOL + 0.392 HP - 1.29 SP - 1.86 WT
```

Interpretación de los coeficientes de regresión estimados:

$\hat{\beta}_1 =$ -0.0156 significa que el millaje promedio por galón baja en promedio en 0.0156 cuando el volumen interior del carro aumenta en un pie cubico, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_2 =$ 0.392 significa que el millaje promedio por galón aumenta en promedio en 0.392 cuando la potencia del motor aumenta en 1 HP, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_3 =$ -1.29 significa que el millaje promedio por galón baja en promedio en 1.29 cuando la velocidad máxima del carro aumenta en 1 milla por hora, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_4 =$ -1.86 significa que el millaje promedio por galón baja en 1.86 cuando el peso del vehículo aumenta en 100 libras, asumiendo que las otras variables permanecen fijas.

En general, un coeficiente de regresión representa el cambio promedio en la variable de respuesta cuando la variable predictora correspondiente se incrementa en una unidad adicional y asumiendo que las otras variables predictoras premanecen fijas.

2.3. Inferencia en Regresión lineal múltiple

En esta sección se harán pruebas de hipótesis e intervalos de confianza acerca de los coeficientes del modelo de regresión poblacional. También se calcularán intervalos de confianza de las predicciones que se hacen con el modelo.

De ahora en adelante vamos a suponer que **e~NI(0,σ^2I$_n$)** o equivalente que **Y~NI(Xβ, σ^2I$_n$).**

Al igual que en regresión lineal simple la variación total de Y se descompone en dos variaciones: una debido a la regresión y otra debido a causas no controlables. Es decir,

SST=SSR+ SSE

Por teoría de modelos lineales se puede determinar que las sumas de cuadrados que aparecen en el análisis de regresión son formas cuadráticas de la variable de respuesta Y. Por lo tanto, éstas se distribuyen como una Ji-cuadrado. Más especificamente, se pueden establecer los siguientes resultados:

i). $\frac{SST}{\sigma^2} \sim \chi^2_{(n-1)}$ Ji-cuadrado no central con n-1 grados de libertad. Los grados de libertad se pueden establecer de la fórmula de cálculo de SST, pues en ella se usan *n* datos, pero en ella aparece un valor estimado ($\bar{y}$) por lo tanto se pierde un grado de libertad.

ii). $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-p-1)}$ Ji-cuadrado con n-p-1 grados de libertad . Para calcular *SSE* se usan *n* datos pero hay presente un estimado $\hat{y}_i$ cuyo cálculo depende a su vez de p+1 estimaciones. Por lo tanto se pierden p+1 grados de libertad. También se puede escribir que $\frac{(n-p-1)s^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$

iii). $\frac{SSR}{\sigma^2} \sim \chi^2_{(p)}$ Ji-cuadrado no central con p grados de libertad

2.3.1 Prueba de hipótesis acerca de un coeficiente de regresión individual

En este caso la hipótesis nula más importante es H_o: $\beta_i=0$ (i=1,2..,p), o sea la variable X_i no es importante en el modelo, versus la hipótesis alterna H_a:$\beta_i\neq 0$; la variable X_i si merece ser considerada en el modelo. La prueba estadística es la prueba de t, definida por

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s\sqrt{C_{ii}}}$$

donde el error estándar de $\hat{\beta}_i$ se calcula por $se(\hat{\beta}_i) = s\sqrt{C_{ii}}$, C_{ii} es el i-ésimo elemento de la diagonal de **(X'X)$^{-1}$.** Esta t se distribuye como una t de Student con n-p-1 grados de libertad. R al igual que otros programas de computadoras, da el "P-value" de la prueba t. Para el ejemplo anterior se obtiene lo siguiente

```
> summary(l6)

Call:
lm(formula = mpg ~ vol + hp + sp + wt)

Residuals:
   Min     1Q Median     3Q    Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
vol          -0.01565    0.02283  -0.685    0.495
hp            0.39221    0.08141   4.818 7.13e-06 ***
sp           -1.29482    0.24477  -5.290 1.11e-06 ***
wt           -1.85980    0.21336  -8.717 4.22e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los "P-values" de la prueba de t sugieren que la variable ***VOL*** no contribuye al modelo, pues se acepta la hipótesis nula de que dicho coeficiente es cero. Las otras tres variables *WT, HP y SP* si parecen ser importantes en el modelo ya que los "P-values" de la prueba t correpondientes son menores que .05.

2.3.2 Prueba de Hipótesis de que todos los coeficientes de regresión son ceros.

En este caso la hipótesis nula es Ho: $\beta_1=\beta_2=\ldots=\beta_p=0$, o sea que el modelo no sirve, versus la hipótesis alterna H_a: Al menos uno de los coeficientes es distinto de cero, o sea al menos una de las variables del modelo puede ser usada para explicar la variación de Y.
En la sección 2.2.3 se mencionó que $E(s^2)=\sigma^2$. La suma de cuadrados del error tiene n-p-1 g.l. Nuevamente usando esperado de una formas cuadrática se puede mostrar que

$$E(SSR)=E[\mathbf{Y'(H-11'/n)Y}]=p\sigma^2+\boldsymbol{\beta}'\mathbf{X}'\mathbf{(H-11'/n)X}\boldsymbol{\beta} \qquad \mathbf{(2.11)}$$

Donde **1** es un vector columna de n unos Si la hipótesis nula se cumpliera entonces $E(MSR)=\sigma^2$
La suma de cuadrados de Regresión tiene p grados de libertad que es igual al número de variables predictoras en el modelo.
Se puede mostrar que si la hipótesis nula es cierta entonces :

$$F=\frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}}=\frac{MSR}{MSE}$$

se distribuye como una F con p grados de libertad en el numerador y n-p-1 g.l en el denominador.

La prueba de F se obtiene al hacer la tabla del análisis de varianza para la regresión múltiple, la cual se muestra a continuación:

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados Medios	F
Regresión	SSR	P	MSR=SSR/p	F=MSR/MSE
Error	SSE	n-p-1	MSE=SSE/n-p-1	
Total	SST	n-1		

Para el ejemplo 1, usando todas las variables predictoras, se tiene,

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733, Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

Notar que la desviación estimada del error es s=3.653= $\sqrt{MSE} = \sqrt{13.3}$. El "P-value" de la Prueba de F es 0.0000, lo cual lleva a la conclusión de que al menos una de las variables predictoras presentes en el modelo es importante para predecir MPG.

El coeficiente de Determinación R^2 tiene la misma interpretación que en regresión lineal simple y se calcula por $R^2 = \frac{SSR}{SST}$.

El R^2=87.3%, lo que quiere decir que hay un ajuste bastante bueno asumiendo que no hay datos contaminados en el conjunto de datos. El 87.3% de la variación del millaje promedio por galón es explicada por su relación lineal con VOL, HP, SP y WT. El R-Sq(adj) llamado el R^2 ajustado será definido más adelante en el capitulo de selección de variables.

La suma de cuadrados de regresión puede ser particionada en tantas partes como variables predictoras existen en el modelo. Esto es llamado un particionamiento secuencial de la suma de cuadrados de regresión y sirve para determinar la contribución de cada una de las variables predictoras al comportamiento de Y. Formalmente,

$SSR(\beta_1,\beta_2,,\ldots.\beta_p/\ \beta_0)=SSR(\beta_1/\ \beta_0)+SSR((\beta_2,/\beta_1,\beta_0)+\ldots\ldots\ldots\ldots..SSR(\beta_p/\beta_{p-1},,\ldots.\beta_1,\beta_0)$

Aquí $SSR(\beta_k/\beta_{k-1},,\ldots.\beta_1,\beta_0)$ significa el incremento en la suma de cudrados de regresión cuando la variable X_k es incluida en el modelo, el cual ya contiene las variables predictivas $X_1,\ldots X_{k-1}$
La función **anova** de **R** produce estas sumas parciales. Para el ejemplo anterior se obtiene lo siguiente:

```
> l6

Call:
lm(formula = mpg ~ vol + hp + sp + wt)

Coefficients:
(Intercept)         vol          hp          sp          wt
  192.43775    -0.01565     0.39221    -1.29482    -1.85980
```

```
> anova(l6)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
vol        1 1101.6  1101.6  82.563 8.172e-14 ***
hp         1 4731.1  4731.1 354.584 < 2.2e-16 ***
sp         1  233.6   233.6  17.509 7.515e-05 ***
wt         1 1013.8  1013.8  75.979 4.221e-13 ***
Residuals 77 1027.4    13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La Suma de Cuadrados de Regresión es 7080.1=1101.6+4731.1+233.6+1013.8. 233.6 significa que la suma de cuadrado de regresión aumenta en 233.6 cuando la variable SP es añadida al modelo, después que las variables VOL y HP ya están incluidas. El problema ahora es tratar de establecer pruebas para determinar si una variable predictora o un subconjunto de ellas efectivamente debe estar o no en el modelo.

Las sumas de cudrados de regresión secuenciales varia si se cambia el orden de las anteriores predictoras al momento de ajustar el modelo. Asi, si elegimos el orden WT, HP, SP y al final VOL se obtiene el siguiente resultado.

```
> l6<-lm(mpg~wt+hp+sp+vol)
> anova(l6)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)
wt         1 6641.5  6641.5 497.7630 < 2.2e-16 ***
hp         1   35.4    35.4   2.6516    0.1075
sp         1  397.0   397.0  29.7522 5.739e-07 ***
vol        1    6.3     6.3   0.4698    0.4951
Residuals 77 1027.4    13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notar que las variables HP y VOL no son significativas. Es claro que la variable VOL es la menos importante de las variables predictoras.

2.3.3 Prueba de hipótesis para un subconjunto de coeficientes de regresión

Algunas veces estamos interesados en probar si algunos coeficientes del modelo de regresión son iguales a 0 simultáneamente. Por ejemplo, si el modelo tiene *p* variables predictoras, quisieramos probar si los k primeros coeficientes son ceros, o sea H_o: $\beta_1=\ldots..=\beta_k=0$. De rechazarse la hipótesis nula implicaría que las k primeras variables predictoras pueden ser excluidas del modelo.
Al modelo en donde se consideran todas las p variables se le llama el ***modelo completo*** y al modelo que queda asumiendo que la hipótesis nula es cierta se le llama el ***modelo reducido*.**
Es decir, que el modelo reducido sería

$$Y=\beta_{k+1}X_{k+1}+\beta_{k+2}X_{k+2}+\ldots..+\beta_pX_p+e \qquad (2.12)$$

Para probar si la hipótesis nula es cierta se usa una prueba de F que es llamada F-parcial. La prueba de F parcial se calcula por

$$Fp = \frac{\frac{SSR(C) - SSR(R)}{k}}{\frac{SSE(C)}{n-p-1}} = \frac{\frac{SSR(C) - SSR(R)}{k}}{MSE(C)}$$

donde SSR(C) y MSE(C) representan la suma de cuadrados de regresión y el cuadrado medio del error del modelo completo respectivamente, y SSR(R) es la suma de cuadrados de regresión del modelo reducido.

SSR(C)=SSR($\beta_1,\beta_2,\ldots\beta_p/\beta_o$) y
SSR(R) =SSR($\beta_{k+1},\beta_{k+2},\ldots,\beta_p/\beta_o$)

SSR(C) –SSR(R)=SSR($\beta_1,\beta_2,\ldots\beta_k/\beta_{k+1},\beta_{k+2},\ldots\beta_p$)

Esta última diferencia representa el incremento en la suma de cuadrados de regresión cuando $X_1,\ldots X_k$ son añadidas al modelo en el cual ya están presentes $X_{k+1},\ldots X_p$ y la constante

Si F_p es mayor que $F_{1-\alpha}$ usando k grados de libertad para el numerador y n-p-1 para el denominador entonces se rechaza H_o, en caso contrario se acepta.

R no tiene una opción que haga directamente la prueba de F parcial. Hay que calcular los dos modelos de regresión y usar las sumas de cuadrados de regresión de ambos modelos para calcular la prueba de F parcial.

Ejemplo 3: En el ejemplo 1, probar que las variables VOL y HP no son importantes y pueden ser excluidas del modelo

Haciendo el análisis de regresión sin incluir VOL y HP se obtiene:

```
> l2

Call:
lm(formula = mpg ~ sp + wt, data = millaje)

Coefficients:
(Intercept)           sp           wt
   75.64938     -0.09816     -0.99738

> anova(l2)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
sp         1 3842.6  3842.6  219.49 < 2.2e-16 ***
wt         1 2881.8  2881.8  164.61 < 2.2e-16 ***
Residuals 79 1383.0    17.5
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Luego SSR(R)=6724.4 y SSR(C)-SSR(R)=7080.1-6724.4=355.7.

```
> #Suma de cuadrados de regresion del modelo completo
> a=sum(anova(l1)$Sum[-(p+1)])
> #Suma de cuadrados de regresion del modelo reducido
> b=sum(anova(l2)$Sum[-(k+1)])
> #Cuadrado Medio del error del modelo completo
> c=anova(l1)$Mean[p+1]
> #Calculo del F parcial
> fp<-((a-b)/2)/c
> fp
[1] 13.32720
```

Luego la F-parcial será

$$F_p = \frac{\frac{355\ .63}{2}}{13\ .34} = \frac{177\ .81}{13\ .34} = 13\ .33$$

Usando un nivel de significación del 5%, Hay que comparar Fp con F(.95,2,77).

```
> #Hallando el percentil de la F con alpha=.05
> qf(.95,k,n-p-1)
[1] 3.115366
```

Luego Fp> F(.95,2,77)=3.11 por lo tanto se rechaza la prueba y se concluye que VOL y HP no pueden ser eliminadas simúltanemanente, al menos una de ellas es importante.

2.3.4 Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Supongamos que se desea predecir el valor medio de la variable de respuesta Y para una combinación predeterminada de las variables predictoras $X_1,\ldots\ldots X_p$,.Consideremos el vector $\mathbf{x'}_o = (1, x_{1,0},\ldots.x_{p,0})$ donde $x_{1,0},\ldots.x_{p,0}$ son los valores observados de $X_1,\ldots..X_p$ respectivamente.

El valor predicho para el valor medio de la variable de respuesta Y será $\hat{y}_o = \mathbf{x'_o}\,\hat{\boldsymbol{\beta}}$. De donde, $Var(\hat{y}_o) = \mathbf{x'_o}\,\mathbf{Var}((\hat{\boldsymbol{\beta}})\mathbf{x_o} = \sigma^2\mathbf{x'_o}\,(\mathbf{X'X})^{-1}\mathbf{x_o}$. En consecuencia asumiendo que los errores están normalmente distribuidos se tiene que un intervalo del 100(1-α)% para el valor medio de Y dado que $\mathbf{x}=\mathbf{x_o}$ es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)}\, s\sqrt{\mathbf{x'_o}\,(\mathbf{X'X})^{-1}\mathbf{x_o}} \qquad (2.13)$$

También usando la misma derivación que se hizo en el caso de regresión lineal simple se llega a establecer que un intervalo de confianza (más conocido como intervalo de predicción) del 100(1-α)% para el valor individual de Y dado $\mathbf{x}=\mathbf{x_o}$ es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)}\, s\sqrt{1+\mathbf{x'_o}\,(\mathbf{X'X})^{-1}\mathbf{x_o}} \qquad (2.14)$$

Ejemplo 4: Usando el conjunto de datos **millaje,** hallar un intervalo de confianza del 95% para el millaje promedio por galón de todos los vehículos con capacidad interior de 90 pies cúbicos, un HP de 50 una velocidad máxima de 1200 millas por galón y un peso de 20,000 libras. Hallar un intervalo de predicción para el millaje de un carro con las mismas características anteriores.

Usando R se obtiene

```
>   # hallando el intervalo de confianza del 95% para el valor medio
>   sp<-100
>   wt<-20
>   vol<-90
>   hp<-50
>   nuevo<-as.data.frame(cbind(sp,wt,vol,hp))
>   nuevo
  sp wt vol hp
1 100 20  90 50
>   predict.lm(l1,nuevo,se.fit=T,interval=c("confidence"),level=.95)
$fit
       fit     lwr     upr
[1,] 43.9624 42.41585 45.50894

$se.fit
[1] 0.7766682

$df
[1] 77

$residual.scale
[1] 3.652755

>  #Hallando el ntervalo de prediccion del 99% para los mismos datos
> predict.lm(l1,nuevo,se.fit=T,interval=c("prediction"),level=.99)
$fit
       fit     lwr     upr
[1,] 43.9624 34.09908 53.82571

$se.fit
[1] 0.7766682

$df
[1] 77

$residual.scale
[1] 3.652755
```

Hay un 95% de confianza de que el millaje promedio de todos los carros con las características dadas caiga entre 42.41 y 45.50 millas por galón. Hay un 99% de confianza de que el rendimiento de millas por galón de un carro con las características anteriores caiga entre 34.09 y 53.82

2.3.5 La prueba de Falta de Ajuste

Es una prueba que se usa para determinar si la forma del modelo que se está considerando es adecuada. Es decir, si la regresión debe o no incluir términos potencias o interacciones entre las variables predictoras. En el caso de regresión simple la prueba requiere que haya por lo menos un valor de la variable predictora con varias observaciones de y. En regresión múltiple se debe suponer que hay m combinaciones distintas de las n observaciones de las p variables predictoras y que por cada una de esas combinaciones hay n_i (i=1,…m) observaciones de la variable de respuesta, es decir, $\sum_{i=1}^{m} n_i = n$.

La Suma de Cuadrados del Error se particiona de la siguiente manera

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i-\hat{y}_i)^2 \qquad (2.15)$$

donde $\hat{y}_i$ es el valor predicho por el modelo de regresión para la i-ésima combinación de las variables predictoras , mientras que $\bar{y}_i$ es el valor promedio de la variable predictora para la i-ésima combinación.

La primera suma de cuadrados del lado derecho es llamado la **Suma de Cuadrados del Error Puro (SSPE)** y tiene n-m grados de libertad. Si no hubiera varios valores de la variable de respuesta por cada combinación de las predictoras esta suma sería cero. Se puede demostrar que el valor esperado del cuadrado medio del error puro es igual a la varianza poblacional σ^2 , sea o no sea el modelo de regresión adecuado.

La segunda suma de cuadrados que también puede ser escrita como $\sum_{i=1}^{m} n_i(\bar{y}_i-\hat{y}_i)^2$ es llamada la **Suma de Cuadrados de Falta de Ajuste (SSLOF)** y tiene m-p-1 grados de libertad. Si el modelo especificado es correcto entonces el valor esperado del cuadrado medio de Falta de Ajuste es igual a σ^2. Si le faltan términos al modelo (por ejemplo: potencias, productos de variables, etc.) entonces el cuadrado medio del la falta de ajuste sobreestimará a σ^2.

En resumen, la hipótesis nula será Ho: El modelo es adecuado (no hay falta de ajuste) versus Ha: el modelo no es adecuado y la prueba estadística es una prueba de F dada por

$$F = \frac{SSLOF/(m-p-1)}{SSPE/(n-m)} = \frac{MSLOF}{MSPE}$$

que se distribuye como una F(m-p-1,n-m) si la hipótesis nula es cierta. La hipótesis nula es rechazada si el valor de la prueba estadstica es mayor que F(1-α,m-p-1, n-m).

R no tiene una función para calcular directamente la prueba de bondad de ajuste. Hay que introducir una variable adicional que identifique los valores de y correspondiente al mismo valor de la variable predictora.

Ejemplo 5: Usando el conjunto de datos **millaje,** hacer una prueba de Falta de Ajuste si se considera la variable de respuesta MPG y la variable predictora HP.

Usando R:

```
> millajelf=millaje[,c(1,5)]
```

```
> table(millajelf$hp)

 49  52  53  55  62  66  70  73  74  78  80  81  84  90  92  93  95  96  98 100
  1   1   1   3   2   1   2   3   1   2   1   2   2   4   7   1   5   1   1   4
102 103 115 120 130 140 145 150 160 162 165 175 180 236 238 245 263 280 295 322
  5   1   5   1   3   4   1   1   1   2   4   1   1   1   1   1   1   1   1   1

>#  Hay m=40 valores distintos de la predictora
>.# anadiendo una columna  var3 que identifica a que grupo pertenece cada
>#  observación
> millajelf[1:10,]
   mpg hp var3
1  65.4 49    1
2  56.0 55    4
3  55.9 55    4
4  49.0 70    7
5  46.5 53    3
6  46.2 70    7
7  45.4 55    4
8  59.2 62    5
9  53.3 62    5
10 43.4 80   11
……
……
> #haciendo la regresion lineal simple
> l1=lm(mpg~hp,data=millajelf)
> l1

Call:
lm(formula = mpg ~ hp, data = millajelf)

Coefficients:
(Intercept)          hp
   50.0661      -0.1390

> anova(l1)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
hp         1 5058.0  5058.0  132.69 < 2.2e-16 ***
Residuals 80 3049.4    38.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Haciendo el analisis de varianza de claificacion simple de mpg
>#  entre los 40 grupos diferentes
>l2=lm(mpg~factor(var3),data=millajelf)
> anova(l2)
Analysis of Variance Table

Response: mpg
             Df Sum Sq Mean Sq F value    Pr(>F)
factor(var3) 39 7794.4   199.9  26.809 < 2.2e-16 ***
Residuals    42  313.1     7.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> # Haciendo el anova  para comparar  los dos modelos .  Se extrae la suma de cuadrados del
>#  error Puro y la suma de cuadrados de falta de Ajuste.
>#
>anova(l1,l2)
Analysis of Variance Table

Model 1: mpg ~ hp
Model 2: mpg ~ factor(var3)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     80 3049.44
2     42  313.11 38   2736.33 9.6592 1.703e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

MINITAB y otros programas estadísticos dan el "P-value" de esta prueba. MINITAB además da una prueba de Falta de Ajuste que no requiere que hayan varios valores de la variable de respuesta para cada combinación. Se debe usar la siguiente secuencia **Stat>Regression>Regression** y oprimiendo el botón de **options** se elige luego **Lack of Fit y Data Subsetting Lack of Fit Test.**

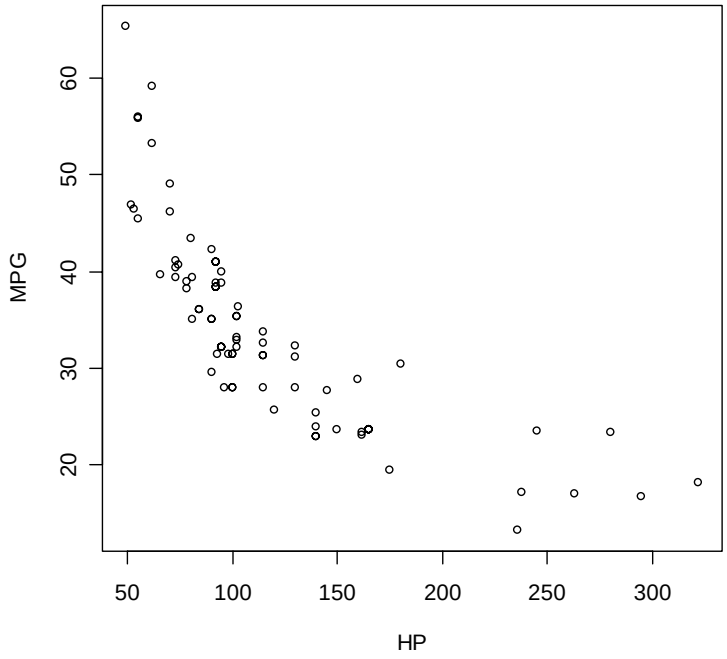
```
The regression equation is
MPG = 50.1 - 0.139 HP

Predictor        Coef       StDev          T        P
Constant       50.066       1.569      31.90    0.000
HP           -0.13902     0.01207     -11.52    0.000

S = 6.174       R-Sq = 62.4%     R-Sq(adj) = 61.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      5058.0      5058.0    132.69    0.000
Residual Error    80      3049.4        38.1
  Lack of Fit     38      2736.3        72.0      9.66    0.000
  Pure Error      42       313.1         7.5
Total             81      8107.5

23 rows with no replicates

Unusual Observations
Obs         HP        MPG         Fit   StDev Fit    Residual    St Resid
  1         49     65.400      43.254       1.068      22.146        3.64R
  2         55     56.000      42.420       1.013      13.580        2.23R
  3         55     55.900      42.420       1.013      13.480        2.21R
  8         62     59.200      41.447       0.953      17.753        2.91R
 71        245     23.500      16.005       1.687       7.495        1.26 X
 72        280     23.400      11.140       2.080      12.260        2.11RX
 78        322     18.100       5.301       2.565      12.799        2.28RX
 80        263     17.000      13.503       1.888       3.497        0.59 X
 81        295     16.700       9.054       2.252       7.646        1.33 X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

Lack of fit test

```
Possible curvature in variable HP (P-Value = 0.000)
Possible lack of fit at outer X-values       (P-Value = 0.000)
Overall lack of fit test is significant at P = 0.000
```

Si usamos el "P-value" de la prueba F de Falta de Ajuste se concluye que se rechaza la hipótesis nula, es decir, hay suficiente evidencia para concluir que faltan términos en el modelo.

La prueba de Falta de ajuste que da MINITAB es más informativa aún, dice que hay una posible curvatura en HP (ver el plot), que hay outliers en la dirección de la variable predictora y finalmente da una prueba de ajuste global.



Consideremos ahora todas las variable predictoras

```
> millajep=millaje[,2:5]
> dim(unique(millajep))
[1] 70  4

>#  Hay m=70 valores distintos de la predictora
>.# anadiendo una columna  var4 que identifica a que grupo pertenece cada
>#  observación
> millajelf=cbind(millaje,var4)
> millajelf[1:10,]
   mpg  sp   wt vol hp var4
1  65.4  96 17.5  89 49    1
2  56.0  97 20.0  92 55    2
3  55.9  97 20.0  92 55    2
4  49.0 105 20.0  92 70    3
5  46.5  96 20.0  92 53    4
```

```
6  46.2 105 20.0  89 70    5
7  45.4  97 20.0  92 55    2
8  59.2  98 22.5  50 62    6
9  53.3  98 22.5  50 62    6
10 43.4 107 22.5  94 80    7
>#  Haciendo  ;la regresion lineal multiple
> l2=lm(mpg~sp+wt+vol+hp,data=millajelf)
> anova(l2)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)
sp         1 3842.6  3842.6 287.9944 < 2.2e-16 ***
wt         1 2881.8  2881.8 215.9879 < 2.2e-16 ***
vol        1   46.0    46.0   3.4451   0.06727 .
hp         1  309.7   309.7  23.2093 7.131e-06 ***
Residuals 77 1027.4    13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Haciendo el analisis de varianza de claificacion simple de mpg
>#  entre los 70 grupos diferentes

> l3=lm(mpg~factor(var4),data=millajelf)
> anova(l3)
Analysis of Variance Table

Response: mpg
             Df Sum Sq Mean Sq F value    Pr(>F)
factor(var4) 69 7990.3   115.8  11.859 2.130e-05 ***
Residuals    12  117.2     9.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Haciendo el anova  para comparar  los dos modelos .   Se extrae la suma de cuadrados del
>#  error Puro y la suma de cuadrados de falta de Ajuste.
>#
> anova(l2,l3)
Analysis of Variance Table

Model 1: mpg ~ sp + wt + vol + hp
Model 2: mpg ~ factor(var4)
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     77 1027.38
2     12  117.18 65    910.20 1.434 0.2518
```

En MINITAB se obtienen los siguientes resultados

```
The regression equation is
MPG = 192 - 0.0156 VOL + 0.392 HP - 1.29 SP - 1.86 WT

Predictor        Coef       StDev          T        P
Constant       192.44       23.53       8.18    0.000
VOL          -0.01565     0.02283      -0.69    0.495
HP            0.39221     0.08141       4.82    0.000
SP            -1.2948      0.2448      -5.29    0.000
WT            -1.8598      0.2134      -8.72    0.000
```

```
S = 3.653       R-Sq = 87.3%     R-Sq(adj) = 86.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4      7080.1      1770.0    132.66    0.000
Residual Error    77      1027.4        13.3
  Lack of Fit     65       910.2        14.0      1.43    0.252
  Pure Error      12       117.2         9.8
Total             81      8107.5

62 rows with no replicates

Source       DF      Seq SS
VOL           1      1101.6
HP            1      4731.1
SP            1       233.6
WT            1      1013.8

Unusual Observations
Obs        VOL        MPG         Fit   StDev Fit    Residual    St Resid
  1         89     65.400      53.415       1.267      11.985        3.50R
  8         50     59.200      47.235       1.213      11.965        3.47R
 29        102     29.500      38.511       0.623      -9.011       -2.50R
 72         50     23.400      19.912       1.692       3.488        1.08 X
 78         50     18.100      20.612       2.033      -2.512       -0.83 X
 80         50     17.000      20.778       1.593      -3.778       -1.15 X
 81        119     16.700      19.301       1.871      -2.601       -0.83 X
 82        107     13.200      12.710       1.636       0.490        0.15 X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Lack of fit test
Possible interactions with variable HP (P-Value = 0.100)
Possible interactions with variable SP (P-Value = 0.093)
Possible curvature in variable WT (P-Value = 0.008)
Possible lack of fit at outer X-values       (P-Value = 0.000)
Overall lack of fit test is significant at P = 0.000
```

La clásica prueba de Falta de Ajuste acepta la hipótesis nula, es decir, que no hay suficiente evidencia para concluir que haya Falta de Ajuste. Sin embargo, la prueba de Falta de ajuste de MINITAB concluye que hay interacción entre las variables predictoras HP y SP, que hay que transformar WT y además hay outliers.

EJERCICIOS

1. Propiedades de la matriz HAT $\mathbf{H=X(X'X)^{-1}X'}$

a) La traza de una matriz es igual a la suma de los elementos que están en su diagonal. Probar que Traza(H)=p' con p'=p+1, donde p es el número de variables predictoras.

b) Probar que $\frac{1}{n} \leq h_{ii} \leq 1$, donde h_{ii} es el i-ésimo elemento de la diagonal de H.

c) Probar que $\mathbf{H1_n=1_n}$ aqui $\mathbf{1_n}$ es un vector columna cuyo elementos son todos unos

2.Usar el conjunto de datos **Fuel** con variable de respuesta Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas. Los datos están disponible en la página de internet del texto

a) Hallar la variable que tiene correlación más slta con la variable de respuesta
b) Hacer un plot matricial para ver si no hay outliers y determinar si el coeficiente de correlación es confiable.
c) Hacer una regresión lineal de Y versus la variable determinada en los pasos a y b y tratar otros modelos: cuadrático, exponencial, logaritmico para mejorar el R^2, si es posible
d) Hallar un Intervalo de Confianza del 99% para el valor medio y el valor Predicho de Y, escogiendo un valor adecuado de la variables predictora usando el modelo lineal . Trazar las bandas de confiianza. Comentar sus resultados.
e) Hallar el modelo de regresion múltiple considerando todas las variable predictoras e interpretar los coeficientes de regresión.
f) Interpretar el coeficiente de Determinación R^2.
g) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.
h) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.

3 Uso de la factorizacion QR en Regresión Supongamos que tenemos una matriz ortogonal **Q** de orden nx p' (es decir, **Q'Q=I**) y una matriz triangular superior **R** tal que **QR=X**)
a) Probar que **R'R=X'X**

b) Escribir el estimador minimocuadrático $\widehat{\boldsymbol{\beta}}$ en terminos de Q, R y Y. Cual sería la ventaja de usar esta fórmula con respecto a la fórmula original.?

c) Expresar $\widehat{\mathbf{Y}}$ y $\widehat{\mathbf{e}}$ en terminos de Y y Q.

4. Usar el conjunto de datos **Highway**, con variable de respuesta es TASA y todas las otras como variables predictoras para responder las siguientes preguntas. Los datos están en la página de internet del texto.

a)Hallar la variable que tiene correlación más slta con la variable de respuesta
b)Hacer un plot matricial para ver si no hay outliers y determinar si el coeficiente de correlación es confiable
c) Hacer una regresión lineal de Y versus la variable determinada en los pasos a y b y tratar otros modelos: cuadrático, exponencial, logaritrmico para mejorar el R^2, si es posible
d) Hallar el modelo de regresion múltiple considerando todas las variable predictoras e interpretar los coeficientes de regresión.
e) Interpretar el coeficiente de Determinación R^2.
f) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.

g) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.
h) Hallar las dos variables que están menos correlacionadas con la variable de respuesta y probar la hipótesis de que ambas variables deben ser excluidas simúltaneamente del modelo.
i) Hallar un Intervalo de Confianza para el valor medio de Y y el valor Predicho del 99% para Y, escogiendo valores adecuados de las variables predictoras. Comentar sus resultados.
j) Hacer un análisis de falta de ajuste usando como variable predictora, aquella obtenida en a).

5. Considerando un modelo de regresión lineal multiple probar que $E[s^2]=\sigma^2$ donde s^2 es la varianza estimada del error definida en la ecuación 2.9.

6. Verificar la identidad de la ecuación 2.11

7. **Efecto de subajuste**. Supongamos que se ajusta el modelo $\mathbf{Y=X\beta+e}$ donde **X** es una matriz n x r cuando en realidad el modelo verdadero incluye s adicionales variables predictoras contenidas en la matriz **Z**. Es decir, que el verdadero modelo es $\mathbf{Y=X\beta+Z\gamma+ e}$. Mostrar que en general el estimador mínimos cuadrático $\hat{\boldsymbol{\beta}}$ usando el modelo reducido es sesgado. Asímismo mostrar que el estimador de la varianza es sesgado. Bajo que condiciones ambos estimadores serían insesgado?

8. Supongamos que se ha obtenido la siguiente regresión usando una muestra de 75 observaciones

$$Y= -5.16 + .325X_1 + 5.55X_2 + .3X_3 + .01X_4 + 8.75X_5 - .97X_6$$

a) Interpretar cualquiera de los coeficientes de las variables predictoras
b) Hallar el valor de la prueba estadística de F si el coeficiente de determinación $R^2=.95$
c) Explicar detalladamente como se probaría la hipótesis $Ho:\beta_1=\beta_3=\beta_4=\beta_6$

9. Supongamos que realmente el modelo $Y=X\beta+e$ (1) ajusta a nuestro conjunto de datos. Reescribamos el modelo anterior por

$$Y=X_1\beta_1+X_2\beta_2+e \quad , \text{ donde } X=(X_1 \mid X_2)$$

X_1 es de orden n x k y X_2 es de orden n x (p-k), n es el número de observaciones y p es el número de parametros del modelo, es decir el número de variables predictoras mas el intercepto. Consideremos que en lugar del modelo (1) se usa el siguiente modelo para ajustar los datos

$$Y=X_1\beta_1+e \quad (2)$$

a) Hallar el esperado del estimador mínimo cuadrático de β_1 usando el modelo (2), pero considerando que realmente (1) es el que se cumple.
b) Hallar el esperado del estimador minimocuadrático de la varianza estimada usando el modelo (2).

CAPÍTULO 3

DIAGNÓSTICOS DE REGRESIÓN

En este capítulo se estudiarán diversos diagnósticos de regresión que nos permitiran verificar si las suposiciones del modelo de regresión se cumplen. Algunos de estos diagnósticos están basados en medidas que envuelven residuales y otros en plots de los residuales.

3.1 Residuales y detección de "outliers".

Consideremos el modelo de regresión lineal múltiple $\mathbf{Y=XB+e}$, donde $E(\mathbf{e})=\mathbf{0}$ y $Var(\mathbf{e})=\sigma^2\mathbf{I}$.
Luego, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, pero como $\widehat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}X'Y}$, se tiene que $\widehat{\mathbf{Y}} = \mathbf{X(X'X)^{-1}X'Y} = \mathbf{HY}$, ésta es la razón por la que a **H** se le llama la matriz HAT (sombrero), ya que actúa como una transformación de **Y** a $\widehat{\mathbf{Y}}$. En particular, $\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$, donde h_{ij} es el elemento de la matríz H que está en la i-ésima fila y j-ésima columna.

Luego, el vector de residuales $\widehat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \mathbf{(I - H)Y}$. En particular, $\widehat{e}_i = y_i - \sum_{j=1}^{n} h_{ij} y_j$.

3.1.1 Media y Varianza del vector de residuales

Notar que

$$E(\widehat{\mathbf{e}}) = \mathbf{(I - H)}E(\mathbf{Y}) = \mathbf{(I - H)XB} = \mathbf{XB - HXB} = \mathbf{XB - X(X'X)^{-1}X'XB} = \mathbf{XB - XB} = \mathbf{0}$$

Por otro lado,

$$Var(\widehat{\mathbf{e}}) = Var[\mathbf{(I - H)Y}] = \mathbf{(I - H)}Var(\mathbf{Y})\mathbf{(I - H)'} = \sigma^2\mathbf{(I - H)(I - H)'} = \sigma^2\mathbf{(I - H)^2} = \sigma^2\mathbf{(I - H)}$$

Aqui se ha usado el hecho que **I-H** es simétrica e idempotente, como se vió en la sección 2.2.3.

En particular, $Var(\widehat{e}_i) = \sigma^2(1 - h_{ii})$ **.** Esta varianza es estimada por $s^2(1\text{-}h_{ii})$.

Asimismo, $Cov(\widehat{e}_i, \widehat{e}_j) = -h_{ij}\sigma^2$

Notar que :

a) Tanto los errores e_i como los residuales tienen media 0.
b) La varianza de los errores es constante, pero la de los residuales no lo es.
c) Los errores no están correlacionados, pero los residuales si lo están.

3.1.2 Residuales Estudentizados internamente

Para reducir el efecto de las varianzas de los residuales es más conveniente trabajar con versiones estandarizadas de ellos. Así, el ***residual estudentizado*** **internamente** se define por

$$r_i^* = \frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}} \qquad (3.1)$$

La covarianza de los residuales estudentizados es igual a

$$Cov(r_i^*, r_j^*) = Cov(\frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}}, \frac{\hat{e}_j}{\sigma\sqrt{1-h_{jj}}}) = \frac{Cov(\hat{e}_i, \hat{e}_j)}{\sigma^2\sqrt{(1-h_{ii})(1-h_{jj})}} = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}$$

En algunos programas estadísticos como MINITAB y el toolbox estadístico de MATLAB los r_i^* son llamados **residuales estandarizados**.

3.1.3 "Outliers", puntos de leverage alto y valores influenciales

Una observación (y*,x*$_1$,……..x*$_p$) es considerado un "outlier" si está bastante alejado de la mayoría de los datos sea en la dirección vertical o en la horizontal. Sin embargo, la mayoría de los textos llaman "outlier" a un valor alejado solamente en la dirección vertical y **punto de leverage alto** a una observación alejada en la dirección horizontal.
Una observación (y*,x*$_1$,……..x*$_p$) es considerado un **valor influencial** si su presencia afecta tremendamente el comportamiento del modelo. Por ejemplo, en el caso de regresión simple remover un valor influencial podría cambiar dramáticamente el valor de la pendiente.

Consideremos el siguiente conjunto de datos, consistente de 8 observaciones

X	4	5	7	9	12	14	16	35
Y	6	7	12	15	18	21	28	65

La figura 3.1 muestra el plot de los datos. Notar que el punto O es un "outlier" y punto de leverage alto, pero a través de cálculos mostraremos que no es un valor influencial.

Primero, calcularemos la ecuación de regresión con el dato "outlier"

```
> l1=lm(y~x)
> summary(l1)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-2.8825 -0.3140  0.4765  1.1130  1.4595

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -2.80152    1.03618  -2.704   0.0354 *
x            1.90600    0.06567  29.026 1.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.727 on 6 degrees of freedom
Multiple R-Squared: 0.9929,     Adjusted R-squared: 0.9918
F-statistic: 842.5 on 1 and 6 DF,  p-value: 1.108e-07
```

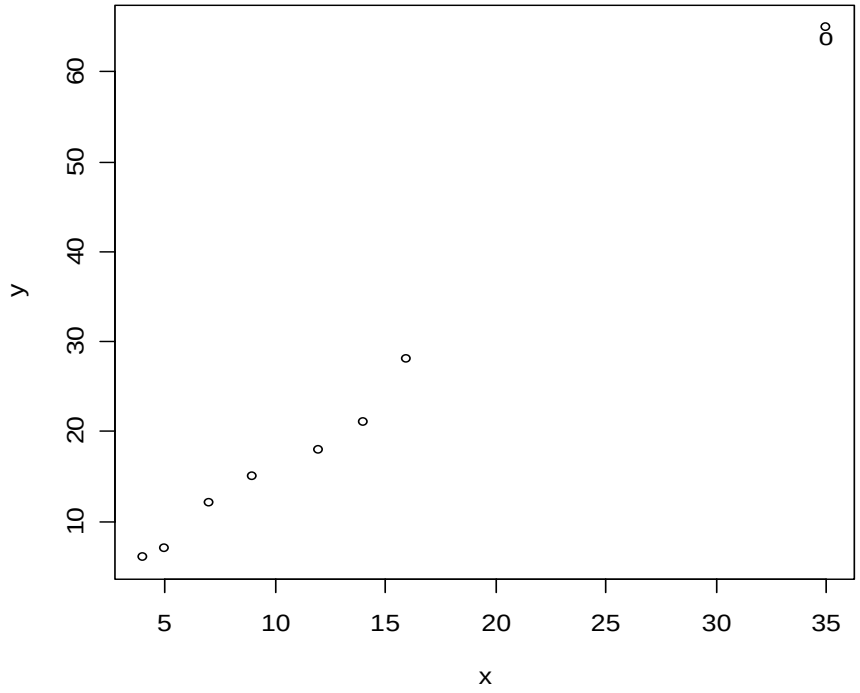


Figura 3.1. Ejemplo de una observación que es "outlier" y punto leverage alto pero que no es influencial.

Por otro lado la regresión sin el dato "outlier" es:

```
> x1=x[-8]
> y1=y[-8]
> l2=lm(y1~x1)
> summary(l2)

Call:
lm(formula = y1 ~ x1)

Residuals:
     1      2      3      4      5      6      7
 0.1034 -0.5818  1.0477  0.6773 -1.3784 -1.7489  1.8807
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8443     1.3465  -0.627    0.558
x1            1.6852     0.1286  13.101 4.62e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.442 on 5 degrees of freedom
Multiple R-Squared: 0.9717,     Adjusted R-squared: 0.966
F-statistic: 171.6 on 1 and 5 DF,  p-value: 4.625e-05
```

Notar que la pendiente y el R^2 han cambiado solo ligeramente. En consecuencia, la observación es un "outlier" y punto de leverage alto pero no es es influencial.

Supongamos ahora que al conjunto de datos anterior y al cual se le eliminó el "outlier", se le agrega el dato (35,22) que es considerado un punto de leverage alto. El plot del conjunto de datos es mostrado en la figura 3.2, donde la observación 0 representa el dato con leverage alto. Mostraremos que esta observación si resulta ser influencial.

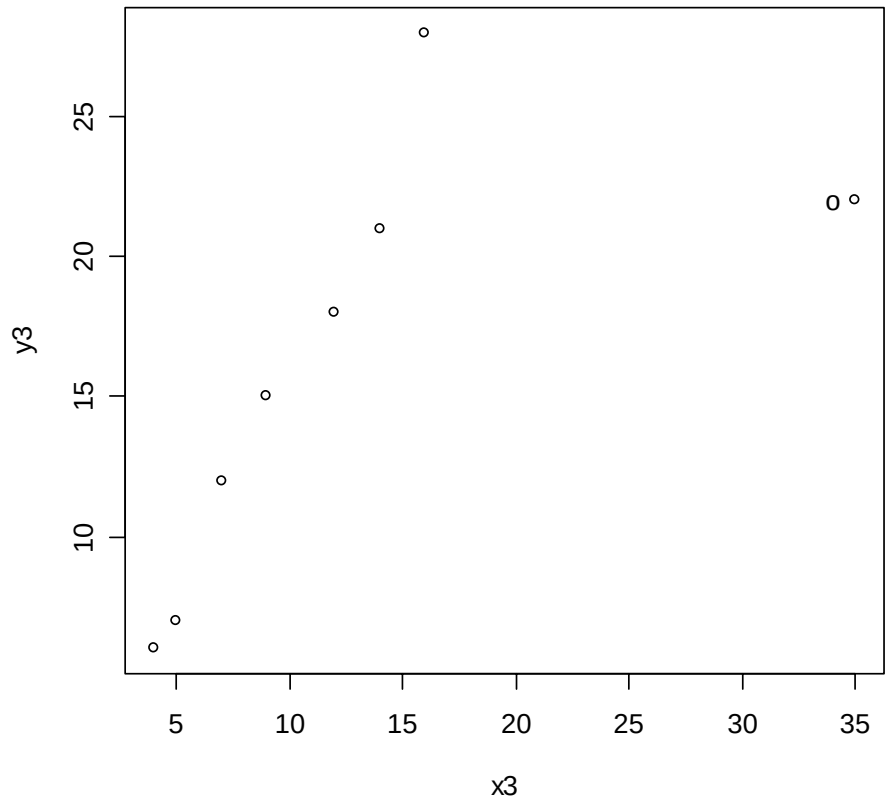


Figura 3.2. Ejemplo de una observación que es punto de leverage alto y que también es influencial.

La ecuación de regresión considerando el dato de leverage alto es

```
> l3=lm(y3~x3)
> summary(l3)

Call:
lm(formula = y3 ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7487 -5.1957 -0.1435  2.7556 10.1772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4642     3.6183   2.616   0.0398 *
x3            0.5224     0.2293   2.278   0.0629 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.03 on 6 degrees of freedom
Multiple R-Squared: 0.4638,     Adjusted R-squared: 0.3745
F-statistic: 5.191 on 1 and 6 DF,  p-value: 0.06295
```

Se puede observar el gran efecto sobre el R^2 que baja de 97.2% a 46.4% y un cambio drástico en la pendiente que cambia de 1.69 a 0.522..

En consecuencia un "outlier" vertical y/o punto de leverage alto puede ser influencial o no serlo. Por otro lado si una observación es influencial entonces es un "outlier" vertical o un punto de leverage alto.

3.1.4 Residuales estudentizados externamente

Supongamos que la i-ésima observación es eliminada del conjunto de datos y que se ajusta el modelo lineal con las n-1 observaciones que quedan. Sean $\widehat{\boldsymbol{\beta}}_{(i)}$ y $s^2_{(i)}$ las estimaciones de los parámetros del modelo y de la varianza de los errores respectivamente. Usando la siguiente identidad debido a Gauss

$$(\mathbf{X'}_{(i)}\,\mathbf{X}_{(i)})^{-1} = (\mathbf{X'X})^{-1} + \frac{(\mathbf{X'X})^{-1}\mathbf{x}_i\mathbf{x'}_i\,(\mathbf{X'X})^{-1}}{1-h_{ii}} \qquad (3.2)$$

donde $\mathbf{X}_{(i)}$ representa a la matriz $\mathbf{X}$ sin su i-ésima fila $\mathbf{x'}_i$., se puede establecer las siguientes relaciones entre $\hat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}}_{(i)}$ y entre s^2 y $s^2_{(i)}$

i) $\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \dfrac{(\mathbf{X'X})^{-1}\mathbf{x}_i\hat{e}_i}{1-h_{ii}}$

ii) $s^2_{(i)} = \dfrac{n-p-1}{n-p-2}s^2 - \dfrac{\hat{e}_i^2}{(n-p-2)(1-h_{ii})}$

La identidad de Gauss es un caso particular de la **Identidad de Sherman-Morrison-Woodburry** (1950)

$$\left(\mathbf{A} \pm \mathbf{uv'}\right)^{-1} = \mathbf{A}^{-1} \mp \frac{\mathbf{A^{-1}uv'A^{-1}}}{1 \pm \mathbf{v'A^{-1}u}} \qquad (3.3)$$

donde **A** es una matriz cuadrada nosingular n x n, y **u** y **v** son dos vectores de dimensión n.
En efecto, puesto que $\mathbf{X}_{(i)}'\mathbf{X}_{(i)} = \mathbf{X'X} - \mathbf{x}_i\mathbf{x}_i'$, donde $\mathbf{x'}_i$ es la i-ésima fila de **X**, se puede tomar **A**=**X'X** y **u**=**v**=$\mathbf{x_i}$. y se obtiene (3.2).

Si $\tilde{y}_i$ representa el valor estimado de la variable de respuesta para la i-ésima observación entonces $\tilde{y}_i = \mathbf{x'_i}\,\hat{\boldsymbol{\beta}}_{(i)}$. Como la i-ésima observación no fue usada en la estimación del modelo entonces y_i y $\tilde{y}_i$ son independientes. Luego la varianza del residual $y_i - \tilde{y}_i$ está dada por

$$Var(y_i - \tilde{y}_i) = Var(y_i) + Var(\tilde{y}_i) = \sigma^2 + \sigma^2 \mathbf{x'_i}\,(\mathbf{X}'_{(i)}\,\mathbf{X}_{(i)})^{-1}\mathbf{x}_i \quad (3.4)$$

Estimando σ^2 por $s^2_{(i)}$ y considerando que si y_i no es un outlier entonces $E(y_i - \tilde{y}_i)=0$ se obtiene

$$t_i = \frac{y_i - \tilde{y}_i}{s_{(i)}\sqrt{1 + \mathbf{x'_i}\,(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{x_i}}} \qquad (3.5)$$

t_i es llamado un **residual estudentizado externamente** y tiene n-p-2 grados de libertad.

Propiedad: Relación entre el residual usual y el residual usando un modelo eliminando la i-ésima observación

$$y_i - \tilde{y}_i = \frac{\hat{e}_i}{1 - h_{ii}} \quad (3.6)$$

Prueba: Sustituyendo $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}$ en

$$y_i - \tilde{y}_i = y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}_{(i)} \qquad (3.7)$$

y usando luego la identidad de Gauss (3.2) se obtiene

$$y_i - \tilde{y}_i = y_i - \mathbf{x}_i'\left[\left(\mathbf{X'X}\right)^{-1} + \frac{\left(\mathbf{X'X}\right)^{-1}\mathbf{x_i}\mathbf{x}_i'\left(\mathbf{X'X}\right)^{-1}}{1 - h_{ii}}\right]\mathbf{X}'_{(i)}\mathbf{y}_{(i)}$$

$$= y_i - \mathbf{x}_i'\left(\mathbf{X'X}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)} - \frac{h_{ii}\mathbf{x}_i'\left(\mathbf{X'X}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}}{1 - h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - (1-h_{ii})\mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{X'_{(i)}}\mathbf{y_{(i)}} - h_{ii}\mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{X'_{(i)}}\mathbf{y_{(i)}}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - \mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{X'_{(i)}}\mathbf{y_{(i)}}}{1-h_{ii}}$$

Si se usa luego el hecho que $\mathbf{X'_{(i)}}\mathbf{y_{(i)}} + \mathbf{x_i}y_i = \mathbf{X'y}$, la anterior relación es equivalente a .

$$y_i - \tilde{y}_i = \frac{(1-h_{ii})y_i - \mathbf{x_i'}(\mathbf{X'X})^{-1}(\mathbf{X'y} - \mathbf{x_i}y_i)}{1-h_{ii}}$$

como $\mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{X'y} = \hat{y}_i$ se obtiene

$$y_i - \tilde{y}_i = \frac{(1-h_{ii})y_i - \hat{y}_i + h_{ii}y_i}{1-h_{ii}}$$

$$= \frac{y_i - \hat{y}}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}}$$

Lo cual concluye la prueba.

Asímismo, se puede establecer la siguiente relación entre los distintos tipos de residuales

$$t_i = \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}} = r_i^*\left(\frac{n-p-2}{n-p-1-r_i^{*2}}\right)^{1/2} \qquad (3.8)$$

3.2 Diagnósticos para detectar “outliers” y puntos de leverage alto

Ahora consideraremos diagnósticos basados en medidas y que servirán para detectar si una observación es un ”outlier” o un punto de leverage alto. Los diagnósticos más básicos son:

Si $|h_{ii}|>2p/n$ (algunos usan 3p/n. Aquí p es el número de parámetros) entonces la i-ésima observación es considerado un “punto leverage” y pudiera ser influencial

Si $|t_i|>2$ (o si $|r_i|>2$) entonces la i-ésima observación es considerada un “outlier” y también puede ser influencial.

A continuación definiremos otros diagnósticos más sofisticados:

i) La Distancia Cook (Cook, 1977**):** Mide el cambio que ocurriría en el vector $\hat{\boldsymbol{\beta}}$ de coeficientes estimados de regresión (y por lo tanto en el valor ajustado de la variable de respuesta) si la i-ésima observación fuera omitida. Se calcula por

$$CD_i^2 = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{ps^2} = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{ps^2} = r*_i^2 \frac{h_{ii}}{p(1-h_{ii})} \quad (3.9)$$

Notar que si el residual estandarizado es muy grande y si el valor leverage es grande, es decir si la observación está bien alejado en la dirección vertical y horizontal entonces su distancia Cook es bien grande. En general un CD_i^2 > 1 indica que la i-ésima observación es potencialmente influencial. Una observación con CD_i^2 <0.1 no merece ninguna discusión y si su CD_i^2 <.0.5 merece un poco de atención. Más especificamente una observación con CD_i^2 > F(0.50,p,n-p) es considerado como un valor influencial, la razón es que $\boldsymbol{\beta}$ cae en un elipsoide de confianza centrado en $\hat{\boldsymbol{\beta}}$ de radio F(α,p,n-p). Aquí p es el número de coeficientes en el modelo. Sin embargo si todos los CD_i^2 son menores que 1 es mejor plotear los valores CD_i^2 para detectar si hay observaciones con valores grandes comparados con los demás.

ii) DFFITS (Belsley, Kuh, y Welsch, 1980). Es similar a la Distancia Cook, excepto por un factor de escala y el remplazo de la varianza estimada s^2 por $s_{(i)}^2$, la varianza estimada del error excluyendo la i-ésima observación en los cálculos. Más precisamente,

$$DFFITS_i^2 = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{s_{(i)}^2} = t_i^2 \frac{h_{ii}}{(1-h_{ii})} \quad (3.10)$$

Un $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ indica un posible valor influencial. Notar que

$$CD_i^2 = \frac{r_i^2}{pt_i^2} DFFITS_i^2 \quad (3.11)$$

iii) DFBETAS (Belsley, Kuh, y Welsch, 1980**).** Mide la influencia de la i-ésima observación en cada uno de los coeficientes de regresión. Se calcula por

$$(DFBETAS)_{ji} = \frac{\beta_j - \beta_{j,(i)}}{s_{(i)}\sqrt{c_{jj}}} \quad (3.12)$$

(i=1,..n, j=0,….p), donde c_{jj} es el j-ésimo elemento de la diagonal de $\mathbf{(X'X)^{-1}}$.

Un $|DFBETAS|_{ji} > \frac{2}{\sqrt{n}}$ indica un posible valor influencial.

iv) COVRATIO (Belsley, Kuh, y Welsch, 1980**)**

Mide el efecto en la variabilidad de los coeficientes de regresión al remover la i-ésima observación. Se define por

$$COVRATIO_i = \frac{\det[s^2_{(i)}(X'_{(i)}X_{(i)})^{-1}]}{\det[s^2(X'X)^{-1}]} \qquad (3.13)$$

para i=1,….n. Donde det[A] significa el determinante de la matriz A. Usando propiedades de determinantes, se puede obtener la siguiente equivalente fórmula

$$(COVRATIO)_i = \left(\frac{s^2_{(i)}}{s^2}\right)^p \frac{1}{(1-h_{ii})} \qquad (3.14)$$

Si $(COVRATIO)_i$ >1+3p/n o si $(COVRATIO)_i$<1-3p/n entonces la i-ésima observación tiene un valor influencial grande.

Ejemplo 1: Aplicar los diagnósticos de regresión al conjunto de datos **millaje.**

Aquí hemos usado R, SAS tambien da una lista completa de diagnósticos pero MINITAB no da los DFBETAS ni los COVRATIO.

Obs	Dep Var MPG	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	**65.4000**	**53.4146**	**1.267**	**11.9854**	**3.426**	**3.499**	||******|	**0.335**
2	**56.0000**	**49.7766**	**1.029**	**6.2234**	**3.505**	**1.776**	|***|	**0.054**
3	**55.9000**	**49.7766**	**1.029**	**6.1234**	**3.505**	**1.747**	|***|	**0.053**
4	49.0000	45.3013	0.742	3.6987	3.577	1.034	|**|	0.009
5	**46.5000**	**50.2870**	**1.092**	**-3.7870**	**3.486**	**-1.086**	**|	**0.023**
6	46.2000	45.3482	0.725	0.8518	3.580	0.238	||	0.000
7	**45.4000**	**49.7766**	**1.029**	**-4.3766**	**3.505**	**-1.249**	**|	**0.027**
8	**59.2000**	**47.2349**	**1.213**	**11.9651**	**3.445**	**3.473**	|******|	**0.299**
9	**53.3000**	**47.2349**	**1.213**	**6.0651**	**3.445**	**1.760**	|***|	**0.077**
10	43.4000	41.9529	0.639	1.4471	3.596	0.402	||	0.001
11	41.1000	44.4650	0.646	-3.3650	3.595	-0.936	*|	0.006
12	**40.9000**	**39.5790**	**1.263**	**1.3210**	**3.428**	**0.385**	**|**	**0.004**
13	40.9000	38.8124	0.976	2.0876	3.520	0.593	|*	0.005
14	40.4000	44.4650	0.646	-4.0650	3.595	-1.131	**|	0.008
15	39.6000	45.6039	0.735	-6.0039	3.578	-1.678	***|	0.024
16	39.3000	44.4650	0.646	-5.1650	3.595	-1.437	**|	0.013
17	38.9000	42.5103	0.605	-3.6103	3.602	-1.002	**|	0.006
18	**38.8000**	**39.5790**	**1.263**	**-0.7790**	**3.428**	**-0.227**	**|**	**0.001**
19	38.2000	42.5103	0.605	-4.3103	3.602	-1.197	**|	0.008
20	42.2000	38.4951	0.631	3.7049	3.598	1.030	|**	0.007
21	40.9000	38.0473	0.651	2.8527	3.594	0.794	|*	0.004
22	40.7000	42.5157	0.729	-1.8157	3.579	-0.507	*|	0.002

23	40.0000	37.8978	0.682	2.1022	3.589	0.586	|	|* |	0.002
24	39.3000	40.2540	0.504	-0.9540	3.618	-0.264	|	| |	0.000
25	38.8000	38.0856	0.647	0.7144	3.595	0.199	|	| |	0.000
26	**38.4000**	**38.8139**	**1.152**	**-0.4139**	**3.466**	**-0.119**	**|**	**| |**	**0.000**
27	38.4000	37.7657	0.849	0.6343	3.553	0.179	|	| |	0.000
28	38.4000	38.0473	0.651	0.3527	3.594	0.098	|	| |	0.000
29	**29.5000**	**38.5108**	**0.623**	**-9.0108**	**3.599**	**-2.504**	| ***** |	**|**	**0.038**
30	**46.9000**	**43.5274**	**1.173**	**3.3726**	**3.459**	**0.975**	**|**	**|* |**	**0.022**
31	36.3000	34.9973	0.659	1.3027	3.593	0.363	|	| |	0.001
32	36.1000	39.0892	0.661	-2.9892	3.592	-0.832	|	*| |	0.005
33	36.1000	39.2925	0.549	-3.1925	3.611	-0.884	|	*| |	0.004
34	35.4000	36.0564	0.512	-0.6564	3.617	-0.181	|	| |	0.000
35	35.3000	35.8061	0.649	-0.5061	3.595	-0.141	|	| |	0.000
36	35.1000	39.4107	0.564	-4.3107	3.609	-1.194	|	**| |	0.007
37	35.1000	37.8083	0.448	-2.7083	3.625	-0.747	|	*| |	0.002
38	35.0000	37.9647	0.497	-2.9647	3.619	-0.819	|	*| |	0.003
39	33.2000	34.1686	0.598	-0.9686	3.603	-0.269	|	| |	0.000
40	32.9000	34.1686	0.598	-1.2686	3.603	-0.352	|	| |	0.001
41	32.3000	30.8137	0.828	1.4863	3.558	0.418	|	| |	0.002
42	32.2000	34.8852	0.512	-2.6852	3.617	-0.742	|	*| |	0.002
43	32.2000	34.9947	0.465	-2.7947	3.623	-0.771	|	*| |	0.002
44	32.2000	34.0747	0.524	-1.8747	3.615	-0.519	|	*| |	0.001
45	32.2000	35.2763	0.576	-3.0763	3.607	-0.853	|	*| |	0.004
46	31.5000	35.5677	0.478	-4.0677	3.621	-1.123	|	**| |	0.004
47	31.5000	34.4756	0.454	-2.9756	3.624	-0.821	|	*| |	0.002
48	31.4000	34.2879	0.491	-2.8879	3.620	-0.798	|	*| |	0.002
49	31.4000	34.9234	0.444	-3.5234	3.626	-0.972	|	*| |	0.003
50	31.2000	30.9076	0.857	0.2924	3.551	0.082	|	| |	0.000
51	33.7000	29.7337	0.610	3.9663	3.601	1.101	|	|** |	0.007
52	32.6000	29.7337	0.610	2.8663	3.601	0.796	|	|* |	0.004
53	31.3000	29.7337	0.610	1.5663	3.601	0.435	|	| |	0.001
54	31.3000	29.3738	0.633	1.9262	3.598	0.535	|	|* |	0.002
55	**30.4000**	**23.9641**	**1.094**	**6.4359**	**3.485**	**1.847**	**|**	**|*** |**	**0.067**
56	28.9000	26.4784	0.715	2.4216	3.582	0.676	|	|* |	0.004
57	28.0000	27.4881	0.670	0.5119	3.591	0.143	|	| |	0.000
58	28.0000	31.4862	0.929	-3.4862	3.533	-0.987	|	*| |	0.013
59	28.0000	29.7337	0.610	-1.7337	3.601	-0.481	|	| |	0.001
60	28.0000	30.4341	0.853	-2.4341	3.552	-0.685	|	*| |	0.005
61	28.0000	28.8107	0.737	-0.8107	3.578	-0.227	|	| |	0.000
62	27.7000	27.1006	0.599	0.5994	3.603	0.166	|	| |	0.000
63	25.6000	24.7507	0.902	0.8493	3.540	0.240	|	| |	0.001
64	25.3000	23.2965	0.787	2.0035	3.567	0.562	|	|* |	0.003
65	23.9000	23.4217	0.741	0.4783	3.577	0.134	|	| |	0.000
66	23.6000	23.4906	0.679	0.1094	3.589	0.030	|	| |	0.000
67	**23.6000**	**24.0105**	**1.520**	**-0.4105**	**3.321**	**-0.124**	**|**	**| |**	**0.001**
68	**23.6000**	**23.0093**	**0.628**	**0.5907**	**3.598**	**0.164**	**|**	**| |**	**0.000**
69	23.6000	22.8059	0.726	0.7941	3.580	0.222	|	| |	0.000
70	23.6000	22.8684	0.684	0.7316	3.588	0.204	|	| |	0.000
71	23.5000	20.7522	1.295	2.7478	3.415	0.805	|	|* |	0.019
72	**23.4000**	**19.9118**	**1.692**	**3.4882**	**3.237**	**1.077**	**|**	**|** |**	**0.063**
73	23.4000	22.7989	0.824	0.6011	3.559	0.169	|	| |	0.000

74	23.1000	22.8458	0.781	0.2542	3.568	0.071	|	|	0.000
75	**22.9000**	**18.7231**	**1.281**	**4.1769**	**3.421**	**1.221**	|	|** |	**0.042**
76	22.9000	19.2081	1.113	3.6919	3.479	1.061	|	|** |	0.023
77	19.5000	18.6925	0.888	0.8075	3.543	0.228	|	|	0.001
78	**18.1000**	**20.6117**	**2.033**	**-2.5117**	**3.035**	**-0.828**	|	*| |	**0.061**
79	17.2000	19.0194	1.082	-1.8194	3.489	-0.521	|	*| |	0.005
80	**17.0000**	**20.7779**	**1.593**	**-3.7779**	**3.287**	**-1.149**	|	**| |	**0.062**
81	**16.7000**	**19.3010**	**1.871**	**-2.6010**	**3.137**	**-0.829**	|	*| |	**0.049**
82	**13.2000**	**12.7102**	**1.636**	**0.4898**	**3.266**	**0.150**	|	| |	**0.001**

Obs	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	VOL Dfbetas	HP Dfbetas	SP Dfbetas	WT Dfbetas
1	**3.7900**	**0.1204**	**0.5107**	**1.4021**	**1.1286**	**0.3421**	**1.1002**	**-1.0801**	**-1.2007**
2	**1.8014**	**0.0794**	**0.9408**	**0.5289**	**0.4117**	**0.1168**	**0.3936**	**-0.3937**	**-0.4227**
3	**1.7712**	**0.0794**	**0.9472**	**0.5200**	**0.4048**	**0.1148**	**0.3871**	**-0.3871**	**-0.4156**
4	1.0346	0.0413	1.0383	0.2146	0.0142	0.0628	0.0176	-0.0026	-0.0793
5	**-1.0877**	**0.0894**	**1.0852**	**-0.3408**	**-0.2762**	**-0.0706**	**-0.2633**	**0.2655**	**0.2751**
6	0.2365	0.0394	1.1072	0.0479	0.0028	0.0098	0.0030	0.0001	-0.0162
7	**-1.2534**	**0.0794**	**1.0468**	**-0.3680**	**-0.2865**	**-0.0813**	**-0.2739**	**0.2739**	**0.2941**
8	**3.7569**	**0.1103**	**0.5120**	**1.3229**	**0.6042**	**-0.9409**	**0.4143**	**-0.5252**	**-0.2150**
9	**1.7852**	**0.1103**	**0.9771**	**0.6286**	**0.2871**	**-0.4471**	**0.1969**	**-0.2496**	**-0.1021**
10	0.4002	0.0306	1.0897	0.0711	-0.0108	0.0168	-0.0098	0.0144	-0.0103
11	-0.9352	0.0312	1.0407	-0.1679	-0.0799	-0.0135	-0.0724	0.0708	0.0925
12	**0.3833**	**0.1195**	**1.2008**	**0.1412**	**-0.0790**	**-0.0946**	**-0.0883**	**0.0881**	**0.0772**
13	0.5906	0.0714	1.1236	0.1637	-0.0988	0.0457	-0.0888	0.1039	0.0358
14	-1.1327	0.0312	1.0135	-0.2034	-0.0968	-0.0163	-0.0877	0.0858	0.1121
15	-1.6984	0.0405	0.9235	-0.3488	-0.2276	-0.0140	-0.2046	0.2117	0.2174
16	-1.4468	0.0312	0.9620	-0.2598	-0.1236	-0.0208	-0.1121	0.1096	0.1432
17	-1.0022	0.0274	1.0279	-0.1683	0.0055	-0.0228	0.0073	-0.0152	0.0341
18	**-0.2259**	**0.1195**	**1.2084**	**-0.0832**	**0.0465**	**0.0558**	**0.0520**	**-0.0519**	**-0.0455**
19	-1.1999	0.0274	0.9993	-0.2014	0.0066	-0.0272	0.0087	-0.0182	0.0409
20	1.0302	0.0299	1.0267	0.1808	-0.0646	0.0702	-0.0573	0.0697	0.0059
21	0.7918	0.0318	1.0582	0.1434	-0.0703	0.0342	-0.0659	0.0752	0.0262
22	-0.5048	0.0399	1.0934	-0.1029	-0.0643	-0.0462	-0.0627	0.0631	0.0693
23	0.5833	0.0349	1.0817	0.1108	-0.0518	0.0359	-0.0464	0.0551	0.0135
24	-0.2621	0.0191	1.0834	-0.0365	-0.0010	-0.0045	0.0005	-0.0006	0.0053
25	0.1975	0.0314	1.0993	0.0356	-0.0191	-0.0029	-0.0192	0.0209	0.0110
26	**-0.1186**	**0.0995**	**1.1844**	**-0.0394**	**0.0149**	**0.0330**	**0.0191**	**-0.0174**	**-0.0204**
27	0.1774	0.0540	1.1261	0.0424	-0.0138	0.0283	-0.0102	0.0140	-0.0028
28	0.0975	0.0318	1.1019	0.0177	-0.0087	0.0042	-0.0081	0.0093	0.0032
29	**-2.5951**	**0.0291**	**0.7192**	**-0.4492**	**0.1645**	**-0.1604**	**0.1481**	**-0.1778**	**-0.0220**
30	**0.9746**	**0.1032**	**1.1187**	**0.3306**	**0.2568**	**0.0056**	**0.2237**	**-0.2568**	**-0.1632**
31	0.3605	0.0325	1.0941	0.0661	-0.0372	0.0246	-0.0336	0.0380	0.0167
32	-0.8304	0.0327	1.0550	-0.1528	-0.0745	-0.0869	-0.0745	0.0754	0.0828
33	-0.8828	0.0226	1.0379	-0.1343	-0.0712	-0.0194	-0.0619	0.0689	0.0567
34	-0.1803	0.0197	1.0866	-0.0255	0.0115	-0.0017	0.0114	-0.0124	-0.0065
35	-0.1399	0.0316	1.1009	-0.0253	0.0075	-0.0156	0.0056	-0.0075	0.0010

36	-1.1978	0.0238	0.9960	-0.1871	-0.0980	-0.0197	-0.0829	0.0950	0.0717
37	-0.7449	0.0151	1.0452	-0.0921	-0.0066	-0.0029	-0.0007	0.0036	0.0052
38	-0.8175	0.0185	1.0411	-0.1122	-0.0018	0.0484	0.0112	-0.0038	-0.0162
39	-0.2672	0.0268	1.0918	-0.0443	0.0172	0.0274	0.0221	-0.0188	-0.0260
40	-0.3500	0.0268	1.0881	-0.0581	0.0225	0.0359	0.0289	-0.0246	-0.0341
41	0.4155	0.0514	1.1127	0.0967	-0.0773	-0.0125	-0.0742	0.0797	0.0576
42	-0.7403	0.0197	1.0505	-0.1048	-0.0011	-0.0458	0.0004	0.0027	0.0028
43	-0.7693	0.0162	1.0438	-0.0987	0.0024	-0.0136	0.0083	-0.0023	-0.0116
44	-0.5161	0.0206	1.0711	-0.0749	0.0310	0.0333	0.0380	-0.0332	-0.0418
45	-0.8513	0.0249	1.0441	-0.1361	0.0128	0.0818	0.0318	-0.0171	-0.0542
46	-1.1252	0.0171	1.0000	-0.1485	-0.0196	0.0088	-0.0054	0.0187	-0.0107
47	-0.8192	0.0155	1.0377	-0.1027	0.0263	0.0168	0.0341	-0.0280	-0.0375
48	-0.7959	0.0180	1.0430	-0.1079	0.0192	-0.0439	0.0192	-0.0182	-0.0108
49	-0.9715	0.0148	1.0187	-0.1191	0.0061	-0.0042	0.0138	-0.0069	-0.0182
50	0.0818	0.0550	1.1292	0.0197	-0.0156	-0.0056	-0.0154	0.0162	0.0127
51	1.1029	0.0279	1.0144	0.1869	-0.0247	-0.0723	-0.0442	0.0224	0.0905
52	0.7940	0.0279	1.0538	0.1345	-0.0177	-0.0521	-0.0318	0.0162	0.0652
53	0.4326	0.0279	1.0848	0.0733	-0.0097	-0.0284	-0.0173	0.0088	0.0355
54	0.5329	0.0300	1.0802	0.0937	-0.0038	0.0428	-0.0033	-0.0008	0.0107
55	**1.8767**	**0.0897**	**0.9350**	**0.5891**	**-0.3824**	**0.2079**	**-0.3056**	**0.3776**	**0.1730**
56	0.6736	0.0383	1.0775	0.1345	-0.0672	0.0567	-0.0507	0.0648	0.0279
57	0.1416	0.0337	1.1033	0.0264	-0.0114	0.0139	-0.0097	0.0103	0.0079
58	-0.9867	0.0647	1.0710	-0.2595	-0.0242	0.1519	0.0149	0.0245	-0.0963
59	-0.4790	0.0279	1.0818	-0.0812	0.0107	0.0314	0.0192	-0.0097	-0.0393
60	-0.6829	0.0545	1.0951	-0.1640	0.0147	0.0954	0.0382	-0.0143	-0.0853
61	-0.2252	0.0407	1.1091	-0.0464	0.0170	0.0034	0.0207	-0.0156	-0.0283
62	0.1653	0.0269	1.0950	0.0275	-0.0157	0.0071	-0.0136	0.0151	0.0114
63	0.2385	0.0610	1.1327	0.0608	-0.0055	-0.0121	-0.0108	0.0030	0.0284
64	0.5592	0.0464	1.0967	0.1233	-0.0241	0.0444	-0.0217	0.0164	0.0376
65	0.1329	0.0411	1.1120	0.0275	-0.0064	0.0038	-0.0067	0.0049	0.0118
66	0.0303	0.0345	1.1056	0.0057	-0.0009	0.0011	-0.0008	0.0006	0.0018
67	**-0.1228**	**0.1732**	**1.2899**	**-0.0562**	**0.0111**	**0.0513**	**0.0170**	**-0.0125**	**-0.0305**
68	0.1631	0.0296	1.0982	0.0285	-0.0067	0.0034	-0.0054	0.0054	0.0092
69	0.2205	0.0395	1.1079	0.0447	-0.0072	0.0229	-0.0031	0.0046	0.0046
70	0.2026	0.0351	1.1033	0.0386	-0.0071	0.0158	-0.0040	0.0050	0.0065
71	0.8027	0.1258	1.1707	0.3045	-0.0236	0.1381	0.0289	0.0196	-0.0810
72	**1.0786**	**0.2145**	**1.2597**	**0.5637**	**-0.1260**	**-0.2019**	**-0.0820**	**0.1434**	**0.0472**
73	0.1678	0.0509	1.1227	0.0389	-0.0044	0.0252	-0.0005	0.0020	0.0006
74	0.0708	0.0457	1.1184	0.0155	-0.0020	0.0092	-0.0005	0.0011	0.0008
75	**1.2250**	**0.1230**	**1.1039**	**0.4587**	**-0.0118**	**0.2285**	**-0.0020**	**-0.0218**	**0.0669**
76	1.0621	0.0928	1.0932	0.3398	-0.0327	-0.0212	-0.0519	0.0138	0.1490
77	0.2265	0.0591	1.1307	0.0567	0.0005	0.0121	0.0020	-0.0042	0.0104
78	**-0.8259**	**0.3097**	**1.4789**	**-0.5532**	**-0.2271**	**0.1223**	**-0.2680**	**0.2168**	**0.2449**
79	-0.5190	0.0877	1.1497	-0.1610	-0.0557	-0.0540	-0.0760	0.0613	0.0679
80	**-1.1518**	**0.1903**	**1.2091**	**-0.5584**	**-0.1786**	**0.3170**	**-0.1807**	**0.1674**	**0.0829**
81	**-0.8274**	**0.2624**	**1.3839**	**-0.4935**	**-0.2092**	**-0.2396**	**-0.2846**	**0.2187**	**0.3211**
82	**0.1490**	**0.2007**	**1.3336**	**0.0747**	**0.0359**	**-0.0182**	**0.0340**	**-0.0386**	**-0.0086**

Nota: Las observaciones en negritas pueden ser influenciales según al menos uno de los diagnosticos que se describen a continuación:

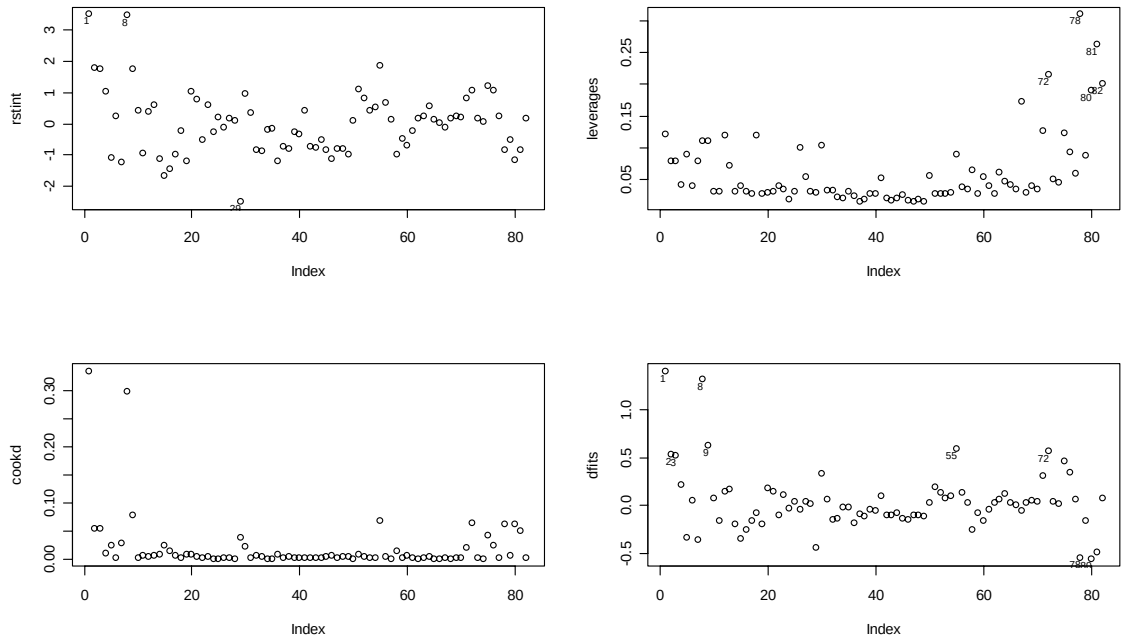
Potentially influential observations of
lm(formula = mpg ~ ., data = millaje) :

	dfb.1_	dfb.sp	dfb.wt	dfb.vol	dfb.hp	dffit	cov.r	cook.d	hat
1	1.13_*	-1.08_*	-1.20_*	0.34_*	1.10_*	1.40_*	0.51_*	0.34	0.12
2	0.41_*	-0.39_*	-0.42_*	0.12	0.39_*	0.53	0.94	0.05	0.08
3	0.40_*	-0.39_*	-0.42_*	0.11	0.39_*	0.52	0.95	0.05	0.08
5	-0.28_*	0.27_*	0.28_*	-0.07	-0.26_*	-0.34	1.09	0.02	0.09
7	-0.29_*	0.27_*	0.29_*	-0.08	-0.27_*	-0.37	1.05	0.03	0.08
8	0.60_*	-0.53_*	-0.21	-0.94_*	0.41_*	1.32_*	0.51_*	0.30	0.11
9	0.29_*	-0.25_*	-0.10	-0.45_*	0.20	0.63	0.98	0.08	0.11
12	-0.08	0.09	0.08	-0.09	-0.09	0.14	1.20_*	0.00	0.12
15	-0.23_*	0.21	0.22	-0.01	-0.20	-0.35	0.92	0.02	0.04
18	0.05	-0.05	-0.05	0.06	0.05	-0.08	1.21_*	0.00	0.12
26	0.01	-0.02	-0.02	0.03	0.02	-0.04	1.18_*	0.00	0.10
29	0.16	-0.18	-0.02	-0.16	0.15	-0.45	0.72_*	0.04	0.03
30	0.26_*	-0.26_*	-0.16	0.01	0.22_*	0.33	1.12	0.02	0.10
55	-0.38_*	0.38_*	0.17	0.21	-0.31_*	0.59	0.94	0.07	0.09
67	0.01	-0.01	-0.03	0.05	0.02	-0.06	1.29_*	0.00	0.17
72	-0.13	0.14	0.05	-0.20	-0.08	0.56	1.26_*	0.06	0.21_*
75	-0.01	-0.02	0.07	0.23_*	0.00	0.46	1.10	0.04	0.12
78	-0.23_*	0.22	0.24_*	0.12	-0.27_*	-0.55	1.48_*	0.06	0.31_*
80	-0.18	0.17	0.08	0.32_*	-0.18	-0.56	1.21_*	0.06	0.19_*
81	-0.21	0.22	0.32_*	-0.24_*	-0.28_*	-0.49	1.38_*	0.05	0.26_*
82	0.04	-0.04	-0.01	-0.02	0.03	0.07	1.33_*	0.00	0.20_*

De acuerdo a los residuales estudentizados internamente o externamente, las observaciones 1, 8 y 29 son “outliers”.

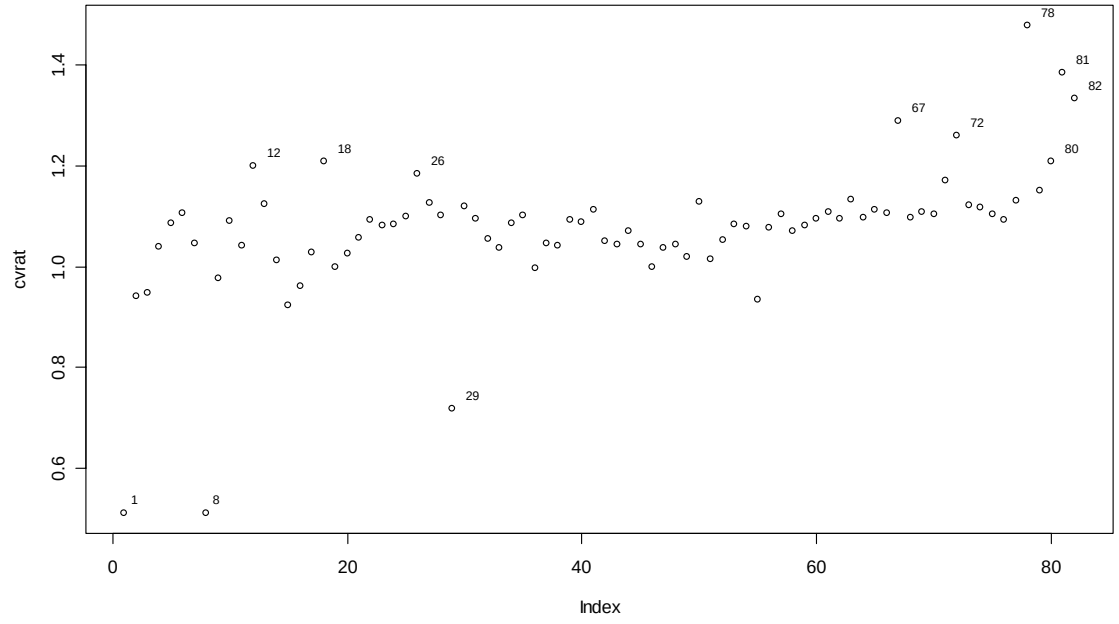
De acuerdo a los valores leverages h_{ii}, las observaciones 72, 78, 80, 81 y 82 son puntos leverages, pues tiene $h_{ii}>3p/n=.1829$

De acuerdo a la Distancia Cook no hay ninguna observación que tenga gran influencia pues todos los CD_i^2 son menores que 1, más aún son menores que F(0.50,5,77)=0.878. Sin embargo, las observaciones 1 y 8 tienen un CD_i^2 mucho mayor que las otras y deberían ser consideradas cuidadosamente.



De acuerdo al DFFITS serían influenciales las observaciones 1, 2, 3, 8, 9, 55, 72 ,78 y 80, pues su $|DFFITS_i| > 2\sqrt{\frac{5}{82}} = .49386$

De acuerdo al COVRATIO serían influenciales las observaciones 1, 8, 12, 18, 26, 29, 72, 78, 80, 81 y 82, pues su COVRATIO>1.1829 ó <0.8171 Los COVRATIO de las observaciones 12, 18, 26 y 29 están bastante cerca de los puntos de corte.



De acuerdo a los DFBETAS una observación es influencial si su valor absoluto es mayor $2/\sqrt{82}$ =0.22086. Las observaciones 1, 8, 9, 80 y 81 parecen afectar el comportamiento del coeficiente β_1, el valor DFBETAS de la observación 75 está muy cerca del punto de corte y no ha sido considerado. Las observaciones 1, 2, 3, 5, 7, 8, 30, 55, 78 y 81 afectan el comportamiento de β_2, en tanto que 1, 2, 3, 5, 7, 8, 9, 30, y 55 parecen tener influencia en β_3 y las observaciones 1, 2, 3, 5, 7, 78 y 81 afectan el comportamiento de β_4.

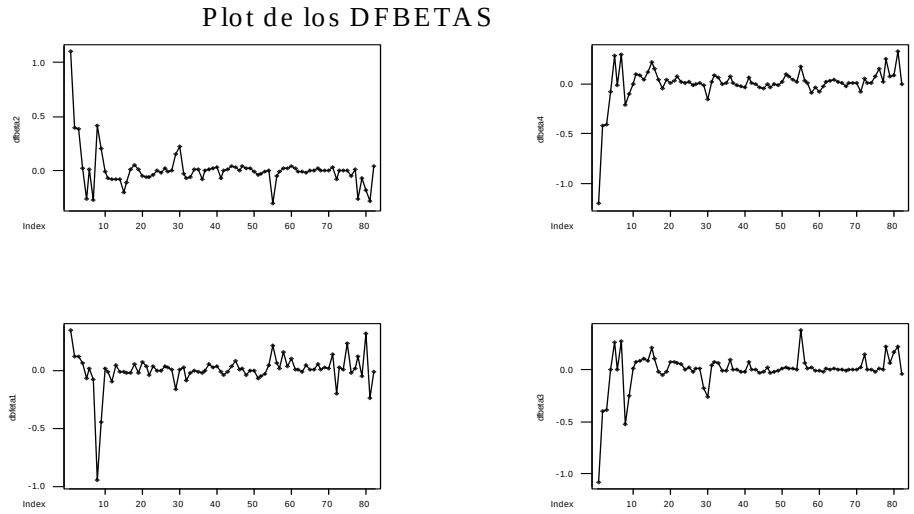


Figura 3.3. Plot de los DFBETAS para el conjunto datos millaje.

En conclusión , las observaciones 1, 2, 3, 5, 7, 8, 9, 30, 55, 72, 78,, 80, 81 y 82 parecen ser las mas influenciales.

3.3 Plot de Residuales para detectar el efecto de variables y casos influenciales

Existen ciertos plot de residuales que se usan para estudiar el efecto de añadir una nueva variable predictora en un modelo. Estos plots tambien permiten detectar la presencia de casos influenciales. Supongamos que queremos ver la importancia de la variable predictora x_j. Consideremos el modelo

$$Y=X_{-j}B_{-j}+\beta_j x_j +e$$

Donde X_{-j} es la matriz X sin incluir la columna j. Se puede mostrar que

$$\hat{\beta}_j = \frac{\mathbf{x}_j'(\mathbf{I}-\mathbf{H}_{-j})\mathbf{Y}}{\mathbf{x}_j'(\mathbf{I}-\mathbf{H}_{-j})\mathbf{x}_j} \qquad (3.15)$$

Definamos los siguientes residuales

a) $\hat{e}_{Y/X_{-j}} = (I-H_{-j})Y$

b) $\hat{e}_{Y/x,X_{-j}} = (I-H)Y$

c) $\widehat{e}_{x_j / X_{-j}} = (I - H_{-j})X_j$

en el caso a) se han considerado en el modelo todas las predictoras excepto x_j, en el caso b) están consideradas todas las variables predictoras y en el caso c) son los residuales de la regresión de x_j versus las otras variables no consideradas en el modelo.

Hay cuatro tipo de plots de residuales que permiten ver el impacto de cada variable predictora x_j en el modelo. Estos son:

a) Plot de Residuales versus las variables predictoras
b) Plot de regresión parcial (o plot de variable añadida)
c) Plot de residuales parciales
d) Plot de residuales parciales aumentados.

a) Plot de residuales versus la variables predictoras.

Aquí se plotea

$\widehat{e}_{Y / x, X_{-j}}$ versus x_j

Si el modelo es adecuado los puntos se deberían alinear a lo largo de una franja horizontal. Si se observa algún patrón no lineal entonces la variable predictora debería ser transformada.

Para el ejemplo de Millaje estos son los plots.de residuales que resultan.

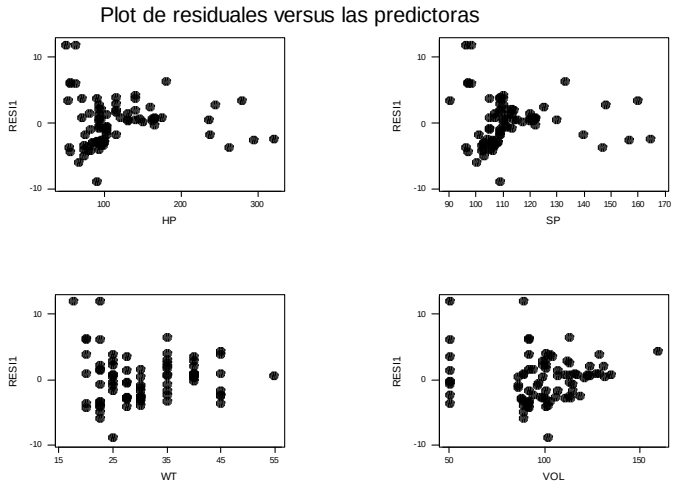


Figura 3.4. Plot de residuales versus las predictoras para el conjunto de datos millaje

Lo que más se destaca en estas gráficas es la presencia de varios valores “outliers” y puntos de leverage alto. También se observa que no todos los puntos se alinean uniformemente alrededor del eje horizontal 0, pero es dificil detectar la tendencia para usar una transformación nolineal.

b) Plots de regresión parciales (o plot de variable añadida)

Aquí se plotea los residuales $\widehat{e}_{Y / X_{-j}}$ versus $\widehat{e}_{x_j / X_{-j}}$

En el plot de regresión parcial se plotea los residuales de la regresión de y considerando todas las variables predictoras excepto x_j versus los residuales de la regresión de x_j contra todas las variables predictoras distintas a ella.
Si la variable x_j entra al modelo en forma lineal entonces su plot de regresión parcial debería mostrar una tendencia lineal que pasa por el origen. Si se observa una tendencia no lineal habría que considerar una transformación de x_j. También se puede localizar a los puntos que afecta el cálculo del coeficiente de regresión correspondiente.

Consideremos por ejemplo el plot de regresión parcial para la variable HP del conjunto de datos Millaje, donde se asume que el modelo contiene ya las otras 3 variables predictoras.

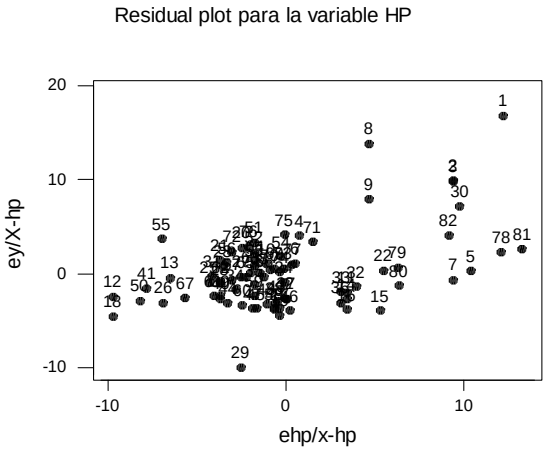


Figura 3.5. Plot de regresión parcial considerando la variable predictora HP

Se observan muchos valores influenciales y la tendencia lineal es bien pobre. Si usamos 1/HP en lugar de HP la cosa no mejora mucho.
En realidad el efecto de este plot se observa mejor si consideramos primero la regresión de MPG con la variable VOL, que es la que tiene menos correlación y si consideramos luego añadir WT. El plot de regresión parcial que se obtiene es

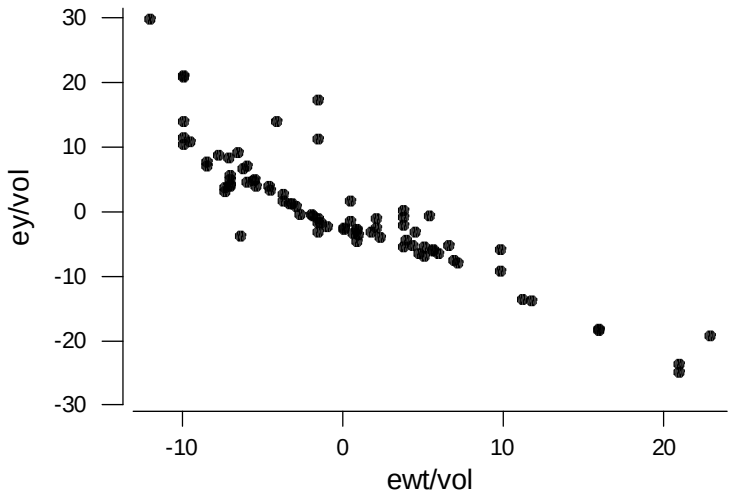


Figura 3.6. Plot de regresión parcial considerando la variable WT asumiendo que el modelo solo contiene a VOL.

Se puede observar que hay bastante linealidad en el plot, y que la linea estimada pasaría por el origen. Luego, se debería usar una regresión lineal múltiple con dos variables predictoras.
Consideremos la misma situación anterior pero en lugar de WT ahora queremos incluir HP. El plot de la Figura 3.7 ya no se ve tan lineal sino parece como una rama de una hipérbola equilatera.

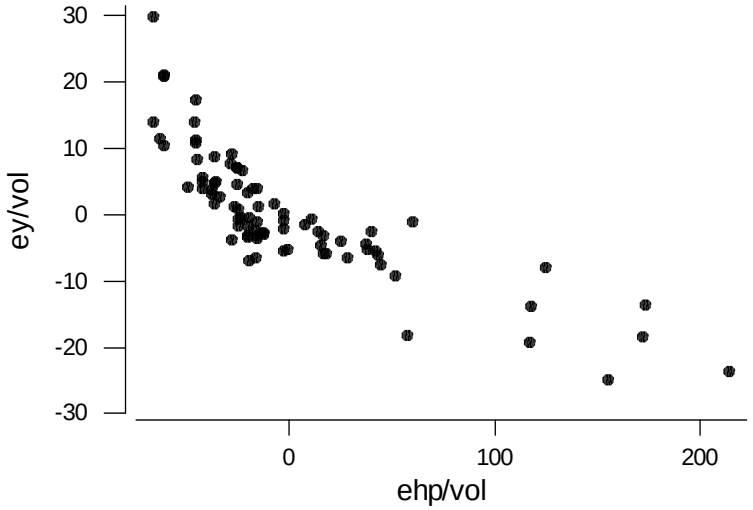


Figura 3.7. Plot de regresión parcial considerando la variable HP asumiendo que el modelo solo contiene a VOL.

Así que sería mejor usar 1/HP en lugar de HP en el modelo

c) Plot de residuales parciales o de residuales más componente

Aquí se plotea
$\hat{e}_{Y/x,X_{-j}} + x_j\beta_j$ versus x_j

Es más efectivo para detectar nolinealidad que el plot de regresión parcial No es muy adecuado para detectar casos influenciales.

d) Plot de residuales parciales aumentados

Aquí se plotea

$$\widehat{e}_{y/X_{-j},x_j^2} + x_j\beta_j + x_j^2\beta_{jj} \text{ versus } x_j$$

Este plot fue propuesto por Mallows (1986) y es el más adecuado para cotejar si la variable x_j debe entrar en forma cuadrática al modelo.

3.4 Plot de residuales para detectar Normalidad

La suposición de la normalidad de los errores es bién importante para el proceso de hacer inferencia en regresión lineal múltiple. Al igual que en regresión lineal simple esto puede ser cotejado haciendo un plot de normalidad para los errores estudentizados internamente.

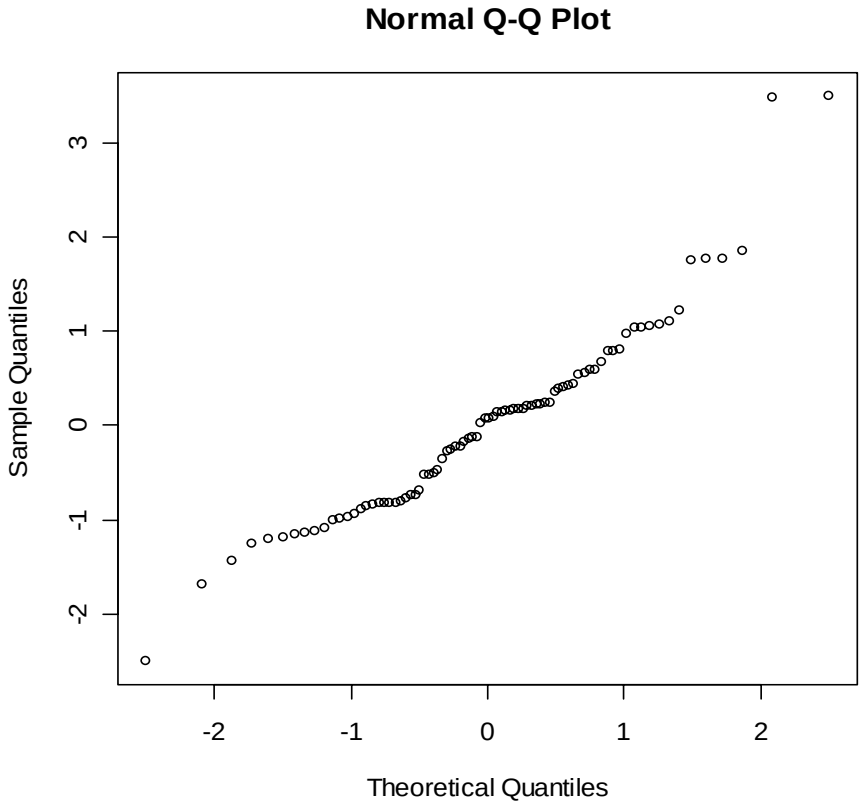


Figura 3.8. Plot de normalidad para los residuales del conjunto de datos millaje.

El plot de Normalidad consiste en un plot de los **scores normales** (estadísticos de orden normales) versus los residuales estandarizados ordenados. Los scores normales representan los valores esperados de observaciones ordenadas que provienen de una distribución normal estándar. El i-ésimo score normal es aproximado en forma bastante precisa por

$$z_{(i)} = \Phi^{-1}(\frac{i-3/8}{n+1/4})$$

donde Φ representa la función de distribución acumulada de una normal estándar y n (n>5) es el número de observaciones en la muestra.
Para que haya normalidad los puntos deben estar alineados alrededor de una recta que pasa por el origen. En la figura 3.8 se muestra el plot de normalidad de los residuales correspondientes a la regresión lineal multiple del conjunto de datos **millaje**, se observa que los puntos están bastante alineados pero se observan varios "outliers" en ambos extremos de la distribución.
Si la tendencia de los puntos es curveada entonces la distribución es asimétrica. El tipo de asimetría es determinada por el lado donde está la parte curveada. Un plot de normalidad que produce una curva en forma de S indica que la distribución tiene una cola pesada o liviana dependiendo de la forma de la S. Si la S es alargada entonces la cola es liviana.

También se podría aplicar una prueba noparamétrica como la de Kolmogorov-Smirnov o Shapiro-Wilks para detectar normalidad.

```
> l1=lm(mpg~.,data=millaje)
> resi=rstandard(l1)
> ks.test(resi,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  resi
D = 0.0881, p-value = 0.519
alternative hypothesis: two-sided

> boxplot(resi)
> shapiro.test(resi)

        Shapiro-Wilk normality test

data:  resi
W = 0.945, p-value = 0.001542
```

El "p-value" de la prueba de Kolmogorov-Smirnov es mayor que 0.05 por lo tanto se acepta la hipótesis de que hay normalidad de los residuales. Sin embargo, la prueba de Shapiro-Wilks indica que no hay normalidad puesto que el "p-value" de la prueba aes pequeño.

3.5 Detectando varianza no constante

La suposición de que en el modelo de regresión lineal múltiple, los errores tienen varianza constante es importante para que los estimadores mínimos cuadráticos sean óptimos. Por lo general varianza no constante viene acompañado del hecho que no hay normailidad.

Para detectar si la varianza es constante o no se hace un plot de resiiduales estudentizados versus los valores ajustados $\hat{y}_i$'s. Si los puntos aparecen alineados arbitrariamente alrededor de una franja horizontal centrada en la linea horizontal en cero entonces hay indicación de varianza constante (homocedasticidad). Si los puntos forman algún tipo de patrón como el que se muestra

en la figura 3.9 entonces indica una violación de la suposición de homocedasticidad. Aquí la varianza varía en forma proporcional a la media de la variable de respuesta Y. Este plot es típico cuando los errores siguen una distribución Poisson o log-normal.

Algunas veces la varianza puede variar de acuerdo a los variables de una variable predictora. Para detectar esta situación hay que hacer un plot de residuales versus cada variable predictora.
Si hay indicación de que la varianza poblacional σ^2 no es constante entonces hay dos remedios posibles:

i) Usar mínimos cuadrados ponderados donde los pesos que se usan son hallados en base a los datos tomados.
ii) Transformar la variable de respuesta Y usando tranfomación que estabiliza la varianza

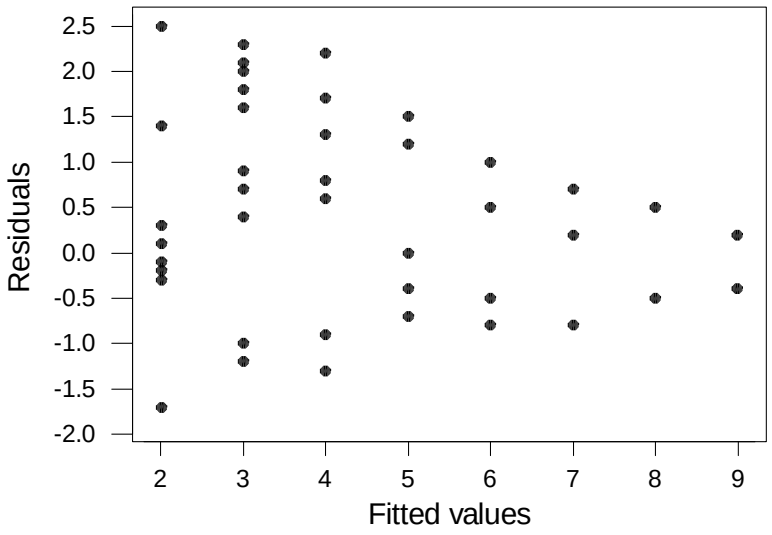


Figura 3.9. Este plot muestra que la varianza de los errores no es constante y que varia en forma proporcional a la media de la variable de respuesta

Las medidas remediales para varianza no constante serán discutidas en el capítulo 4 del texto.

3.6 Errores correlacionados en Regresión

Una de las suposiciones que se hace en regresión lineal es que Cov(e_i, e_j)=E($e_i e_j$)=0 para $i \neq j$ .Es decir que los errores no se correlacionan entre si . Hay un caso en regresión cuando la variable predictora es tiempo, donde puede haber una dependencia del comportamiento con respecto al tiempo . Por ejemplo, las ventas de una compañia de ropas pueden seguir un patrón que depende de la epoca del año y pudiera ocurrir entonces que que E(e_i, e_{i+k}) $\neq 0$ para un cierto k en este caso se dice que los errores tiene una correlación serial y estan autocorrelacionados . Si se gráfica los residuales versus la variable predictora tiempo y se observa mucho cambio de signo

entonces la autocorrelación es negativa si el cambio de signo no es muy frecuente entonces la autocorrelacion es positiva.

Ejemplo 2. Consideremos las siguientes series de tiempo

year	y1	y2	y3
1	10	15	5
2	20	10	10
3	35	18	15
4	50	12	18
5	12	24	20
6	24	15	32
7	40	32	37
8	50	18	39
9	14	40	42
10	25	20	35
11	40	55	30
12	60	25	27

cuyas graficas aparecen en la siguiente figura

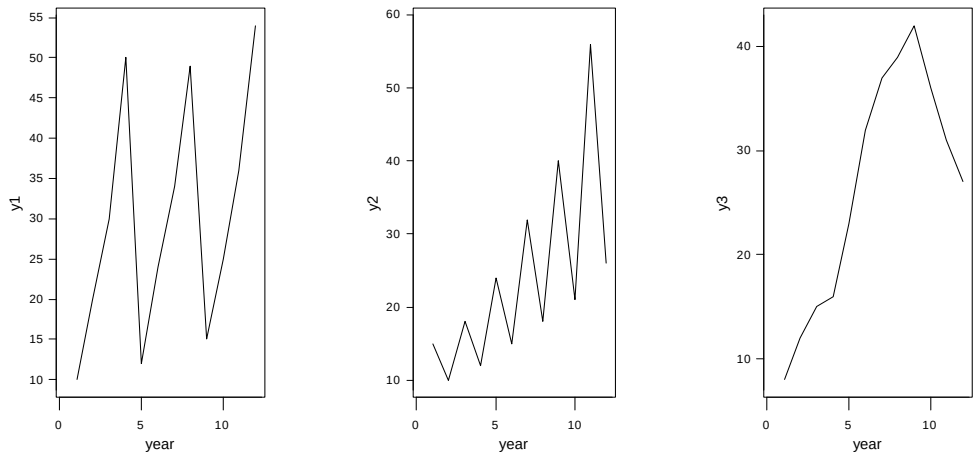


Figura 3.10. Gráfica de las 3 series de tiempo del ejemplo

En los dos primeros plots la autocorrelación es negativa y en la última es positiva
Los plots de residuales correspondientes se muestran en la figura 3.11

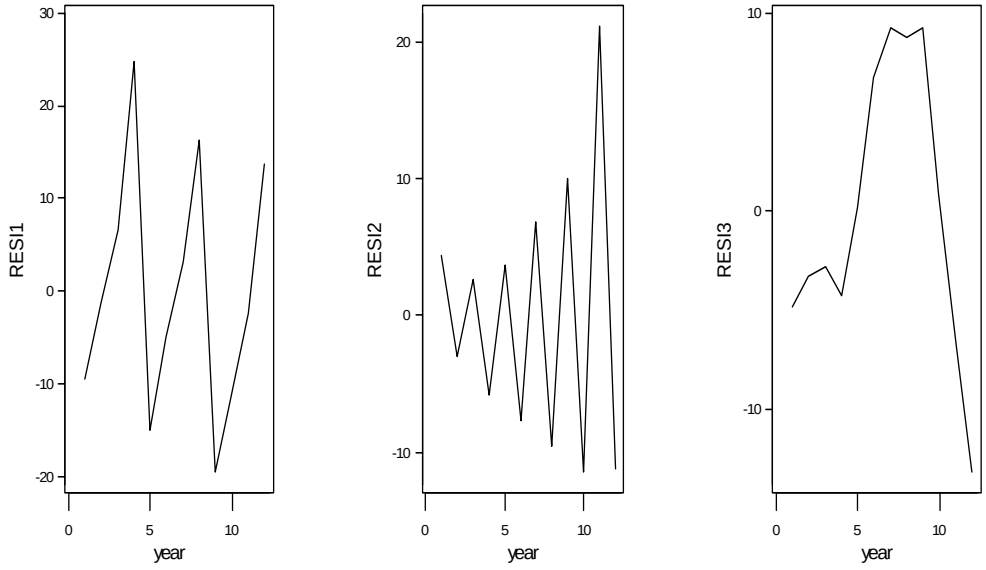


Figura 3.11. Plot de los residuales de las tres series de tiempo del ejemplo 2

Cuando los residuales cambian frecuentemente de signo hay autocorrelación negativa y si hay un conglomerado de residuales de un mismo signo antes de cambiar a otro entonces la autocrrelación es positiva. Lo anterior se puede observar más claramente si se plotea los residuales en el tiempo t versus los residuales en el tiempo t-1. La siguiente figura muestra estos plots para los datos correspondiente a las dos últimas graficas de la figura anterior.

El Modelo autorregresivo de primer orden para los errores se define

$$e_t=\rho e_{t-1}-u_t$$

donde se supone que las u_t son variables aleatorias distrbuidas normalmente con media 0 y varianza constante.

La prueba de Durbin-Watson se usa para detectar si hay correlación positiva. Es decir para probar Ho: $\rho=0$ vs Ha: $\rho>0$.
La prueba está dada por

$$D=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}e_t^2}$$

Existen tablas de la prueba, que dan valores límites DL y DU para que se rechaze la hipótesis nula. La tabla considera número de casos:(n), número de predictoras y tres niveles de significación. Las decisiones se toman así: Se rechaza Ho si D<DL, se acepta Ho si D>DU y la prueba no lleva a ninguna conclusión si DL<D<DU.

La prueba de Durbin-Watson no detecta autocorrelaciones de segundo orden o mayor. R no tiene una función que calcule este estadístico, pero hemos construido una function **dw** para calcularla. Sin embargo, el estadistico es dado por muchos programas entre ellos MINITAB.

Para los datos cuyas graficas de residuales aparecen en la figura 3.11 se obtienen los siguientes resultados para el estadístico de Durbin-Watson.

```
> l1=lm(y1~year,data=corrdata)
> dw(l1$residuals)
[1] 1.835639
> l2=lm(y2~year,data=corrdata)
> dw(l2$residuals)
[1] 3.537003
> l3=lm(y3~year,data=corrdata)
> dw(l3$residuals)
[1] 0.4571476
>
```

Buscando en la tabla de Durbin-Watson con n=12 (aprox con n=15), k=1 y $\alpha=0.5$ resulta ser que $D_L=1.08$ y $D_U=1.36$, por lo tanto, no se rechaza la hipótesis nula en los dos primeros casos ya que DW=1.835 y DW=3.53 respectivamente son mayores que 1.36, y se concluye que no hay autocorrelación de primer orden entre los errores. En el último caso si se rechaza la hipótesis nula puesto que DW=.457 < 1.08 y se concluye que hay autocorrelación positiva de primer orden.

Si se desea probar una hipótesis de dos lados Ho: $\rho=0$, versus Ha:$\rho\neq0$ entonces se rechaza Ho: si $D<D_L$ ó $4-D<D_L$, al nivel de significación de 2α. Si $D>D_U$ y $4-D>D_U$ entonces no se rechaza. Para cualquier otro valor de D la prueba no llega a ninguna conclusión.

Una regla práctica es que cuando el estadístico de Durbin-Watson sale cerca de 2 entonces es probable que no hay autocorrelación.

Si hubiera autocorrelación positiva de primer orden entonces una forma de resolver el problema sería considerar el modelo

$$y_t=\beta_o+\beta_1 t+\beta_2 y_{t-1}+e_t$$

donde y_{t-1} son los valores de la variable de respuesta en el tiempo anterior. Estos modelos son llamados modelos de series de tiempo y son discutidos en textos especializados.

Ejercicios

1. Probar las siguientes identidades

i) $\hat{\boldsymbol{\beta}}_{(\mathbf{i})} = \hat{\boldsymbol{\beta}} - \dfrac{(\mathbf{X'X})^{-1}\mathbf{x_i}\hat{e}_i}{1-h_{ii}}$

ii) $s^2_{(i)} = \dfrac{n-p-1}{n-p-2}s^2 - \dfrac{\hat{e}_i^2}{(n-p-2)(1-h_{ii})}$

2. Probar las siguientes relaciones

i) $t_i = \dfrac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$

ii) $t_i = r_i^*\left(\dfrac{n-p-2}{n-p-1-r_i^{*2}}\right)^{1/2}$

3. Probar la siguiente formula equivalente para calcular COVRATIOS

$$(COVRATIO)_i = \left(\frac{s^2_{(i)}}{s^2}\right)^p \frac{1}{(1-h_{ii})}$$

4. Usar el conjunto de datos **Fuel** con variable de respuesta: Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto.

a) Cotejar las suposiciones del modelo de regresión multiple mediante un plot de residuales.
b) Determinar outliers y puntos con leverage alto usando los diagnosticos de regresión.
c) Usar plot de residuales para evaluar el efecto de añadir la segunda variable predictora con la correlación más alta con Fuel al modelo que ya tiene incluido la variable más altamente correlacionada con Fuel.

5. Usar el conjunto de datos **Headcirc** con variable de respuesta: headcirc (circunferencia de la cabeza del bebe) para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto

a) Cotejar las suposiciones del modelo de regresión multiple mediante un plot de residuales.
b) Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
c) Usar plot de residuales para evaluar el efecto de añadir la segunda variable predictora con la correlación más alta con Headcirc al modelo que ya tiene incluido la variable más altamente correlacionada con Headcir.

6. Usar el conjunto de datos **Grasa** con variable de respuesta: grasa (porcentaje de grasa en el cuerpo) para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto.

a) Cotejar las suposiciones del modelo de regresión multiple mediante un plot de residuales.
b) Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
c)Usar plot de residuales para evaluar el efecto de añadir al modelo con la variable predoctora más altamente correlacionada con grasa, la segunda variable más talamente correlacionada con grasa.

7. Supongamos que ajustamos un modelo de regresion múliple con intercepto y se define la distancia (cuadrada) Mahalanobis de $\mathbf{x_i}=(x_{i,1},\ldots x_{i,p})$ a $\bar{\mathbf{x}} = \sum_{i=1}^{n} \frac{\mathbf{x}_i}{n}$ por

$D_i^2 = (\mathbf{x_i} - \bar{\mathbf{x}})\mathbf{C}^{-1}(\mathbf{x_i} - \bar{\mathbf{x}})$, donde $\mathbf{C^{-1}}$ es la inversa de la matriz de covarianza de las x's.

Establecer una relación entre los valores leverages h_{ii} y D_i^2.

CAPÍTULO 4
TRANSFORMACIONES EN REGRESIÓN

4.1 Transformaciones para linealizar modelos

Consideremos por ahora solo modelos con una variable predictora. La idea es tratar de aumentar la medida de ajuste R^2 del modelo, sin incluir variables predictoras adicionales. Lo primero que hay que hacer es un plot para obervar el tipo de tendencia, pueden resultar plots como los que aparecen en las figuras 4.1 y 4.2.

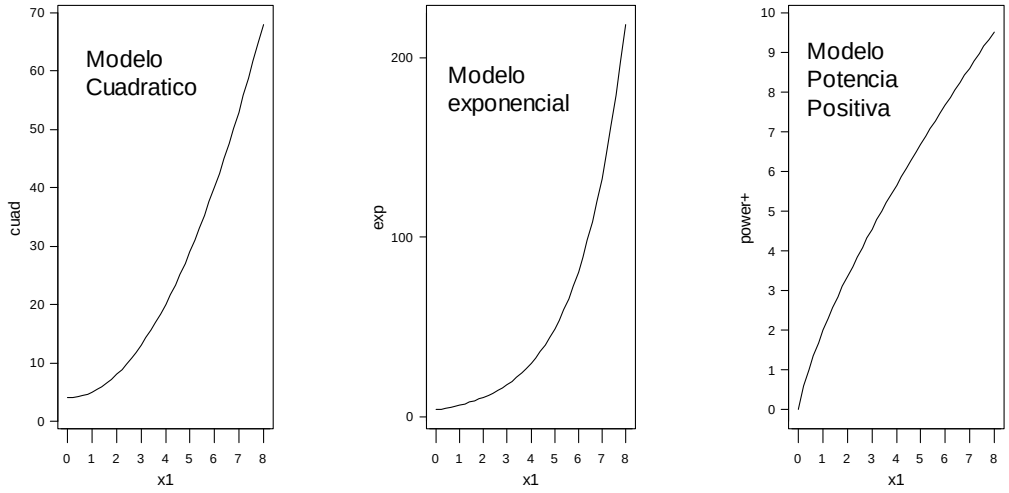


Figura 4.1. Gráficas de tres modelos no lineales.

En la primera gráfica de la figura 4.1 se ha ajustado un modelo cuadrático, que es de la forma general $y=a+bx+cx^2$ y es el caso más sencillo de regresión polinómica. Esto puede ser modelado como una regresión múltiple con dos variables predictoras.
La segunda gráfica corresponde a un modelo exponencial de la forma $y=\alpha e^{\beta x}$ con α y β positivos. Este modelo es muy adecuado para modelar crecimientos poblacionales.
La tercera gráfica corresponde a un modelo potencial o doblemente logarítmico de la forma $y=\alpha x^{\beta}$, con β positivo.

La primera gráfica de la figura 4.2 corresponde a un modelo hiperbólico o inverso de la forma $y=\alpha+\beta/x$, con $x>0$.
La segunda gráfica corresponde a un modelo logarítmico de la forma $y==\alpha+\beta\log(x)$ con $x>0$.
La tercera gráfica corresponde a un modelo potencia pero con $\beta>0$.

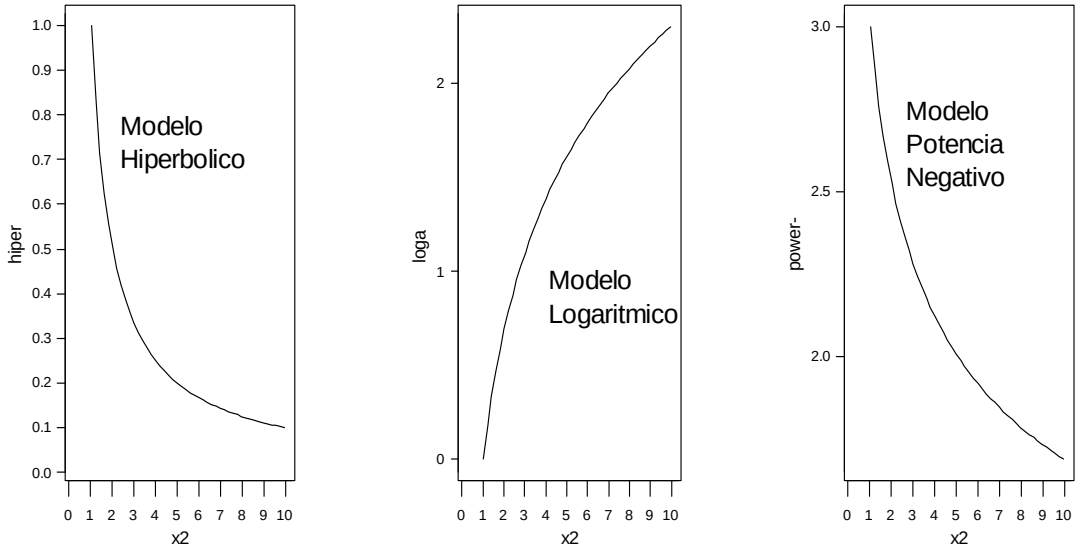


Figura 4.2. Mas gráficas de modelos no lineales

La siguiente tabla muestra las transformaciones de la variable predictora y/o respuesta que se requieren para linealizar varios modelos.

Nombre del modelo	**Ecuación del Modelo**	**Transformación**	**Modelo Linealizado**
Exponencial	$Y=\alpha e^{\beta X}$	Z=LogY X=X	Z=Log α +βX
Logaritmico	Y= α +βLog X	Y=Y W=Log X	Y= α +βW
Doblemente Logarítmico o Potencia	$Y=\alpha X^{\beta}$	Z=Log Y W=Log X	Z= Log α +βW
Hiperbólico	Y= α +β/X	Y=Y W=1/X	Y= α +βW
Doblemente Inverso	Y=1/(α +βX)	Z=1/Y X=X	Z=α +βX

El primer y tercer modelo son válidos bajo la suposición de que los errores son multiplicativos y habría que cotejar haciendo análisis de residuales si el logaritmos de los errores tiene una media de cero y varianza constante. Si los errores no son multiplicativos entonces deberían aplicarse técnicas de regresión no lineal las cuales no son consideradas en este texto.

Ejemplo 1. Los siguientes datos representan como ha cambiado la población en Puerto Rico desde 1930.

```
year  poblacion

1930    1552000
1940    1877800
1950    2218000
```

```
1960    2359800
1970    2716300
1980    3196520
1990    3527796
```

Se desea establecer un modelo para predecir la población de Puerto Rico en el año 2000.

Solución: Observando el diagrama de puntos de población versus años que aparece en la figura de abajo

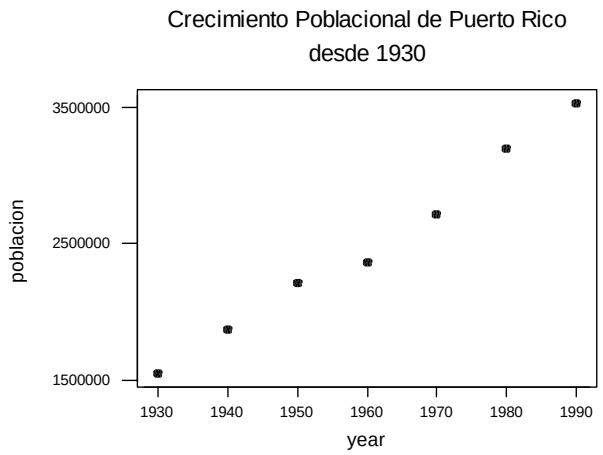


Figura 4.3 Crecimiento poblacional de Puerto Rico desde 1930

El plot sugiere que podemos ajustar los datos al modelo exponencial

$Poblac=\alpha e^{\beta year}$

Y el modelo linealizado da como ecuación

Ln(Poblac) = - 11.4 + 0.0133 year

con un R^2 del 98.9%, mejorando el R^2 del modelo lineal que era de 98.7%. Para predecir la población para el año 2000 se obtiene que

Ln(Poblac)=-11.4 + 0.0133(2000)=-11.4+26.6=15.2

luego $Poblac=e^{15.2}$=3,992,787 será la población de PR estimada para el año 2000.

4.2 Transformaciones de las variables predictoras en regresión múltiple

Supongamos que uno tiene una variable de respuesta Y y varias variables predictoras y desea hacer transformaciones en las variables de respuesta para mejorar la medida de ajuste del modelo. Lo primero que uno intenta es hacer un plot matricial y de allí extraer las relaciones de y con cada una de las variables predictoras. Pero estas transformaciones se pueden ver afectadas por la colinealidad (dependencia lineal) existente entre las variables predictoras. Este mismo problema afecta al plot de regresión parcial o de variables añadidas.

En 1962, Box y Tidwell, propusieron un método para transformar las variables predictoras pero solo usando potencia de ellas. Mas especificamente, ellos consideraron el modelo

$$y = \beta_o + \beta_1 w_1 + ... + \beta_k w_k + e \qquad (4.1)$$

donde $w_j = x_j^{\alpha_j}$ si $\alpha_j \neq 0$ y $w_j=\ln(x_j)$ si $\alpha_j = 0$. El método está basado en el desarrollo en series de Taylor del modelo anterior con respecto a $\boldsymbol{\alpha} = (\alpha_1,.....\alpha_k)$ y alrededor del punto $\boldsymbol{\alpha_0} = (\alpha_{1,0},....., \alpha_{k,0}) = (1,....,1)$. Haciendo las derivaciones respectivas, el modelo (4.1) se reduce a:

$$y \cong \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + (\alpha_1 - 1)\beta_1 x_1 \ln x_1 + (\alpha_2 - 1)\beta_2 x_2 \ln x_2 + ... + (\alpha_k - 1)\beta_k x_k \ln x_k$$

el cual es equivalente a

$$y \cong \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \gamma_1 z_1 + \gamma_2 z_2 + ... + \gamma_k z_k \qquad (4.2)$$

donde $\gamma_j = (\alpha_j - 1)\beta_j$ y $z_j = x_j \ln x_j$ para j=1,2….,k.

El procedimiento para la estimación de los α_j se puede resumir como sigue:

a) Hacer la regresión lineal múltiple considerando las variables predictoras originales x_j y denotar los estimados de los coeficientes por b_j.
b) Hacer una regresión lineal múltiple de y versus las variables predictoras originales mas las variables $z_j=x_j\ln(x_j)$ y denotar los estimados de los coeficientes de z_j por $\hat{\gamma}_j$.
c) Estimar α_j por $\hat{\alpha}_j = \dfrac{\hat{\gamma}_j}{b_j} + 1$

El procedimiento se puede repetir varias veces usando en cada etapa las nuevas variables transformadas y la siguiente relación de recurrencia

$$\hat{\alpha}_j^{(m+1)} = (\frac{\hat{\gamma}_j^{(m)}}{b_j^{(m)}} + 1)\hat{\alpha}_j^{(m)} \qquad (4.3)$$

Terminando el proceso cuando $|\alpha_j^{(m+1)} - \alpha_j^{(m)}| < TOL$, donde TOL es una cantidad de tolerancia muy cercana a cero.

Sin embargo, muy a menudo un solo paso es suficiente.

Ejemplo 2. Aplicar la técnica sugerida por Box and Tidwell al conjunto de datos **millaje.**

Solución. Usando R se obtiene

```
> l1<-lm(mpg~.,data=millaje)
> betas<-l1$coeff
> betas
 (Intercept)          sp          wt         vol          hp
192.43775332  -1.29481848  -1.85980373  -0.01564501   0.39221231
> summary(l1)

Call:
lm(formula = mpg ~ ., data = millaje)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
sp          -1.29482    0.24477  -5.290 1.11e-06 ***
wt          -1.85980    0.21336  -8.717 4.22e-13 ***
vol         -0.01565    0.02283  -0.685    0.495
hp           0.39221    0.08141   4.818 7.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733,     Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

Notar que la predictora VOL no es significativa.

La regresión con las variables originales resulta ser

```
MPG = 192.4 - 0.0156 VOL + 0.392 HP - 1.294 SP - 1.859 WT
```

Ahora creamos cuatro variables predictoras $z_1=x_1 lnx_1$, $z_2=x_2 lnx_2$, $z_3=x_3 lnx_3$ y $z_4=x_4 lnx_4$. Haciendo la regresión múltiple con las 8 variables predictoras se obtiene

```
> l2<-lm(mpg~.,data=millaje1)
> betas2<-l2$coeff
> betas2
 (Intercept)          sp          wt         vol          hp          z1
1048.2022263  -38.8522423  -17.9023484   -1.0023285    5.4675149    6.3624693
         z2          z3          z4
  3.3262799    0.1803016   -0.8006012
> summary(l2)
```

```
Call:
lm(formula = mpg ~ ., data = millaje1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0797 -1.4479 -0.1852  1.4320 10.1958

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1048.2022   268.3693   3.906 0.000208 ***
sp           -38.8522    11.8106  -3.290 0.001546 **
wt           -17.9023     4.3238  -4.140  9.2e-05 ***
vol           -1.0023     0.5916  -1.694 0.094470 .
hp             5.4675     1.8491   2.957 0.004185 **
z1             6.3625     1.9713   3.228 0.001871 **
z2             3.3263     0.8739   3.806 0.000291 ***
z3             0.1803     0.1086   1.660 0.101185
z4            -0.8006     0.2744  -2.917 0.004690 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.247 on 73 degrees of freedom
Multiple R-Squared: 0.905,      Adjusted R-squared: 0.8946
F-statistic: 86.97 on 8 and 73 DF,  p-value: < 2.2e-16
```

La ecuación de regression estimada resulta ser

$$MPG = 1048.2 - 38.852\,SP - 17.902\,WT - 1.002\,VOL + 5.467\,HP + 6.362\,x_1 lnx_1 + 3.326\,x_2 lnx_2 + 0.180\,x_3 lnx_3 - 0.800\,x_4 lnx_4$$

Notar que tanto VOL como la variable z_3 , relacionalda a ella, son no significativas.

Aplicando el paso c) del algoritmo se tendría que

```
> gammas<-betas2[c(6:9)]
> #Hallando los alfas
> alfas<-(gammas/betas1)+1
> alfas
       z1         z2         z3         z4
 -3.9137925  -0.7885113 -10.5245410  -1.0412443
```

Haciendo la regresión con las nuevas variables $vol^{-10.52}$, $hp^{-1.04}$, $sp^{-3.91}$ y $wt^{-0.79}$ se obtiene

```
> sp1<-millaje1$sp^alfas[1]
> wt1<-millaje1$wt^alfas[2]
> vol1<-millaje1$vol^alfas[3]
> hp1<-millaje1$hp^alfas[4]
> #regresion con todas las variables transformadas
> l3<-lm(millaje1$mpg~sp1+wt1+vol1+hp1)
> summary(l3)
```

```
Call:
lm(formula = millaje1$mpg ~ sp1 + wt1 + vol1 + hp1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.34348 -1.62938 -0.07744  1.35872 10.15980

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.298e+00  4.420e+00  -0.520 0.604656
sp1         -1.465e+08  4.698e+08  -0.312 0.755972
wt1          3.329e+02  9.382e+01   3.548 0.000665 ***
vol1         1.843e+18  8.827e+17   2.088 0.040082 *
hp1          1.668e+03  8.078e+02   2.065 0.042325 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.095 on 77 degrees of freedom
Multiple R-Squared: 0.909,      Adjusted R-squared: 0.9043
F-statistic: 192.4 on 4 and 77 DF,  p-value: < 2.2e-16
```

Hay problemas con la variable transformada de VOL, su coeficiente estimado es enormemente grande, porque todas sus entradas se hacen demasiado pequeñas.

Repitiendo el proceso, eliminado VOL antes de aplicar el método de Box and Tidwell se obtiene que

```
> millaje2<-cbind(millaje2,z11,z21,z31)
> l21<-lm(mpg~.,data=millaje2)
> betas22<-l21$coeff
> gammas1<-betas22[c(5:7)]
> #Hallando los alfas1
> alfas1<-(gammas1/betas12)+1
> alfas1
      z11       z21       z31
-4.3033518 -0.9219605 -1.0966294
```

Luego, α_1=-1.09, α_2 =-4.30 y α_3=-0.92

```
> #Creando las nuevas variables
> sp11<-millaje2$sp^alfas1[1]
> wt11<-millaje2$wt^alfas1[2]
> hp11<-millaje2$hp^alfas1[3]
> l5<-lm(millaje2$mpg~sp11+wt11+hp11)
> summary(l5)

Call:
lm(formula = millaje2$mpg ~ sp11 + wt11 + hp11)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.60068 -1.61086  0.08952  1.18229 12.43902

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.084e+00  3.681e+00   0.566  0.57286
sp11        -1.507e+09  2.800e+09  -0.538  0.59186
wt11         4.503e+02  1.345e+02   3.348  0.00125 **
hp11         2.146e+03  1.052e+03   2.040  0.04475 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.167 on 78 degrees of freedom
Multiple R-Squared: 0.9035,     Adjusted R-squared: 0.8998
F-statistic: 243.4 on 3 and 78 DF,  p-value: < 2.2e-16
```

Luego la regresión estimada es

```
MPG = 2.084 + 2146 hp11 + 450.3 wt11 -1.507e+09 sp11
```

Observese que la predictora SP11 no es significativa y podríamos sacarla del modelo. El cual se reduciría ahora a

```
> #Haciendo la regresion con solo las dos variables significativas
> l6<-lm(millaje2$mpg~wt11+hp11)
> summary(l6)

Call:
lm(formula = millaje2$mpg ~ wt11 + hp11)

Residuals:
    Min      1Q  Median      3Q     Max
-8.67047 -1.66461  0.04419  1.21415 12.53739

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.3744     1.8522   0.202     0.84
wt11         511.1816    72.4007   7.060 5.73e-10 ***
hp11        1600.0983   280.2234   5.710 1.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.153 on 79 degrees of freedom
Multiple R-Squared: 0.9031,     Adjusted R-squared: 0.9007
F-statistic: 368.3 on 2 and 79 DF,  p-value: < 2.2e-16
```

```
La ecuación de regression estimada es:
MPG = 0.374 + 1600.0 hp1 + 511.1 wt1
```

Donde $hp1=1/hp^{1.09}$ y $wt1=1/wt^{0.92}$. En la sección 2.3.5 habíamos llegado a establecer que el mejor modelo era de MPG versus hpo=1/hp y wto=1/wt. Los resultados eran como sigue:

```
> reg1=lm(mpg~hp0+wt0)
> summary(reg1)

Call:
lm(formula = mpg ~ hp0 + wt0)

Residuals:
    Min      1Q  Median      3Q     Max
-8.70343 -1.70579 -0.01131  1.20593 12.60609

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.217      1.573   0.773    0.442
hp0         1131.489    200.586   5.641 2.54e-07 ***
wt0          610.370     87.913   6.943 9.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.165 on 79 degrees of freedom
Multiple R-Squared: 0.9024,     Adjusted R-squared: 0.8999
F-statistic: 365.3 on 2 and 79 DF,  p-value: < 2.2e-16
```

Notar pues que la transformación de Box y Tidwell parece ser bastante eficiente.

4.3. Transformaciones para mejorar la normalidad de la variable de respuesta

En 1964, Box y Cox introdujeron una transformación de la variable de respuesta con el objetivo de satisfacer la suposición de normalidad del modelo de regresión. La transformación es de la forma y^{λ} (transformación potencia), donde λ es estimada con los datos tomados. Más especificamente, la transformación está definida por $w=\dfrac{y^{\lambda}-1}{\lambda}$ si $\lambda\neq 0$ y $w=\ln(y)$ si $\lambda=0$. Notar que $\lim_{\lambda \to 0} \dfrac{y^{\lambda}-1}{\lambda} = \ln y$. En la figura 4.4 se muestra la gráfica de la transformación Box-Cox para 5 distintos valores de lambda.

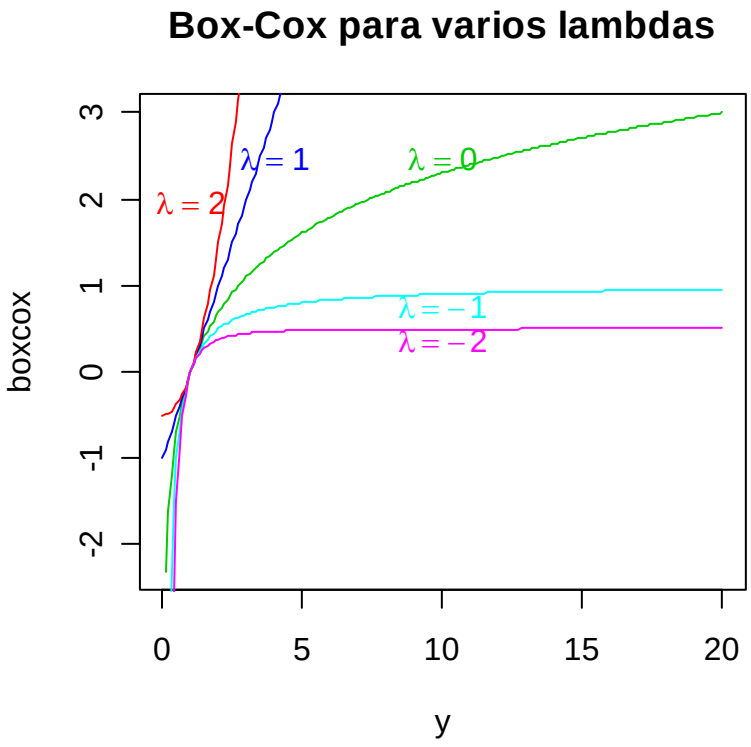


Figure 4.4 Transformación Box-Cox para varios valores de lambda

El parámetro λ se estima, usando el método de Máxima verosimilitud, conjuntamente con los coeficientes del modelo de regresión lineal múltple

$$w = \beta_o + \beta_1 x_1 + ... + \beta_k x_k + e \qquad (4.4)$$

La transformación estandarizada de los w's se define por

$$z_i = \frac{w_i}{\tilde{y}^{\lambda-1}} \qquad (4.5)$$

donde $\tilde{y} = (\prod_{i=1}^{n} y_i)^{1/n}$, es la media geométrica de las y's . Luego, el modelo (4.4) se convierte en **z=Xβ+*e* .** El método asume que para algún λ las z_i's son normales e independientes con varianza común σ^2 .
Escribiendo la función de verosimilitud, correspondiente al modelo transformado, en términos de las z_i's se tiene que

$$L(\boldsymbol{\beta},\lambda) = \frac{e^{-\frac{1}{2\sigma^2}\mathbf{e'e}}}{(2\pi\sigma^2)^{n/2}} = \frac{e^{-\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{X\beta})'(\mathbf{z}-\mathbf{X\beta})}}{(2\pi\sigma^2)^{n/2}}$$

Luego se puede establecer que el máximo del logaritmo de la función de verosimílitud está dado por:

$$LnL(\hat{\boldsymbol{\beta}},\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (4.6)$$

donde $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'z}$, y $\hat{\sigma}^2 = SSE/n = (\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})/n$. Luego,

$$LnL(\hat{\boldsymbol{\beta}},\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n}{2} \equiv -\frac{n}{2}\ln(\hat{\sigma}^2) \qquad (4.7)$$

Claramente (4.7) depende de λ puesto que $\hat{\sigma}^2$ depende de **z** y ésta a su vez de λ.

El procedimiento para estimar el parámetro λ es el siguiente:

1) Seleccionar una conjunto de valores de λ entre –2 y 2, usualmente entre 10 y 20 valores
2) Para cada valor de λ, ajustar el modelo

$$\mathbf{z}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}$$

3) Plotear max[Ln L($\boldsymbol{\beta}$,λ)] versus λ.
4) Escoger como parámetro λ aquel valor que da el mayor valor para max[Ln L($\boldsymbol{\beta}$,λ)].

Varios programas estadísticos, entre ellos S-Plus y R, tienen funciones que permiten estimar el parámetro λ de la transformación Box-Cox. Además del plot del paso 3 producen un intervalo de confianza para λ.

Ejemplo 3. Aplicar la transformación de Box y Cox al conjunto de datos **millaje**

Solución: Haremos uso de R, cuya librería MASS incluye la función **boxcox**. Los resultados usando las variables originales son como sigue:

```
> reg1<-lm(MPG~VOL+HP+SP+WT,data=MILLAJE)
> summary(reg1)

Call: lm(formula = MPG ~ VOL + HP + SP + WT, data = MILLAJE)
Residuals:
   Min     1Q Median    3Q   Max
-9.011 -2.773 0.2733 1.836 11.99

Coefficients:
             Value Std. Error   t value  Pr(>|t|)
(Intercept) 192.4378   23.5316    8.1778    0.0000
        VOL  -0.0156    0.0228   -0.6854    0.4951
         HP   0.3922    0.0814    4.8176    0.0000
         SP  -1.2948    0.2448   -5.2899    0.0000
```

```
WT   -1.8598    0.2134    -8.7166    0.0000

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733
F-statistic: 132.7 on 4 and 77 degrees of freedom, the p-value is 0

Correlation of Coefficients:
   (Intercept)     VOL      HP      SP
VOL  0.1049
 HP  0.9814      0.2324
 SP -0.9961     -0.1501 -0.9837
 WT -0.8658     -0.4260 -0.9228  0.8555
```

Aplicando la función boxcox

```
> boxcox(reg1,lambda=seq(-.6,.6,length=20),plotit=T)
```

Se obtiene el plot de la siguiente figura, donde el parámetro λ puede ser estimado por –0.22, con un intervalo de confianza de (-0.54,0.10)

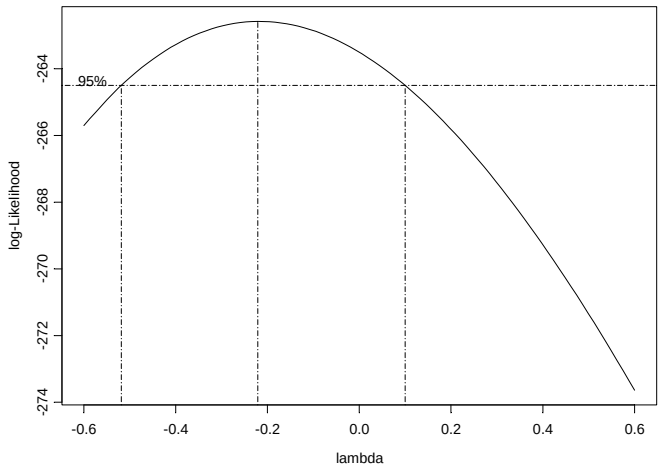


Figura 4.5. Plot de log-likelihood para varios valores de λ

Ahora veremos el efecto de la transformación

```
> millaje1=millaje
> millaje1$mpg<-((millaje$mpg)^-0.22-1)/-0.22
> reg2<-lm(mpg~vol+hp+sp+wt,data=millaje1)
> summary(reg2)

Call: lm(formula = mpg ~ vol + hp + sp + wt, data = millaje1)
Residuals:
    Min      1Q    Median      3Q    Max
 -0.128 -0.0237 -0.004189 0.01595 0.1096
Coefficients:
```

```
             Value Std. Error  t value Pr(>|t|)
(Intercept)  3.2290   0.2600   12.4214  0.0000
      vol   -0.0001   0.0003   -0.4866  0.6279
       hp    0.0004   0.0009    0.4573  0.6488
       sp   -0.0033   0.0027   -1.2145  0.2283
       wt   -0.0152   0.0024   -6.4641  0.0000

Residual standard error: 0.04035 on 77 degrees of freedom
Multiple R-Squared: 0.9252
F-statistic: 238.2 on 4 and 77 degrees of freedom, the p-value is 0
Correlation of Coefficients:
   (Intercept)     vol        hp         sp
vol  0.1049
hp   0.9814      0.2324
sp  -0.9961     -0.1501 -0.9837
wt  -0.8658     -0.4260 -0.9228  0.8555
```

Los plots para cotejar normalidad de los residuales se muestra en la figura 4.5

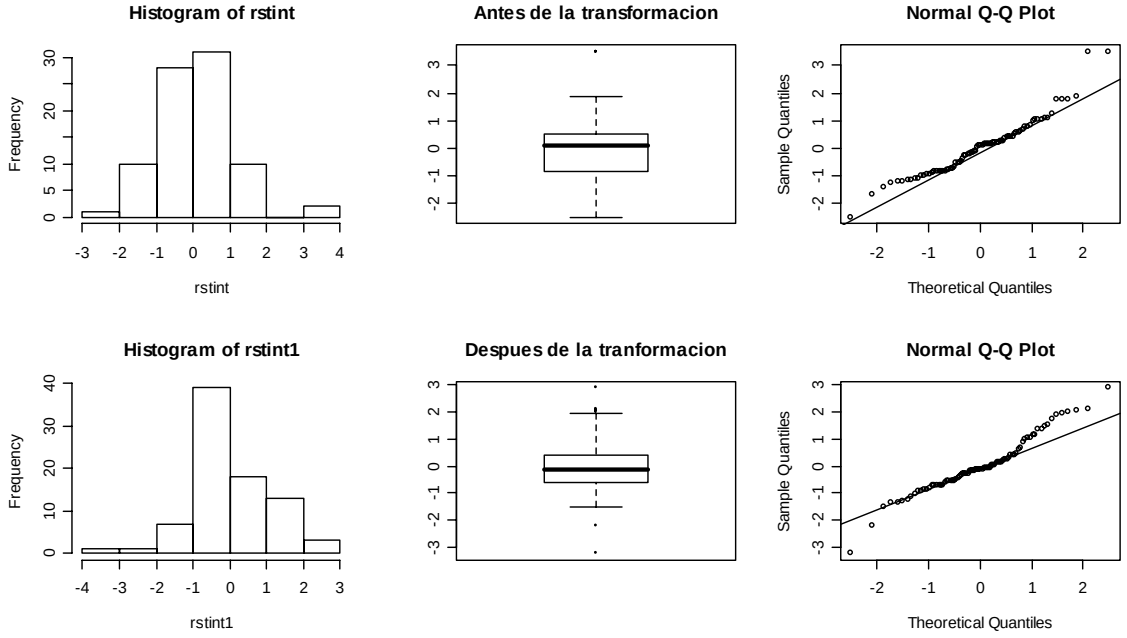


Figura 4.6. Plots para ver el efecto de la transformación Box-Cox en la distribución de los residuales de la regresión para el conjunto de datos millaje.

Notar que los puntos están mejor alineados que en plot con las variables originales (ver figura 3.8) especialmente en la parte central. Se observan claramente dos "outliers" inferiores y uno superior. Notar que el R^2 ha subido de 87.33% a 92.52%, mejorando el efecto de transformar las variables predictoras que se llevó a cabo en el ejemplo 2.

4.4 Transformaciones para estabilizar la varianza.

Algunas veces el comportamiento de la varianza varía según la variable de respuesta. Una de las medidas remediales para hacer constante la varianza es transformar la variable de respuesta. La siguiente tabla muestra las transformaciones de la variable de respuesta que hay que hacer para hacer que la varianza sea constante

Situación	**Transformación**
$Var(e_i) \propto E(y_i)$	$\sqrt{y}$
Igual que el caso anterior	$\sqrt{y}+\sqrt{y+1}$
$Var(e_i) \propto (E(y_i))^2$	Log(Y)
Igual que el caso anterior	Log(y+1)
$Var(e_i) \propto (E(y_i))^4$	1/y
Igual que el caso anterior	1/(y+1)
$Var(e_i) \propto E(y_i)[1-E(y_i)]$	$Sin^{-1}(\sqrt{y})$

Las transformaciones se justifican de la siguiente manera:

Expandiendo en series de Taylor una función h(Y) alredededor de $\mu=E(Y)$ se obtiene

$$h(Y) \approx h(\mu) + h'(\mu)(Y-\mu) + h''(\mu)(Y-\mu)^2/2 \qquad (4.8)$$

Tomando varianza a ambos lados y considerando solamente la aproximación lineal se obtiene

$$Var(h(Y)) \approx [h'(E(y)]^2 Var(Y) \qquad (4.9)$$

Por ejemplo, si $Var(Y) \propto [E(y)]^2$ se tendrá que $[h'(E(Y))]^2 \approx$constante$/[E(y)]^2$. Luego, $h'(\mu) \approx 1/\mu$, de donde por integración resulta $h(\mu) \approx \log(\mu)$.

Haciendo un plot de residuales versus los valores ajustados de Y se puede estimar la transformación más adecuada. Si hubiera valores repetidos de Y entonces es mejor agrupar la variable y calcular medias y desviaciones estándar para cada grupo y luego estimar la mejor linea que pasa por los puntos ($\log \overline{Y}_g$, $\log S_g^2$).

Ejemplo 4. Aplicar una transformación para establizar la varianza en el modelo de regresión para el conjunto de datos **millaje**

Solución. Si observamos el plot de residuales versus valores ajustados por el modelo de regresión, el cual aparece en la figura 4.6 podemos ver que la varianza está cambiando de alguna manera con los valores $\hat{y}$ Se ha explorado varias transformaciones del tipo potencia para la variable de respuesta y la que ha dado mejores resultados es la transformación $h(y)=y^{-1/2}$ que es

aquella correspondiente a la situación cuando la varianza de los errores es proporcional al cubo de la media de la variable de respuesta.

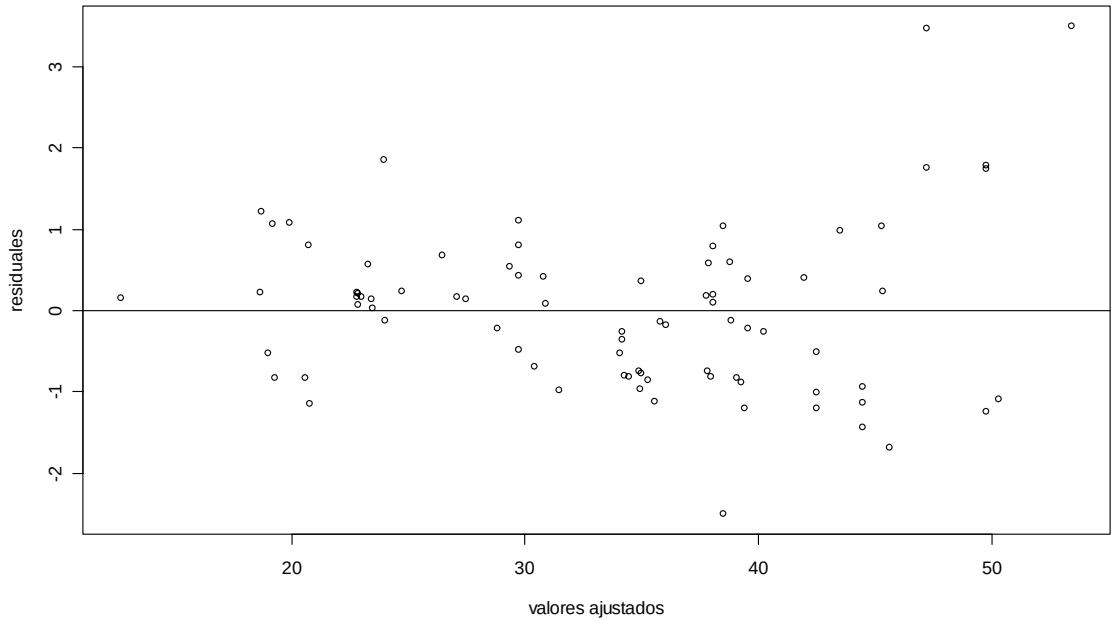


Figura 4.7. Plot de residuales estandarizados versus valores ajustados para el conjunto de datos **millaje**

```
> # El lsfit indica que la varianza es proporcional a la media al cuadrado
> # una transformacion logaritmica en la variable de respuesta es recomendada
> mpglog<-log(millaje$mpg)
> millaje1<-cbind(millaje,mpglog)

> l2<-lm(mpglog~sp+wt+vol+hp,data=millaje1)
> summary(l2)

Call:
lm(formula = mpglog ~ sp + wt + vol + hp, data = millaje1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.273816 -0.058032 -0.008837  0.038624  0.253079

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.7725247  0.5647743  10.221 5.49e-16 ***
sp          -0.0130542  0.0058747  -2.222   0.0292 *
wt          -0.0370088  0.0051209  -7.227 3.08e-10 ***
vol         -0.0003088  0.0005478  -0.564   0.5746
hp           0.0029479  0.0019540   1.509   0.1355
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08767 on 77 degrees of freedom
Multiple R-Squared: 0.9211,     Adjusted R-squared: 0.917
F-statistic: 224.8 on 4 and 77 DF,  p-value: < 2.2e-16
```

```
# Considerando que la varianza es proporcional a la media al cubo
# una transformacion  h(y)=y^-0.5 es realizada
mpg05<-millaje$mpg^-0.5
millaje2<-cbind(millaje,mpg05)
l3<-lm(mpg05~sp+wt+vol+hp,data=millaje2)
summary(l3)
```

```
> summary(l3)

Call:
lm(formula = mpg05 ~ sp + wt + vol + hp, data = millaje2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.019083 -0.003005  0.001039  0.003944  0.024431

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.141e-02  5.014e-02   1.823   0.0722 .
sp          -7.386e-05  5.215e-04  -0.142   0.8878
wt           2.398e-03  4.546e-04   5.275 1.18e-06 ***
vol          1.751e-05  4.863e-05   0.360   0.7198
hp           1.621e-04  1.735e-04   0.935   0.3529
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007783 on 77 degrees of freedom
Multiple R-Squared: 0.9266,     Adjusted R-squared: 0.9228
F-statistic: 243.1 on 4 and 77 DF,  p-value: < 2.2e-16
```

El plot de residuales versus valores ajustados es como en la Figura 4.7. Notar que no se observa ningún patron de los puntos y hay dos “outliers” bien distinguibles.

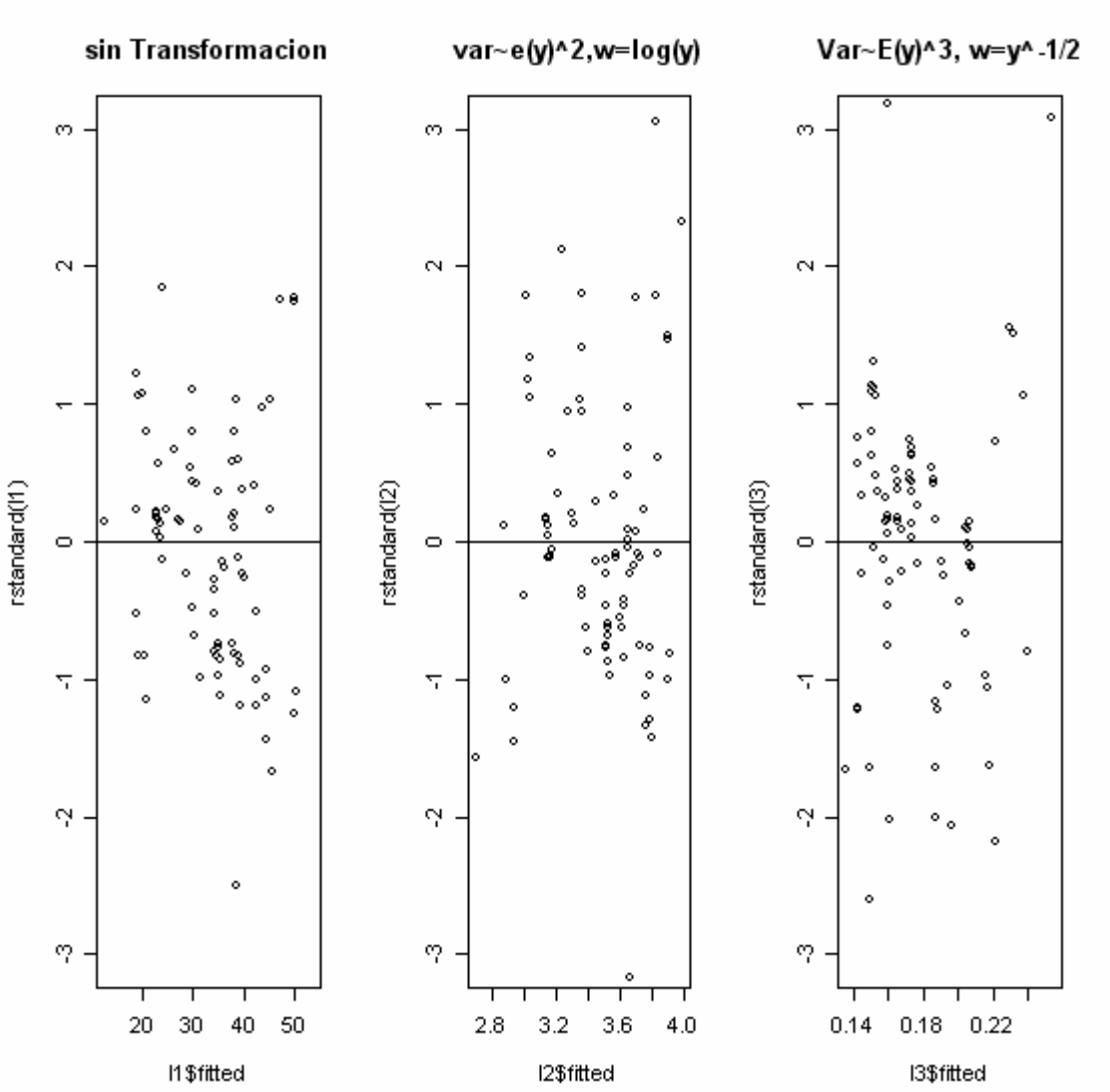


Figura 4.8. Plot de residuales versus valores ajustados después de la transformación.

4.5 Mínimos cuadrados ponderados.

Otra manera de tratar de remediar la falta de homogeneidad de varianza de los errores es usar mínimos cuadrados ponderados, suponiendo que los errores son todavía no correlacionados. En este caso se minimiza $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$, donde w_i representa el peso asignado a la i-ésima observación. Por ejemplo, si en el plot de residuales versus la variable predictora se observa que la dispersión aumenta cuando x aumenta sería conveniente usar $w_i = \frac{1}{\sigma_i^2}$. Aquí, σ_i^2 son las varianzas poblacionales de la Y para cada observación x_i en caso de regresión lineal simple o para

cada combinación de las variables predictoras en el caso de regresión lineal múltiple. Obviamente estas varianzas no son conocidas y deben ser estimadas por sus varianzas muestrales s_i^2. Si hay solamente una observación *y* para el valor x_i entonces se consideran valores de y corrrespondientes a valores cercanos a x_i. En otras palabras la variable x es considerada agrupada.
Esta no es la única manera de escoger los pesos, en regresión robusta que será tratada en el capítulo 8, se hacen distintos cálculos de los pesos con la idea de dar a las observaciones anómalas un menor peso. El cálculo de los pesos está basado mayormente en los diagnósticos de regresión.

Consideremos el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e} \qquad (4.10)$$

con $\mathbf{Var(e)=V}\sigma^2$, donde **V** es una matriz diagonal. Es decir,

$$\mathbf{V} = \begin{bmatrix} k_1^2 & 0 & . & . & 0 \\ 0 & k_2^2 & . & . & 0 \\ 0 & 0 & k_3^2 & . & 0 \\ . & . & . & . & 0 \\ 0 & 0 & 0 & 0 & k_n^2 \end{bmatrix}$$

Sea $\mathbf{W}=(\mathbf{V}^{1/2})^{-1}$, claramente, $\mathbf{W'W=V^{-1}}$. Multiplicando ambos lados del modelo lineal (4.10) por W se obtiene

$$\mathbf{Wy} = \mathbf{WX\beta} + \mathbf{We} \qquad (4.11)$$

Sea $\mathbf{y^*=Wy}$, $\mathbf{e^*=We}$ y $\mathbf{X^*=WX}$, entonces el modelo (4.11) se convierte en el modelo

$$\mathbf{y^*=X^*\beta} + \mathbf{e^*} \qquad (4.12)$$

Notar que $\text{Var}(\mathbf{e^*})=\text{Var}(\mathbf{We})=\mathbf{W}\text{var}(\mathbf{e})\mathbf{W'}=\mathbf{WVW'}\sigma^2=\mathbf{I}\sigma^2$, así que la varianza de los errores es constante. Luego el estimador mínimo cuadrático de $\boldsymbol{\beta}$ será

$$\boldsymbol{\beta}^* = (\mathbf{X^{*\prime}X})^{-1}\mathbf{X^{*\prime}Y^*} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}Y}.$$

Se puede ver fácilmente que $\mathbf{E}(\boldsymbol{\beta}^*)=\boldsymbol{\beta}$ y que

$$\mathbf{Var}(\boldsymbol{\beta}^*)=(\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}}Var(\mathbf{Y})\mathbf{V^{-1}X}(\mathbf{X'V^{-1}X})^{-1} = (\mathbf{X'V^{-1}X})^{-1}\sigma^2$$

Ejemplo 5: Consideremos las variables MPG y WT del conjunto de datos **millaje** y que además la primera y última observacion han sido eliminadas. El plot de residuales versus la variable predictora WTO (WT excluyendo la primera y última observación) aparece en la fiura 4.9.

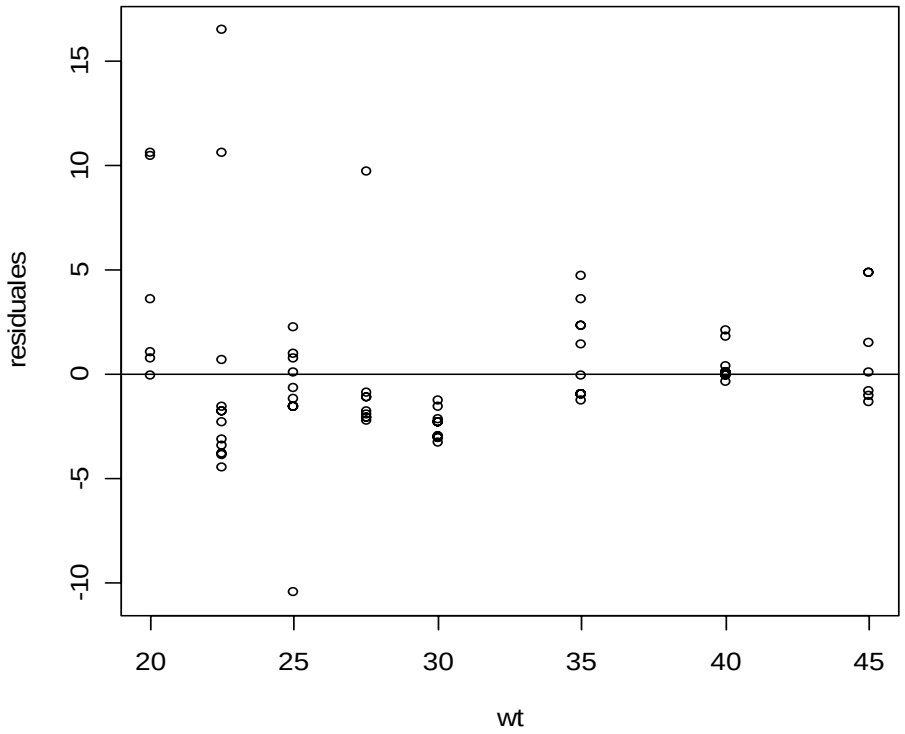


Figura 4.9. Plot de residuales versus la variable WTO donde se observa que la varianza no es homogénea

Aunque es díficil verlo en forma definitiva la variabilidad de los residuales está disminuyendo cuando la variable predictora aumenta. Algo más formal sería calcular la varianza de las y's por cada valor de X. esto produce la siguiente tabla de valores.

X_i	n_i	s_i^2
20.0	6	23.8987
22.5	12	42.7533
25.0	10	12.0049
27.5	9	14.5586
30.0	12	0.3961
35.0	12	4.4627
40.0	12	0.5973
45.0	7	7.2948

Haciendo un plot de x_i versus s_i^2 parece haber una buena relación cuadrática entre ambas variables. En la figura 4.9 se observa el plot de puntos y la regresión cuadrática La ecuación del modelo resulta ser

s_i^2 = 148.482 - 7.81488 Xi + 0.103609 Xi**2

El R^2=64.7.
Para determinar los pesos hay dos alternativas:

Primera alternativa: (Myers pag.281 Weisberg pag 85). Usar $w_i = \frac{1}{s_i^2}$, los s_i^2 están dados en la tabla anterior. Para usar esta alternativa debería haber un número razonable de observaciones y's para cada X_i.

Segunda Alternativa: (Draper y Smith, pag 226). Usar la ecuación de la regresión cuadrática para estimar las varianzas muestrales s_i^2 para cada x_i. Luego, escogemos los pesos como el reciproco de la varianzas muestrales estimadas. Es decir, $w_i = \frac{1}{\hat{s}_i^2}$, donde $\hat{s}_i^2$ es el valor correspondiente a un X_i en el modelo cuadrático.

Los resultados que se obtienen en R son los siguientes:

a) Análisis sin usar regresión ponderada.

```
> millaje1<-millaje[-c(1,82),c(1,3)]
> l1=lm(mpg~wt,data=millaje1)
> summary(l1)

Call:
lm(formula = mpg ~ wt, data = millaje1)

Residuals:
   Min     1Q Median     3Q    Max
-10.484  -1.992  -1.017   0.720  16.474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.40171    1.78634   37.73   <2e-16 ***
wt          -1.09671    0.05635  -19.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.825 on 78 degrees of freedom
Multiple R-Squared: 0.8293,     Adjusted R-squared: 0.8271
F-statistic: 378.8 on 1 and 78 DF,  p-value: < 2.2e-16
```

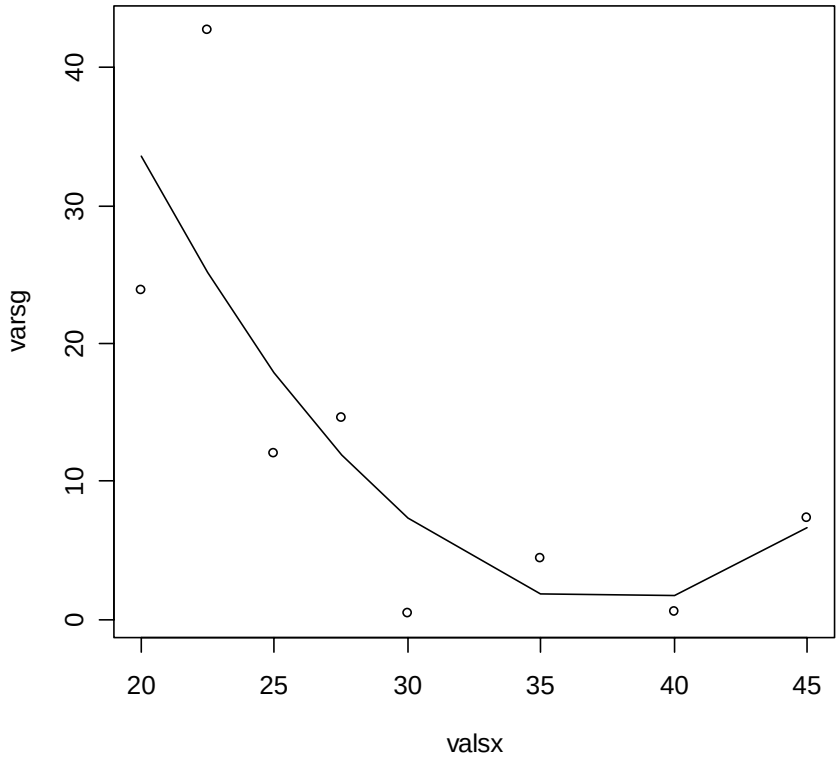


Figura 4.10. Ajuste cuadrático de la varianza versus la variable predictora

B) Análisis de regresión ponderada con la alternativa a).

```
> summary(lw1)

Call:
lm(formula = mpg ~ ., data = millaje1, weights = pesos)

Residuals:
   Min     1Q  Median     3Q    Max
-2.1007 -0.1768  0.1953  1.0526  3.2224

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.51692    1.06539   54.92   <2e-16 ***
wt          -0.86954    0.03107  -27.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.206 on 78 degrees of freedom
Multiple R-Squared: 0.9094,     Adjusted R-squared: 0.9083
```

F-statistic: 783.4 on 1 and 78 DF, p-value: < 2.2e-16

C) Análisis de Regresión ponderada usando la alternativa b)

```
> lw2<-lm(mpg~.,data=millaje1,weights=pesos1)
> summary(lw2)

Call:
lm(formula = mpg ~ ., data = millaje1, weights = pesos1)

Residuals:
   Min     1Q  Median     3Q    Max
-2.3321 -0.7047 -0.3122  0.2965  3.4499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.94848    1.72500   37.65   <2e-16 ***
wt          -1.02365    0.04738  -21.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.082 on 78 degrees of freedom
Multiple R-Squared: 0.8568,     Adjusted R-squared: 0.855
F-statistic: 466.8 on 1 and 78 DF,  p-value: < 2.2e-16
```

Observese que cuando se hace la regresión ponderada con la alternativa a) se obtiene una mejora del 7.0% en el R^2 mientras que con la alternativa b) solo se mejora un 2.8% .

4.6 Mínimos Cuadrados generalizados

Consideremos ahora la situación más general de que los errores no tiene varianza constante y además que son correlacionados. Sea el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

Supongamos ahora que $\mathbf{Var(e)=V}\sigma^2$, donde **V** es una matriz simétrica y definida positiva. Un caso particular de **V** es cuando los errores tienen distinta varianza y no están correlacionados.

Siempre es posible encontrar una matriz nosingular y simétrica **T** tal que $\mathbf{TT=T^2=V.}$ Mutiplicando ambos lados del modelo anterior por $\mathbf{T^{-1}}$ se obtiene

$$\mathbf{T^{-1}y} = \mathbf{T^{-1}X\beta} + \mathbf{T^{-1}e}$$

Sea $\mathbf{e^*}= \mathbf{T^{-1}e,}$ notando que $\text{Var}(\mathbf{e^*})=\text{Var}(\mathbf{T^{-1}e})=\mathbf{T^{-1}}\text{Var}(\mathbf{e})\mathbf{T^{-1}}=\mathbf{I}\sigma^2$ entonces el estimador mínimo cuadrático de $\boldsymbol{\beta}$ se obtiene minimizando

$$\mathbf{e^{*\prime}e^*} = \mathbf{(Y - X\beta)'V^{-1}(Y - X\beta)}$$

lo cual produce $\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$. Se puede ver fácilmente que $\mathbf{E}(\boldsymbol{\beta}^*)=\boldsymbol{\beta}$ y que $\mathbf{Var}(\boldsymbol{\beta}^*)=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Var(\mathbf{Y})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$.

Ejercicios

1. Usar el conjunto de datos **Highway**, con variable de respuesta es RATE y todas las otras como variables predictoras para responder las siguientes preguntas.

a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el R^2 de la regresión usando como predictora la que tiene mas alta correlacion
b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
d) Aplicar una transformación tipo potencia para establizar la varianza
e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.

2. Usar el conjunto de datos **Fuel** con variable de respuesta es Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas.

a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el R2 de la regresión usando como predictora la que tiene mas alta correlacion
b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
d) Aplicar una transformación tipo potencia para estabilizara la varianza
e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.

3. Usar el conjunto de datos **Headcirc** con variable de respuesta es headcirc (circunferencia de la cabeza del bebe) para responder a las siguientes preguntas.

a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el R^2 de la regresión usando como predictora la que tiene mas alta correlacion
b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
d) Aplicar una transformación tipo potencia para estabilizar la varianza.
e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.

4. Verificar la relación 4.7

5. Prueba para detectar varianza no constante (Cook y Weisberg, 1983). Consiste de los siguientes pasos:

a) Calcular la regression de Y versus todas las variables predictoras y guardar los residuales $\hat{e}_i$.

b) Calcular los residuales cuadrados escalados u_i definidos por $u_i = \frac{\hat{e}_i}{\tilde{\sigma}^2}$, donde $\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n}$ es el estimado máximo verosímil de σ^2.

c) Calcular la regresión de u_i versus las $\mathbf{z}_i$ incluyendo el intercepto. Los $\mathbf{z}_i$ son las variables de las que se sospecha que depende la varianza σ^2. Así, $\mathbf{z}_i = \hat{y}_i$ indica que la varianza varia con la variable de respuesta, $\mathbf{z}_i = x_i$ indica que la varianza varía con la predictora x_i. También $\mathbf{z}_i$ pueden contener todas las variables predictoras o un subconjunto de ellas. Guardar la Suma de cuadrado de la regresión (SSR).

d) Calcular la prueba S=SSR/2. Si S es grande entonces hay indicación de varianza no constante. Mas formalmente S se distribuye asintóticamente como una Ji cuadrado con q grados de libertad, donde q es el número de componentes de z_i, bajo la hipótesis nula de varianza constante.
Aplicar la prueba definida por los pasos a –d a los datos de los ejercicios 1 y 3.

6) Deducir que transformación de la variable de respuesta hace que la varianza σ^2 sea constante cuando ella es proporcional a $[E(y)]^4$

CAPÍTULO 5

REGRESIÓN CON VARIABLES CUALITATIVAS

5.1 Regresión con variables predictoras cualitativas.

Frecuentemente se considera que entre las variables predictoras, que explican el comportamiento de la variable de respuesta, hay algunas que son cualitativas o categóricas. Por ejemplo, si en una empresa se trata de explicar el salario de un empleado hay muchas variables predictoras a considerar algunas de ellas cuantitativas y otras cualitativas. Entre las variables cuantitativas estarán años de experiencia en la empresa, años de educación, edad, etc. y entre las variables cualitativas estarán el sexo del empleado, estado civil, jerarquía del empleado, etc.
Cuando una variable cualitativa asume solamente dos valores es llamada variable indicadora, variable binaria o variable "dummy". Estas variables son codificadas numéricamente con 0's y 1's.
Algunas veces la variable cualitativa puede asumir más de dos valores. Por ejemplo, la variable Opinión: A favor, Indeciso, En contra. Se podría codifucar los valores como 0, 1 y 2 pero esto estaría implicando una suposición de ordenamiento y además implicaría que el efecto de cambiar de A favor a Indeciso es lo mismo que cambiar de Indeciso a En contra (o sea se está suponiendo igual espaciamiento). Ambas suposiciones no son justificables. Lo mejor es definir dos variables indicadoras

A_1=1 A favor, 0 en otro caso
A_2= 1 En contra, 0 en otro caso

Usar una tercera variable es redundante puesto que los indecisos pueden ser representados por A_1=A_2=0. Estas variables cualitativas son llamadas mas propiamente variables nominales. A las variables cualitativas donde el orden si interesa se le conoce como variables ordinales y en ese caso es más frecuente codificar la variable como una secuencia ordenada de números enteros.

En un problema de regresión debe haber por lo menos una variable predictora cuantitativa. Si todas las variables predictoras fueran cualitativas entonces el problema se convierte en uno de **diseños experimentales.**

5.1.1 Regresión con una sola variable "dummy"

Consideremos un modelo de regresión con una sola variable cualitativa A y una variable cuantitativa X. Es decir,

$$Y=\beta_0+\beta_1X+\beta_2A + \varepsilon \qquad (5.1)$$

Notar que si A=0 se obtiene el modelo lineal simple

$$Y=\beta_0+\beta_1X + \varepsilon \qquad (5.2)$$

Y que si A=1 se obtiene el modelo

$$Y=(\beta_0+\beta_2)+\beta_1 X+\varepsilon \qquad (5.3)$$

Notar que las lineas estimadas de los modelos (5.2) y (5.3) serán paralelas (igual pendiente). El valor estimado de β_2 representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable "dummy".

Ejemplo 1. En el conjunto de datos **bajopeso**, disponible en la página de internet del texto, se trata de relacionar el peso de lo recién nacidos con los pesos de sus madres y la condición de fumar de las mismas. El conjunto de datos contiene 189 observaciones y será tratado en forma más completa más adelante.

Solución. Considerando que fumar=0 si la persona no fuma y 1 si fuma se obtiene los siguientes resultados

```
> l2<-lm(pbebe~pmama+fuma)
> summary(l2)

Call:
lm(formula = pbebe ~ pmama + fuma)

Residuals:
     Min       1Q   Median       3Q      Max
-2030.90  -445.69    29.16   521.76  1967.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2501.125    230.836  10.835   <2e-16 ***
pmama          4.237      1.690   2.507   0.0130 *
fuma        -272.081    105.591  -2.577   0.0107 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 707.8 on 186 degrees of freedom
Multiple R-Squared: 0.06777,    Adjusted R-squared: 0.05775
F-statistic: 6.761 on 2 and 186 DF,  p-value: 0.001464
```

Notar que el R^2 es bajísimo. El coeficiente de regresión estimado de fuma es –272 y significa que si la mama fuma en promedio el peso del bebe disminuirá en 272 libras.

Podemos hacer la regresión por grupos. Es decir, una regresión para los madres que no fuman y otras para las que si fuman. Se obtienen los siguientes resultados.

Para madres no fumadoras:

```
> l3<-lm(pbebe0~pmama0)
> summary(l3)

Call:
lm(formula = pbebe0 ~ pmama0)

Residuals:
    Min      1Q  Median      3Q     Max
-2029.87 -550.86   28.23  551.78 1976.84

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2350.578    326.583   7.197 7.21e-11 ***
pmama0         5.387      2.439   2.209   0.0292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 740.2 on 113 degrees of freedom
Multiple R-Squared: 0.04139,    Adjusted R-squared: 0.03291
F-statistic:  4.88 on 1 and 113 DF,  p-value: 0.02919
```

Para madres fumadoras:

```
> summary(l4)

Call:
lm(formula = pbebe1 ~ pmama1)

Residuals:
    Min      1Q  Median      3Q     Max
-2038.8  -414.4    35.0   473.6  1490.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2391.962    301.229   7.941 1.98e-11 ***
pmama1         2.965      2.274   1.304    0.196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 656.5 on 72 degrees of freedom
Multiple R-Squared: 0.02307,    Adjusted R-squared: 0.0095
F-statistic:   1.7 on 1 and 72 DF,  p-value: 0.1964
```

Notar que ambos casos los R^2 son más bajos que el R^2 anterior. Los plots pueden verse en la siguiente figura:

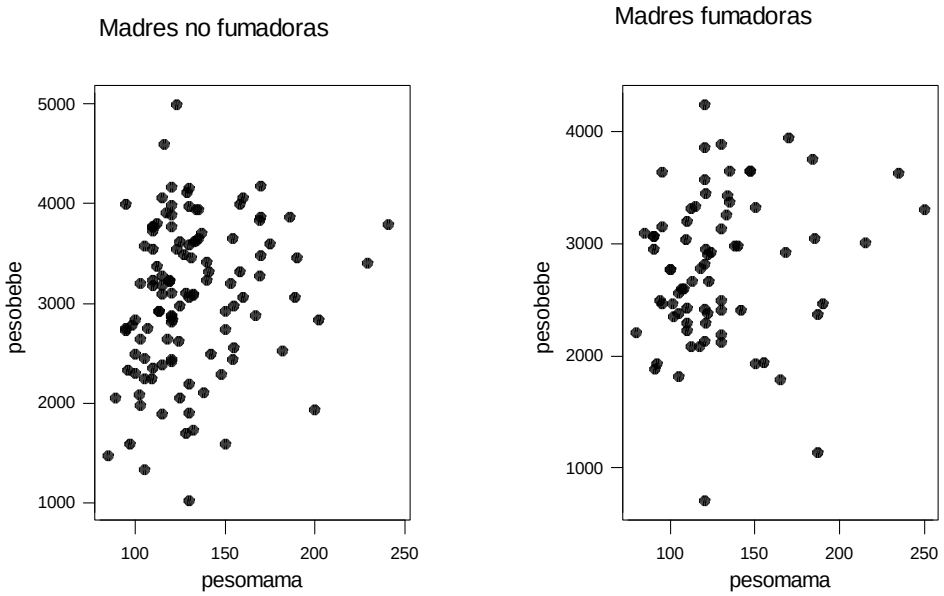


Figura 5.1. Plots de la relación pesobebe versus pesomama según la condición de fumar de la madre

En ambos plots se puede ver que no parece haber relación entre el peso del bebe y peso de la madre aunque esto es más evidente para las madres fumadoras. Los "outliers" parecen afectar más la regresión del peso bebe versus peso mama entre las madres no fumadoras.

Si se desea comparar las pendientes de las línea de regresión de los dos grupos se puede usar una prueba de t similar a la prueba de comparación de dos medias y asumiendo que hay homogeneidad de varianza. También se puede usar una prueba de F parcial o probando la hipótesis Ho: $\beta_3=0$ en el siguiente modelo

$$Y=\beta_0+\beta_1 A+\beta_2 X+\beta_3 AX +e$$

Cuando la hipótesis nula no es rechazada se concluye que la pendiente de regresion de ambos grupos son iguales. Si no hubiera igualdad de varianza de los dos grupos, habria que usar una prueba de t aproximada similar al problema de Behrens-Fisher. Aquí los grados de libertad se aproximan por

$$gl = \frac{(c_1 + c_2)^2}{\frac{c_1^2}{m-1} + \frac{c_2^2}{n-1}}$$

con $c_1 = \frac{s_1^2}{m}$ y $c_2 = \frac{s_2^2}{n}$.

Donde m y n son los grados de libertad de la suma de cuadrados del error en cada modelo, y s_1 y s_2 son las estimaciones de la varianza del error en cada modelo.

Algunas veces se desea comparar lineas de regresión para varios grupos. Supongamos que se tiene una variable predictora continua X para explicar el comportamiento de Y en tres grupos. Luego hay tres modelos de regresión que se pueden comparar. Estos son:

i) $Y=\beta_{01}+\beta_{11}X+\varepsilon$

ii) $Y=\beta_{02}+\beta_{12}X+\varepsilon$

iii) $Y=\beta_{03}+\beta_{13}X+\varepsilon$

Para relacionar las lineas de regresión hay que introducir 3 variables "dummy" para identificar los grupos G_1, G_2, y G_3 y 3 variables adicionales $Z_1=G_1X$, $Z_2=G_2X$, y $Z_3=G_3X$. (Otra alternativa sería usar solo dos variables "dummy"). Hay 4 posibilidades que podrian ocurrir:
a) Que las lineas se intersecten en un punto cualquiera, ya que tendrían diferente intercepto y pendiente. En este caso se ajusta el modelo $Y=\beta_{01}G_1+\beta_{11}Z_1+ \beta_{02}G_2+\beta_{12}Z_2+ \beta_{03}G_3+\beta_{13}Z_3+\varepsilon$
(Usando dos variables "dummy" este modelo sería $Y=\beta_0+\beta_1X+ \beta_{01}G_1+ \beta_{02}G_2+\beta_{11}Z_1+\beta_{12}Z_2+\varepsilon$)
b) Que las lineas sean paralelas (homogeneidad de pendientes). En este caso se ajusta el modelo $Y=\beta_{01}G_1+ \beta_{02}G_2+ \beta_{03}G_3+\beta X+\varepsilon$
c) Que las lineas tengan el mismo intercepto con el eje Y pero distintas pendientes (homogeneidad de interceptos). En este caso se ajusta el modelo $Y=\beta_0+\beta_{11}Z_1+\beta_{12}Z_2+\beta_{13}Z_3+\varepsilon$
d) Que las tres lineas coincidan. En este caso se ajusta el modelo $Y=\alpha+\beta X+\varepsilon$

Para probar la hipótesis Ho: el modelo satisface b) o c) o d) versus
Ha: el modelo satisface a)

Se usa una prueba de F parcial dada por

$$F_m=[(SSE_m-SSE_a)/(gl_m-gl_a)]/[SSE_a/gl_a]$$

Donde m, representa los modelos b,c, o d, y gl_m y gl_a representan los grados de libertad del error del modelo m y del modelo a, respectivamente. La F parcial se distribuye como una f con (gl_m-gl_a, gl_a)grados de libertad.

5.2 Regresión Logística

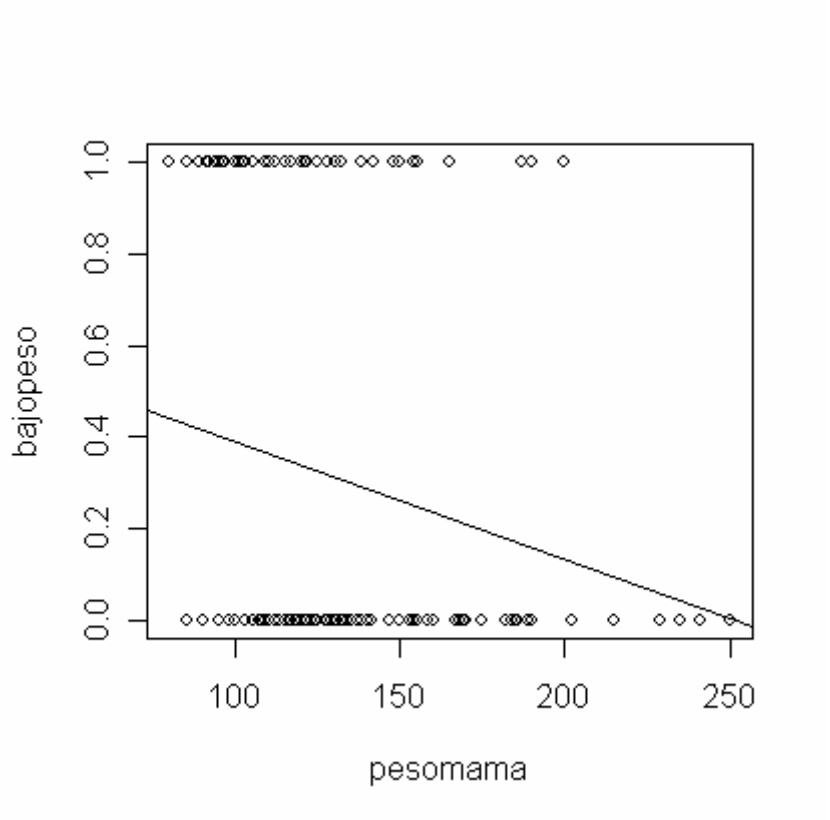
Consideraremos ahora que la variable de respuesta, Y, es una del tipo binario y que se tiene p variables predictoras x's, las cuales son consideradas aleatorias. Es decir, que el conjunto de datos consiste de una muestra de tamaño $n=n_1+n_2$, donde n_1 observaciones son de una clase C_1 y n_2 son de una clase C_2 . Así, para cualquier observación $\mathbf{x}_j$ la variable de respuesta Y es igual a 1 si $\mathbf{x}_j$ es de la clase C_1 , mientras que Y es igual a 0 si $\mathbf{x}_j$ pertenece a la clase C_2 .

Ejemplo 2: El conjunto de datos **bajopeso** contiene los pesos de 189 bebes recién nacidos. Para determinar si el niño es de bajo peso (menos de 2500 gramos) o no lo es, o sea **bajopeso**=1 si peso bebe<2500 y **bajopeso**=0 en otro caso, se han medido las siguientes variables predictoras

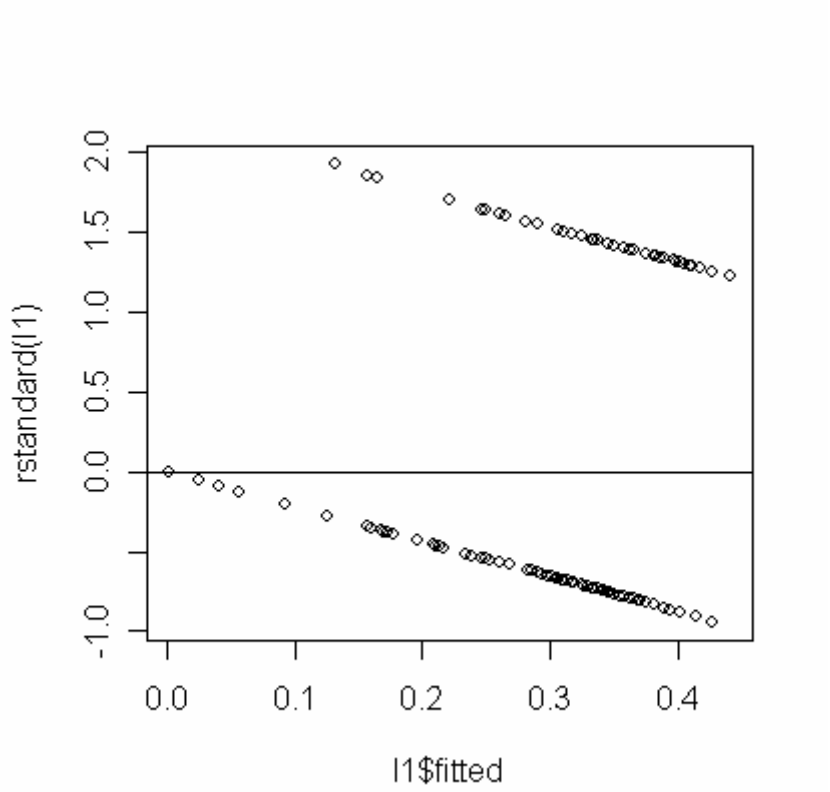
Edad: edad de la madre
Pesomama: peso de la madre en su último período muestral
Raza: raza de la madre:1=blanca, 2=negra, 3=otro

***Fuma**:* 0 ssi la madre no fuma, 1 si lo hace.
Prematur: número de partos prematuros de la madre
Hiperten: 0 si la madre no sufre de hipertensión, 1 si sufre
Uterirrit: *0 si no tiene utero irritado, 1 si lo tiene.*
Chequeos: número de visitas al médico en los tres primeros meses del embarazo.

Suponiendo que **pesomama** es la variable predictora más importante, podriamos explorar su relación con **bajopeso**. Haciendo un plot de **bajopeso** versus **pesomama** y ajustando una línea de regresión lineal simple se obtiene la siguiente figura

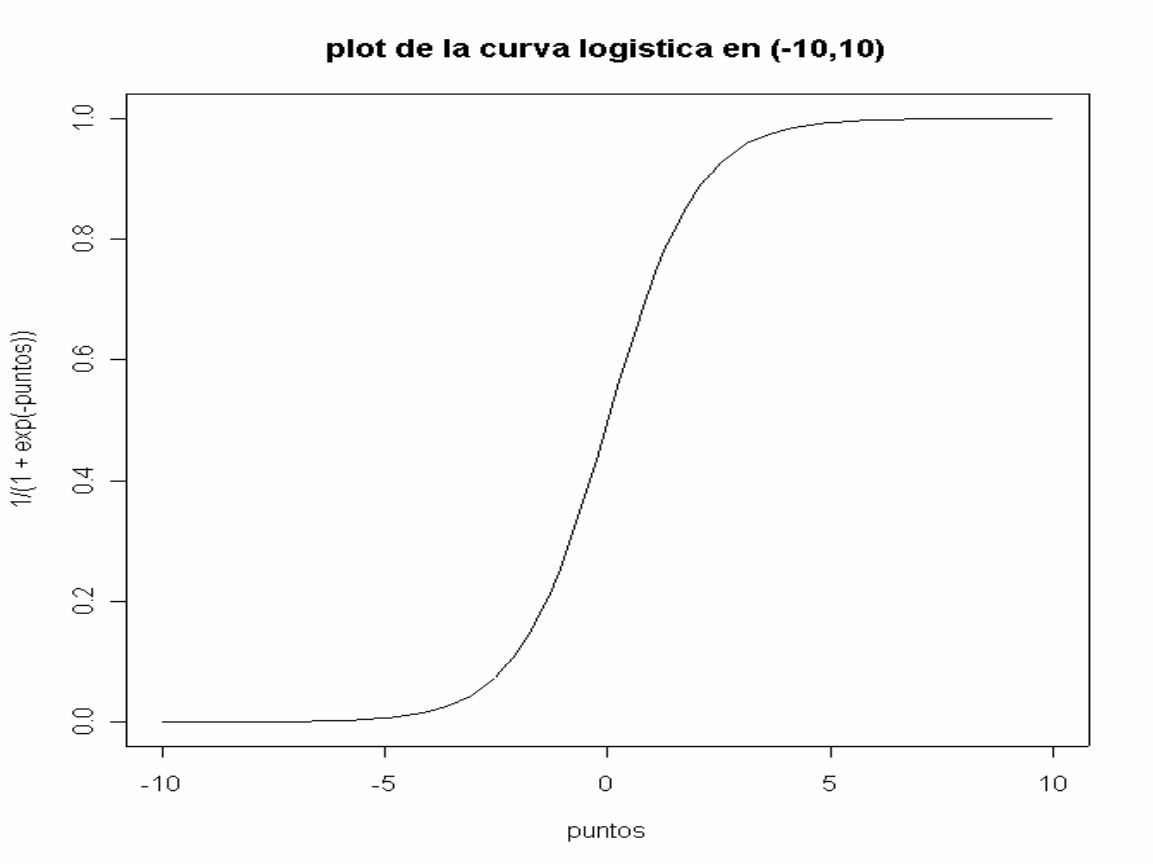


Como se puede ver es imposible que la linea de regresión represente la tendencia de los puntos. Además, la linea de regresión puede predecir valores de **bajopeso** que no son necesariamente 0 y 1 lo cual es totalmente ilógico. Asímismo, la suposición de varianza constante para la variable de respuesta no se cumple, como lo muestra el plot de residuales de la siguiente figura.



Se podría usar mínimos cuadrados reponderados para remediar esta situación pero aún asi conseguir predicciones 0 y 1 usando el modelo lineal sería imposible. Es más conveniente modelar la probabilidad de que la variable de respuesta asuma los valores 0 y 1 basado en las mediciones de las variables predictoras.

Notar que una curva en forma de S ajustaría bien los datos. Por otro lado, existe un modelo bien conocido en crecimiento poblacional cuya curva tiene esta forma y este modelo es llamado el modelo logístico y el cual se muestra en la siguiente figura. Propiamente se ha graficado $f(x) = \frac{1}{1+e^{-x}}$, -10<x<10. Esta curva también puede ser considerada como la gráfica de la distribución acumulada correspondiente a la densidad logística $p(x) = \frac{e^{-x}}{(1+e^{-x})^2}$



Sea $f(\mathbf{x}/C_i)$ (i=1,2) la función de densidad del vector aleatorio p-dimensional **x** en la clase C_i, en el modelo logístico se asume que

$$\log(\frac{f(\mathbf{x}/C_1)}{f\ (\mathbf{x}/C_2)}) = \alpha + \boldsymbol{\beta}'\mathbf{x} \qquad (5.4)$$

Aquí β es un vector de p parámetros y α representa el intercepto.

Por otro lado, sea p=P(Y=1/**x**) la probabilidad a posteriori de que Y sea igual a 1 para un valor observado de **x** ,entonces haciendo uso de probabilidad condicional se tiene que:

$$\frac{p}{1-p} = \frac{\dfrac{P\{Y=1\}f(\mathbf{x}/y=1)}{f(\mathbf{x})}}{\dfrac{P\{Y=0\}f(\mathbf{x}/y=0)}{f(\mathbf{x})}} = \frac{\pi_1 f\ (\mathbf{x}/C_1)}{\pi_2 f\ (\mathbf{x}/C_2)} \qquad (5.5)$$

donde π_i representa la *probabilidad a priori* de que **x** pertenezca a la clase C_i.. La expresión $\frac{p}{1-p}$ es llamado la razón de apuestas (***odds ratio***). Tomando logaritmos en ambos lados de (5.5) se obtiene

$$\log(\frac{p}{1-p}) = \log(\frac{\pi_1}{\pi_2}) + \log\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}$$

Usando la suposición (5.4), la ecuación anterior puede ser escrita como

$$\log(\frac{p}{1-p}) = \alpha + \beta' \mathbf{x} \qquad (5.6)$$

y $\log(\frac{p}{1-p})$ es llamado la transformacion ***logit***.

Despejando p de la expresión anterior se obtiene

$$p = \frac{\exp(\alpha + \beta' \mathbf{x})}{1 + \exp(\alpha + \beta' \mathbf{x})} \qquad (5.7)$$

La ecuación (5.7) representa el modelo de la regresión logística, que fue introducida en 1944 por J. Berkson. Notar que si las variables **x** en cada clase se distribuyen normalmente con igual matriz de covarianza Σ entonces se satisface la suposición (5.4) ya que

$$\log(\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}) = \mathbf{(u_1-u_2)' \Sigma^{-1}(x-1/2(u_1+u_2))} \qquad (5.8)$$

En este caso $\alpha = -\mathbf{(u_1 - u_2)'\Sigma^{-1}(u_1 + u_2)}/2$ y $\boldsymbol{\beta} = \mathbf{(u_1 - u_2)'\Sigma^{-1}}$. La suposición (5.4) se cumple también para otros tipos de distribuciones distintas de la normal multivariada tales como distribuciones de Bernoulli, y mezclas de éstas.

Un coeficiente b_k en el modelo de regresión logística estimado representa el cambio promedio de la logit función cuando la variable X_k cambia en una unidad adicional asumiendo que las otras variables permanecen constantes.
También se puede considerar que $\exp(b_k)$ es una razón de cambio de la razón de apuestas cuando X_k varia en una unidad adicional. Si X_k es binaria entonces $\exp(b_k)$ es el cambio en la razón de apuestas cuando ella asume el valor 1.
Cuando el modelo tiene una sola variable predictora, que además es binaria entonces existe una relación entre la regresión logística y el análisis de una tabla de contingencia 2 X 2.

5.2.1 Estimación del modelo logístico.

El método más usado para estimar α y **β** es el método de máxima verosimilitud. Dada una observación **x**, las probabilidades de que ésta pertenezca a las clases C_1 y C_2 son :

$$P(C_1/\mathbf{x}) = \frac{\exp(\alpha + \beta' \mathbf{x})}{1 + \exp(\alpha + \beta' \mathbf{x})} \qquad (5.9)$$

$$P(C_2/\mathbf{x}) = 1 - P(C_1/\mathbf{x}) = \frac{1}{1 + \exp(\alpha + \beta' \mathbf{x})} \qquad (5.10)$$

respectivamente.

Considerando una muestra de tamaño $n=n_1+n_2$ y un parámetro binomial p igual a $\exp(\alpha+\boldsymbol{\beta}'\mathbf{x})/(1+\exp(\alpha+\boldsymbol{\beta}'\mathbf{x})$ la función de verosimilitud es de la forma

$$L(\alpha,\boldsymbol{\beta})=\prod_{i=1}^{n_1}\frac{\exp(a+\mathbf{x_i}'\boldsymbol{\beta})}{1+\exp(\alpha+\mathbf{x_i}'\boldsymbol{\beta})}\cdot\prod_{j=n_1+1}^{n}\frac{1}{1+\exp(\alpha+\mathbf{x_j}'\boldsymbol{\beta})} \quad (5.11)$$

asumiendo que las primeras n_1 observaciones son de la clase C_1 y las restantes son de la clase C_2. Los estimados $\tilde{\alpha}$ y $\tilde{\boldsymbol{\beta}}$ son aquellos que maximizan la función anterior y son encontrados aplicando métodos iterativos tales como Newton-Raphson (SAS) o minimos cuadrados reponderados iterativos (MINITAB, R/S-Plus).

La solución de la función de verosimilitud puede no ser única si existe una marcada separación entre las dos clases.

Otra forma de hacer la estimación es como sigue: Los parámetros α y $\boldsymbol{\beta}$ pueden ser estimados haciendo la regresión lineal múltiple de logit($\hat{p}$) versus $x_1,x_2,\ldots.x_p$. Usando los resultados de la sección 4.4 para aproximación de la varianza de una transformación se tiene que

$$Var[\ln(\frac{\hat{p}}{1-\hat{p}})]\cong[\frac{1}{p(1-p)}]^2\frac{p(1-p)}{n_1}=\frac{1}{n_1p(1-p)} \quad (5.12)$$

Como $p=p(\mathbf{x})$ se llega a un problema donde la varianza no es constante y se puede usar mínimos cuadrados ponderados con pesos $w_i(\mathbf{x})=n_1\,\hat{p}\,(\mathbf{x})(1-\hat{p}\,(\mathbf{x}))$ para estimar los parámetros α y $\boldsymbol{\beta}$ del modelo logístico.

La regresión logística es un caso particular de los modelos lineales generalizados (GLM) propuesto por Nelder y Wedderburn (1972). Los modelos lineales generalizados extienden los modelos lineales en dos sentidos: Primero con la especificación de una **función link** que relaciona el esperado de la variable de respuesta con las predictoras lineales y segundo con la especificación de una función de distribución de los errores que es distinta de la Gaussiana. La forma de un modelo lineal generalizado es

$$\eta(E(y))=\alpha+\boldsymbol{\beta}\mathbf{x}$$

donde $\eta(.)$ es la función link. En un modelo de regresión clásico, $\eta(t)=t$ y la distribución de los errores es Gaussiana o Normal. En la regresión logística $\eta(t)=\log(t/(1-t))$ y la distribución de los errores es binomial. El modelo de regresión viene dado por $E(y)=\eta^{-1}(\alpha+\boldsymbol{\beta}\mathbf{x})$

En **MINITAB** el menú de **Regresión** contiene tres tipos de regresión logística: regresión logística binaria (aplicada a dos clases), regresión logística ordinal (si hay mas de dos clases) y regresión logística nominal (si hay mas de dos clases no ordenadas). Para ajustar un modelo logístico en SAS se usa el procedimiento **LOGISTIC,** mientras que en **R y S-Plus** se usa el procedimiento **glm** (modelos lineales generalizados) con la opción family=binomial. Aquí **family** representa el tipo de distribución de los errores.

5.2.2 Medidas de Confiabilidad del Modelo

Las siguientes son unas medidas que cuantifican el nivel de ajuste del modelo al conjunto de datos:

a) La Devianza: Es similar a la suma de cuadrados del error de la regresión lineal y se define como el negativo de dos veces la función de verosimilitud maximizada. Para los casos cuando la variable de respuesta Y no está agrupada se tiene que:

$$D = -2\{\sum_{i:y_i=1}^{n} \log(\hat{p}_i) + \sum_{i:y_i=0}^{n} \log(1-\hat{p}_i)\}$$

D es equivalente a la prueba de razón de verosimilitud para probar la validez del modelo logístico. El estadístico D se distribuye como una Ji-Cuadrado con n-p-1 grados de libertad, donde p es el número de variables predictoras. Si D es mayor que una Ji-Cuadrado con n-p-1 grados de libertad para un nivel de significación dado entonces el modelo logístico no es confiable.

b) El Pseudo-R^2. Es similar al R^2 de la regresión lineal se define por

$$Pseudo - R^2 = (1 - \frac{Devianza}{Devianza.Nula})100\%$$

donde la Devianza Nula es la Devianza considerando solamente el intercepto y que se distribuye como una Ji-Cuadrado con n-1 grados de libertad. Para hallar la Devianza Nula se hace una regresión logística considerando que hay una sola variable predictora cuyos valores son todos unos. Notar que Devianza=Devianza Nula-Devianza Residual.

c) El Criterio de Información de Akaike (AIC): Se define por

$$AIC=D+2(p+1)$$

Donde p es el número de variables predictoras. Un modelo es mejor que otro si su AIC es más pequeño.

d) La Prueba de Bondad de Ajuste de Hosmer-Lemeshov. Se aplica cuando los datos son dados en forma agrupada y se define por

$$C = \sum_{i=1}^{g} \frac{(O_i - n'_i\ \overline{p}_i)^2}{n'_i\ \overline{p}_i(1-\overline{p}_i)}$$

Donde g es el número de grupos, n'_i es el numero de observaciones en el i-ésimo grupo

O_i es la suma de las y's en el i-ésimo grupo y $\overline{p}_i$ es el promedio de las proporciones p_i del evento que esta siendo considerado en el i-ésimo grupo.

5.2.3 Estadisticas Influenciales para regresión logística

Existen varios tipos de residuales que permiten cotejar si una observación es influencial o no.

a) Residuales de Pearson: Están definidos por

$$r_i = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1-\hat{p}_i)}}$$

donde, si los valores de la variable de respuesta están agrupadas, y_i representa el número de veces que y=1 entre las m_i repeticiones de X_i , de lo contrario m_i=1 para todo i.

El residual de Pearson es similar al residual estudentizado usado en regresión lineal. Así un residual de Pearson en valor absoluto mayor que 2 indica un dato anormal.

b) Residuales de Devianza: Están dados por

$D_i = -\sqrt{2\,|\log(1-\hat{p}_i)|}$ si y_i=0 y por $D_i = \sqrt{2\,|\log(\hat{p}_i)|}$ si y_i=1.

Si el residual de devianza es mayor que 2 en valor absoluto entonces la observación correspondiente es anormal. Estos son los residuales dados por R.

Ejemplo 3. Aplicar regresión logística a los datos del ejemplo 2.

```
> # Haciendo la regresion logistica simple con la predictora pesomama
> logis1<-glm(bajopeso~pesomama,data=pesobebe,family=binomial)
> summary(logis1)

Call:
glm(formula = bajopeso ~ pesomama, family = binomial, data = pesobebe)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.0951  -0.9022  -0.8018   1.3609   1.9821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
pesomama    -0.01406    0.00617  -2.279   0.0227 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 228.69  on 187  degrees of freedom
AIC: 232.69

Number of Fisher Scoring iterations: 4

> #Haciendo la regresion logistica multiple usando todas las variables predictoras
> logis2<-glm(bajopeso~.,data=pesobebe,family=binomial)
> summary(logis2)

Call:
```

```
glm(formula = bajopeso ~ ., family = binomial, data = pesobebe)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8832  -0.8178  -0.5574   1.0288   2.1451

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.078975   1.276254  -0.062  0.95066
edad        -0.035845   0.036472  -0.983  0.32569
pesomama    -0.012387   0.006614  -1.873  0.06111 .
raza         0.453424   0.215294   2.106  0.03520 *
fuma         0.937275   0.398458   2.352  0.01866 *
prematuros   0.542087   0.346168   1.566  0.11736
hipertensio  1.830720   0.694135   2.637  0.00835 **
uteroirrit   0.721965   0.463174   1.559  0.11906
chequeos     0.063461   0.169765   0.374  0.70854
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 204.19  on 180  degrees of freedom
AIC: 222.19

Number of Fisher Scoring iterations: 4

> #Haciendo otra vez la regresion logistica incluyendo solo las variables mas significativas
> logis3<-glm(bajopeso~pesomama+raza+fuma+hipertensio,data=pesobebe,family=binomial)
> summary(logis3)

Call:
glm(formula = bajopeso ~ pesomama + raza + fuma + hipertensio,
    family = binomial, data = pesobebe)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.7988  -0.8865  -0.5847   1.0997   2.2503

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.357536   1.010584  -0.354  0.72350
pesomama     -0.015354   0.006523  -2.354  0.01858 *
raza          0.489555   0.207324   2.361  0.01821 *
fuma          1.080020   0.383735   2.814  0.00489 **
hipertensio   1.744272   0.687563   2.537  0.01118 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 211.04  on 184  degrees of freedom
AIC: 221.04.
Number of Fisher Scoring iterations: 4
```

Observando el valor de la devianza residual y del AIC el tercer modelo seria el mejor modelo. Notar que la Devianza=Devianza Nula –Devianza Residual=23.63 el cual habría que compararlo con una Ji-Cuadrado con 184 grados de libertad para un nivel de significación dado. Usando un nivel de significación del 5%, la Ji-Cuadrado da 216.64. En consecuencia los datos parecen ajustarse a un modelo logístico.

De acuerdo a los residuales de Pearson las siguientes observaciones pueden ser influenciales

```
> y=pesobebe$bajopeso
>pihat=logis3$fit
>rp=(y-pihat)/sqrt(pihat*(1-pihat))
> rp[abs(rp)>2]
      13       132       147       152       155       170       183
-2.010543  2.539516  3.402709  2.048327  2.700368  2.178067  2.345670
```

De acuerdo a los residuales de devianza las siguientes observaciones pueden ser influenciales

```
>r1=sqrt(2*abs(log(pihat[y==1])))
>r2=-sqrt(2*abs(log(1-pihat[y==0])))
>rd=c(r2,r1)
>rd[abs(rd)>2]
    132      147      155
2.004045 2.250326 2.056837
```

5.2.4 Uso de la regresión logística en Clasificación:

Para efectos de clasificación la manera más facil de discriminar es considerar que si p>0.5 entonces la obervación pertenece a la clase que uno está interesado. Pero algunas veces esto puede resultar injusto sobre todo si se conoce si una de las clases es menos frecuente que la otra.

Metodos alternos son:

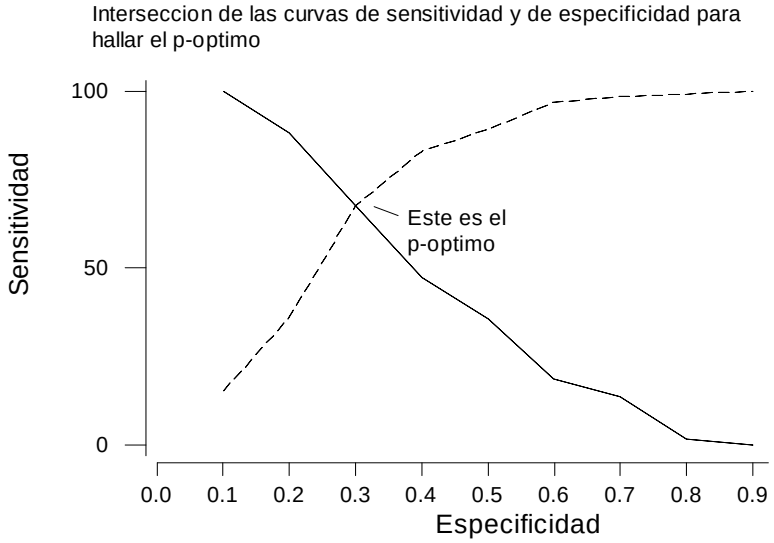
a) Plotear el porcentaje de observaciones que poseen el evento que han sido correctamente clasificadas (**Sensitividad**) versus distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (**especifidad**) versus los mismos niveles de probabilidad anteriormente usados, en la misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtienen intersectando las dos curvas.
b) Usar la curva ROC (receiver operating characteristic curva). En este caso se grafica la sensitividad versus (1-especifidad)100%, y se coge como el p ideal aquel que está más cerca a la esquina superior izquierda, o sea al punto (0,100).

Ahora aplicaremos la regresión logística como un clasificador a los datos del ejemplo anterior. En lo que sigue vamos a considerar los resultados del segundo modelo.
Prediciendo las clases con el segundo modelo usando el método mas simple es decir comparando el valor ajustado por la regresión logística con p=0.5 y asignando la observación a la clase 1 se obtienen 52 de las 189 observaciones mal clasificadas lo cual representa una tasa de mala clasificacion del 27.51%
Haciendo la clasificación con el método mas complicado calculando la sensitividad y especificidad se obtiene la siguiente tabla

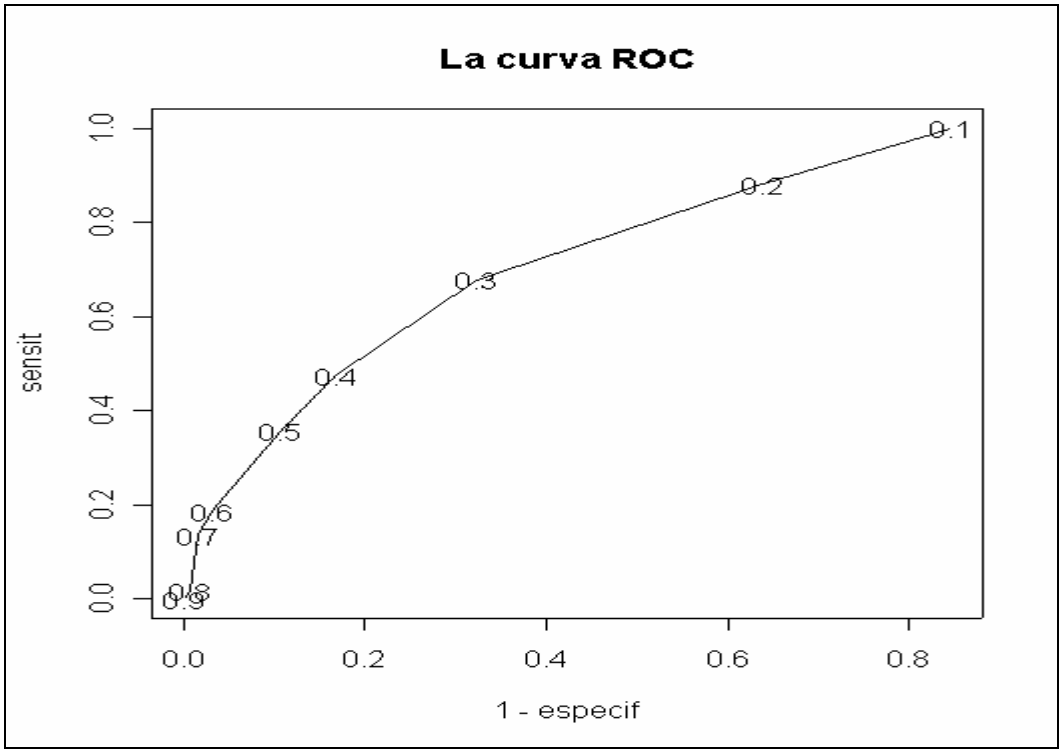
Notar que para p=.30 la curva está mas cerca a la esquina superior izquierda. La tasa de mala clasificación optima es= 0.3227513

Sensitividad	Especifidad	P	(1-especifidad)%
100.00	15.38	0.10	84.62
88.14	36.15	0.20	63.85
79.66	53.85	0.25	46.15
67.80	67.69	0.30	32.31
55.93	77.69	0.35	22.31
47.46	83.08	0.40	16.92
35.59	89.23	0.50	10.77
18.64	96.92	0.60	3.08
13.56	98.46	0.70	1.54
1.69	99.23	0.80	0.77
0.00	100.00	0.90	0.00

Las gráficas de los dos métodos aparecen en la siguiente figura y en ambos caso el p-óptimo a usarse es p=0.3



La regresión logística se puede extender al caso donde hay más de dos clases y recibe el nombre de regresión logística politómica. Este tipo de regresión es estudiada en mas detalle en un curso de clasificación. También existe una relación entre regresión logistica y redes neuronales.



Ejercicios

1. Comparando lineas de regresión. Considerar el conjunto de datos bajopeso disponible en la página de internet del texto y tomar al peso del bebé como Y y a peso de la mamá como X. Comparar las pendientes y los interceptos de la línea de regresión en los tres grupos de raza de la madre.

2. Considerar el conjunto de datos **heartc** disponible en www.uprm.edu/~edgar/datosclass.html en el cual se toman 13 mediciones a 297 pacientes para clasificarlos en propensos o no propensos a sufrir ataque cardíaco. Las clases están en la última columna y están codificadas como 1 y 2.

a) Usar los criterios de la Devianza y del AIC para determinar un modelo de logistico óptimo
b) Determinar la bondad de ajuste del modelo
c) Identificar posibles valores influenciales
d) Determinar la tasa de mala clasificación según las distintas maneras consideradas en el texto.

3**.** Considerar el conjunto de datos **breastw** disponible en www.uprm.edu/~edgar/datosclass.html en el cual se toman 9 mediciones a 699 mujeres para clasificarlas en propensas o no propensas a tener cáncer al seno. Las clases están en la última columna y están codificadas como 1 y 2.

a) Usar los criterios de la Devianza y del AIC para determinar un modelo de logistico óptimo
b) Determinar la bondad de ajuste del modelo
c) Identificar posibles valores influenciales
d) Determinar la tasa de mala clasificación según las distintas maneras consideradas en el texto.

4. Regresión logística con datos agrupados. En una Universidad se registra el número de estudiantes que pasaron con A de acuerdo a las veces que habian tomado de antemano un curso de estadistica. Los resultados se muestran en la siguiente tabla

Veces que habia tomado el curso anteriormente	Número de estudiantes	Estudiantes pasando con A
0	300	30
1	150	25
2	80	20
3	35	5
4	20	3
5	5	1

a) Construir una regresión logística para predecir la probabilidad de que un estudiante obtenga A en la clase de acuerdo a las veces que la ha tomado antes.
b) Probar si la variable predictora: número de veces que el estudiante ha tomado antes el curso e s significativa o no?
c) Determinar la bondad de ajuste del modelo.

CAPÍTULO 6

SELECCIÓN DE VARIABLES EN REGRESIÓN

Selección de variables o también llamado selección de un subconjunto de predictoras es un procedimiento estadístico que es importante por diversas razones, entre estas están:

a) No todas las variables predictoras tienen igual importancia, por lo tanto es más eficiente trabajar con un modelo donde las variables importantes estén presentes y las que tienen poca importancia no aparezcan.
b) Algunas variables pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras, luego se hace necesario eliminarlas ya que son redundantes.
c) Computacionalmente es más fácil trabajar con un conjunto de variables predictoras pequeño.
d) Es más económico recolectar información para un modelo con pocas variables.
e) Si se reduce el número de variables entonces el modelo se hace más **parsimonioso**. Se dice que un modelo es **parsimonioso** si consigue ajustar bien los datos pero usando la menor cantidad de variables predictoras posibles. Es más conveniente porque sus predicciones son más confiables y además es más robusto que el modelo original.

Desde que empezó a trabajarse en esta área en los años 60 y gracias al desarrollo de las computadoras se han introducido muchos métodos de selección de variables. Aquí describiremos sólo algunos de ellos.

6.1 Metodos "Stepwise"

La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo (o excluyendo) una sola variable predictora en cada paso de acuerdo a ciertos criterios. El proceso secuencial termina cuando una regla de parada se satisface.

Hay tres algoritmos más conmúnmente usados, los cuales serán descritos a continuación.

6.1.1 "Backward Elimination" (Eliminación hacia atrás).

En este caso se comienza con el modelo completo y en cada paso se va eliminando una variable. Si resultara que todas las variables predictoras son importantes, es decir tienen "p-value" pequeños para la prueba t, entonces no se hace nada y el mejor modelo es el que tiene todas las variables predictoras disponibles. En caso contrario, en cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de estos requisitos equivalents entre sí:

a) Aquella variable que tiene el estadístico de t, en valor abosluto, más pequeño entre las variables incluidas aún en el modelo. Es decir, aquella variable con el F parcial más pequeño. El F parcial está definido por:

$$F_p=[SSR_k-SSR_{k-1}]/MSE_k$$

donde SSR_k es la suma de cuadrados debido a la regresión con k variables y SSR_{k-1} es la misma suma con k-1 variables. $MSE_k=SSE_k/n-k-1)$ es el cuadrado medio del error del modelo que incluye k variables. Hay que calcular el F_p para cada una de las variables presentes aún en

el modelo y se elimina del modelo aquella variable que dá el F_p mas pequeño. Se puede mostrar que $t^2=F_p$. En realidad todo el proceso se entiende mucho mejor con la t que con la F.

b) Aquella variable que produce la menor disminución en el R^2 al ser eliminada del modelo. Es decir, aquella variable que produce el mas pequeño incremento en la suma de cuadrados del error.
c) Aquella variable que tiene la correlación parcial (en valor absoluto) más pequeña con la variable de respuesta, tomando en cuenta las variables que quedarían en el modelo. La correlación parcial de Y con la variable X_i se define como la correlación entre los residuales de la regresión de Y con todas las variables predictoras, excepto X_i y los residuales de la regresión de X_i con todas las otras restantes variables predictoras.

El método "Backward" padece del efecto de anidamiento ya que toda variable que es eliminada del modelo ya no vuelve a entrar a él.

El proceso termina cuando se cumple una de las siguientes condiciones:

a) Se llega a un modelo con un número prefijado p* de variables predictoras.
b) El valor de la prueba de F parcial para todas las variables incluidas en el modelo son mayores que un número prefijado F-out (por lo general este valor es 4). O en forma equivalente, se para cuando el valor absoluto del estadístico de t para cada variable es mayor que la raíz cuadrada de F-out (por lo general, |t|>2). Algunas veces se prefija de antemano un nivel de significación dado α* (digamos del 10%) para la prueba de t o de F parcial en cada paso y en este caso se termina el proceso cuando todos los p-values son menores que α*.

6.1.2 "Forward Selection" (Selección hacia adelante).

Aquí se empieza con la regresión lineal simple que considera como variable predictora a aquella que esta más altamente correlacionada (sin tomar en cuenta el signo) con la variable de respuesta. Si esta primera variable no es significativa entonces se considera el modelo $\hat{Y} = \overline{Y}$ y se para el proceso, de lo contrario se sigue y en el siguiente paso se añade al modelo la variable que reune cualquiera de estos requisitos equivalents entre sí:

a) Aquella variable que tiene el estadístico de t , en valor absolute, más grande entre las variables no incluidas aún en el modelo. Es decir, la variable con el F-parcial más grande.
b) Aquella variable que produce el mayor incremento en el R^2 al ser añadida al modelo. Es decir, aquella variable que produce la mayor reducción en la suma de cuadrados del error.
c) Aquella variable que tiene la correlación parcial más alta (en valor absoluto) con la variable de respuesta, tomando en cuenta las variables ya incluidas en el modelo.

Aquí también está presente el efecto de anidamiento ya que toda variable que es añadida al modelo ya no puede salir del mismo.

El proceso termina cuando se cunple una de las siguientes condiciones:

a) Se llega a un modelo con un número prefijado p* de variables predictoras.
b) El valor de la prueba de F parcial para cada una de las variables no incluidas aun en el modelo es menor que un número prefijado F-in (por lo general este valor es 4). O en forma equivalente se para cuando el valor absoluto del estadistico de t es menor que la raíz cuadrada de F-in (por lo general, |t|<2). Algunas veces se prefija de antemano un nivel de significación

dado α^* (digamos del 15%) para la prueba de t o de F parcial en cada paso y en este caso se termina el proceso cuando todos los p-values de la prueba t de las variables no incluidas aún son mayores que α^*.

6.1.3 "Stepwise Selección" (Selección Paso a Paso)

Fue introducido por Efroymson (1960) para subsanar el problema de anidamiento de los dos métodos anteriores. Se puede considerar como una modificación del método "Forward". Es decir, se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable en forma similar al método forward, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada. Aqui se usan F-out y F-in con F-in ≤ F-out.
El proceso termina cuando ninguna de las variables, que no han entrado aún, tienen importancia suficiente como para entrar al modelo.

Prácticamente todos los programas estadísticos ejecutan los procedimientos "stepwise". En MINITAB se sigue la secuencia **STAT ▸ Regression ▸ Stepwise**. En S-Plus existe la función **stepwise** que tiene la opción **method**, la cual permite elegir entre el método backward, forward y stepwise (Efroymson).

Ejemplo 1: Aplicar los métodos "stepwise" al conjunto de datos **grasa**. La variable de respuesta grasa: porcentaje de grasa en el cuerpo. En 252 sujetos se midieron las siguientes variables predictoras:

edad (en años)
peso (en libras)
altura (en pulgadas)
cuello (circunferencia en cms)
pecho (circunferencia en cms)
abdomen (circunferencia en cms)
cadera (circunferencia en cms)
muslo (circunferencia en cms)
rodilla (circunferencia en cms)
tobillo (circunferencia en cms)
biceps (circunferencia en cms)
antebrazo (circunferencia en cms)
muneca (circunferencia en cms)

para predecir su porcentaje de grasa en el cuerpo

Primero aplicaremos el método de eliminación hacia atrás con un nivel de significación del 10%.

Stepwise Regression: grasa versus edad, peso, ...

```
Backward elimination.  Alpha-to-Remove: 0.1

 Response is  grasa   on 13 predictors, with N =  252

    Step          1        2        3        4        5        6        7
```

Constant	-18.19	-17.93	-19.69	-26.00	-23.30	-22.66	-33.26
edad	0.062	0.063	0.062	0.065	0.063	0.066	0.068
T-Value	1.92	2.00	2.00	2.11	2.06	2.14	2.21
P-Value	0.056	0.046	0.046	0.036	0.041	0.034	0.028
peso	-0.088	-0.088	-0.093	-0.107	-0.098	-0.090	-0.119
T-Value	-1.65	-1.70	-1.96	-2.55	-2.42	-2.25	-3.51
P-Value	0.100	0.091	0.051	0.011	0.016	0.025	0.001
altura	-0.070	-0.069	-0.064				
T-Value	-0.72	-0.72	-0.69				
P-Value	0.469	0.470	0.493				
cuello	-0.47	-0.47	-0.48	-0.47	-0.49	-0.47	-0.40
T-Value	-2.02	-2.06	-2.08	-2.05	-2.18	-2.08	-1.83
P-Value	0.044	0.040	0.039	0.042	0.030	0.039	0.068
pecho	-0.024	-0.024					
T-Value	-0.24	-0.25					
P-Value	0.810	0.805					
abdomen	0.955	0.954	0.944	0.958	0.949	0.945	0.918
T-Value	11.04	11.09	12.51	13.16	13.18	13.13	13.21
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
cadera	-0.21	-0.21	-0.20	-0.18	-0.18	-0.20	
T-Value	-1.42	-1.42	-1.41	-1.29	-1.32	-1.41	
P-Value	0.156	0.156	0.161	0.199	0.189	0.159	
muslo	0.24	0.24	0.25	0.26	0.27	0.30	0.22
T-Value	1.64	1.72	1.81	1.94	1.99	2.34	1.91
P-Value	0.103	0.086	0.072	0.054	0.048	0.020	0.057
rodilla	0.02						
T-Value	0.06						
P-Value	0.950						
tobillo	0.17	0.18	0.18	0.18			
T-Value	0.79	0.81	0.82	0.85			
P-Value	0.433	0.419	0.412	0.396			
biceps	0.18	0.18	0.18	0.19	0.18		
T-Value	1.06	1.06	1.05	1.10	1.06		
P-Value	0.290	0.289	0.297	0.271	0.289		
antebraz	0.45	0.45	0.45	0.45	0.45	0.52	0.55
T-Value	2.27	2.29	2.28	2.31	2.31	2.77	2.99
P-Value	0.024	0.023	0.023	0.022	0.022	0.006	0.003
muneca	-1.62	-1.62	-1.61	-1.66	-1.54	-1.54	-1.53
T-Value	-3.03	-3.04	-3.04	-3.14	-3.03	-3.02	-3.00
P-Value	0.003	0.003	0.003	0.002	0.003	0.003	0.003
S	4.31	4.30	4.29	4.28	4.28	4.28	4.29
R-Sq	74.90	74.90	74.90	74.85	74.77	74.66	74.45
R-Sq(adj)	73.53	73.64	73.75	73.81	73.84	73.82	73.71
C-p	14.0	12.0	10.1	8.5	7.2	6.4	6.3

Notar que en el último paso los “P-values” de la prueba t son todos menores que 10%, y que los t-values son aproximadamente mayores que 2 en valor absoluto. Mas especificamente el valor

crítico de F corresponde a una F(1,n-k-1,α)=F(1,244,.10)=2.72, aqui k=7 número de variables presentes en el modelo en el paso 7, y el correspondiente valor critico de t es 1.65. Notar que el R^2 ha bajado muy poco. Las variables que quedan en el modelo son: edad, peso, cuello, abdomen, muslo, antebrazo y muñeca. Si se escoge un nivel de significación del 5% entonces el proceso termina en 10 pasos y solo quedan cuatro variables en el modelo: peso, abdomen, antebrazo y muñeca.

Ahora aplicaremos el método de selección hacia adelante, usando un nivel de significación del 15%.

Stepwise Regression: grasa versus edad, peso, ...

```
Forward selection.  Alpha-to-Enter: 0.15

 Response is  grasa   on 13 predictors, with N =  252


    Step          1        2        3        4        5        6        7
Constant     -39.28   -45.95   -27.93   -34.85   -30.65   -25.89   -33.26

abdomen        0.631    0.990    0.975    0.996    1.008    0.945    0.918
T-Value        22.11    17.45    17.37    17.76    17.89    13.82    13.21
P-Value        0.000    0.000    0.000    0.000    0.000    0.000    0.000

peso                   -0.148   -0.114   -0.136   -0.123   -0.094   -0.119
T-Value                 -7.11    -4.84    -5.48    -4.75    -2.98    -3.51
P-Value                 0.000    0.000    0.000    0.000    0.003    0.001

muneca                           -1.24    -1.51    -1.25    -1.59    -1.53
T-Value                          -2.85    -3.40    -2.66    -3.09    -3.00
P-Value                          0.005    0.001    0.008    0.002    0.003

antebraz                                   0.47     0.53     0.57     0.55
T-Value                                    2.60     2.86     3.08     2.99
P-Value                                   0.010    0.005    0.002    0.003

cuello                                             -0.37    -0.40    -0.40
T-Value                                            -1.65    -1.81    -1.83
P-Value                                            0.100    0.072    0.068

edad                                                        0.046    0.068
T-Value                                                      1.60     2.21
P-Value                                                     0.110    0.028

muslo                                                                 0.22
T-Value                                                               1.91
P-Value                                                              0.057

S              4.88     4.46     4.39     4.34     4.33     4.31     4.29
R-Sq          66.17    71.88    72.77    73.50    73.79    74.06    74.45
R-Sq(adj)     66.03    71.65    72.44    73.07    73.26    73.43    73.71
C-p            72.9     20.7     14.2      9.3      8.6      8.0      6.3
```

La siguiente tabla lista los valores de t y los p-values correspondientes para cada una las variables no incluídas en el modelo, cuando se hace la regresión considerando las variables ya incluídas y cada una de las que falta incluir.

	t-value	P-value
altura	-0.48	0.629
pecho	0.20	0.840
cadera	-1.41	0.159
rodilla	-0.01	0.991
tobillo	0.83	0.406
biceps	1.18	0.240

Notar que todas las variables no incluídas aún en el modelo tiene ”P-value” grande, mayor de 0.15, lo cual indica que ellas son no significativas. En consecuencia, el proceso termina. El valor de F critico es F(1,243,.15)=2.085 y el de t crítico correspondiente es 1.44 y se puede ver que todos los t-values en valor absoluto son menores que 1.44. Notar que las variables que quedan en el modelo final son las mismas que en el método de eliminación hacia atrás.

Finalmente, aplicaremos el método “stepwise”, con un nivel de significación del 15% para remover una variable y del 15% para que entre una variable.

Stepwise Regression: grasa versus edad, peso, ...

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is  grasa   on 13 predictors, with N =  252
```

Step	1	2	3	4	5	6	7
Constant	-39.28	-45.95	-27.93	-34.85	-30.65	-25.89	-33.26
abdomen	0.631	0.990	0.975	0.996	1.008	0.945	0.918
T-Value	22.11	17.45	17.37	17.76	17.89	13.82	13.21
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
peso		-0.148	-0.114	-0.136	-0.123	-0.094	-0.119
T-Value		-7.11	-4.84	-5.48	-4.75	-2.98	-3.51
P-Value		0.000	0.000	0.000	0.000	0.003	0.001
muneca			-1.24	-1.51	-1.25	-1.59	-1.53
T-Value			-2.85	-3.40	-2.66	-3.09	-3.00
P-Value			0.005	0.001	0.008	0.002	0.003
antebraz				0.47	0.53	0.57	0.55
T-Value				2.60	2.86	3.08	2.99
P-Value				0.010	0.005	0.002	0.003
cuello					-0.37	-0.40	-0.40
T-Value					-1.65	-1.81	-1.83
P-Value					0.100	0.072	0.068
edad						0.046	0.068
T-Value						1.60	2.21
P-Value						0.110	0.028
muslo							0.22
T-Value							1.91
P-Value							0.057
S	4.88	4.46	4.39	4.34	4.33	4.31	4.29
R-Sq	66.17	71.88	72.77	73.50	73.79	74.06	74.45

```
R-Sq(adj)     66.03    71.65    72.44    73.07    73.26    73.43    73.71
C-p            72.9     20.7     14.2      9.3      8.6      8.0      6.3
```

Notar que el método “stepwise” produjo exactamente los mismos resultados que la selección hacia adelante.

A continuacion se muestra la salida en el R para el método de selección hacia adelante

```
> #llamando a la libreria leaps
> library(leaps)
> # El numero maximo de variables a entrar sera igual al numero de
> # predictoras del conjunto original
> maxvar<-dim(grasa)[2]
> #Aplicando el metodo forward
> freg<-regsubsets(grasa~., data=grasa,method="forward",nvmax=maxvar)
> #Mostrando la salida de todos los pasos con la estadisticas respectiva
> acuforw(freg)
Metodo forward y las estadisticas de cada paso
$which
   (Intercept) edad  peso  altura cuello pecho abdomen cadera muslo rodilla tobillo biceps
1         TRUE FALSE FALSE  FALSE  FALSE FALSE   TRUE   FALSE  FALSE   FALSE   FALSE  FALSE
2         TRUE FALSE  TRUE  FALSE  FALSE FALSE   TRUE   FALSE  FALSE   FALSE   FALSE  FALSE
3         TRUE FALSE  TRUE  FALSE  FALSE FALSE   TRUE   FALSE  FALSE   FALSE   FALSE  FALSE
4         TRUE FALSE  TRUE  FALSE  FALSE FALSE   TRUE   FALSE  FALSE   FALSE   FALSE  FALSE
5         TRUE FALSE  TRUE  FALSE   TRUE FALSE   TRUE   FALSE  FALSE   FALSE   FALSE  FALSE
6         TRUE  TRUE  TRUE  FALSE   TRUE FALSE   TRUE   FALSE  FALSE   FALSE   FALSE  FALSE
7         TRUE  TRUE  TRUE  FALSE   TRUE FALSE   TRUE   FALSE   TRUE   FALSE   FALSE  FALSE
8         TRUE  TRUE  TRUE  FALSE   TRUE FALSE   TRUE    TRUE   TRUE   FALSE   FALSE  FALSE
9         TRUE  TRUE  TRUE  FALSE   TRUE FALSE   TRUE    TRUE   TRUE   FALSE   FALSE   TRUE
10        TRUE  TRUE  TRUE  FALSE   TRUE FALSE   TRUE    TRUE   TRUE   FALSE    TRUE   TRUE
11        TRUE  TRUE  TRUE   TRUE   TRUE FALSE   TRUE    TRUE   TRUE   FALSE    TRUE   TRUE
12        TRUE  TRUE  TRUE   TRUE   TRUE  TRUE   TRUE    TRUE   TRUE   FALSE    TRUE   TRUE
13        TRUE  TRUE  TRUE   TRUE   TRUE  TRUE   TRUE    TRUE   TRUE    TRUE    TRUE   TRUE
   antebrazo muneca
1     FALSE  FALSE
2     FALSE  FALSE
3     FALSE   TRUE
4      TRUE   TRUE
5      TRUE   TRUE
6      TRUE   TRUE
7      TRUE   TRUE
8      TRUE   TRUE
9      TRUE   TRUE
10     TRUE   TRUE
11     TRUE   TRUE
12     TRUE   TRUE
13     TRUE   TRUE

$rss
 [1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849 4455.324 4434.613 4421.330
[11] 4412.655 4411.522 4411.448

$s
 [1] 4.887269 4.464576 4.401902 4.351542 4.336447 4.322660 4.299416 4.290738 4.289625 4.292110
[11] 4.296858 4.305323 4.314360
```

```
$rsq
 [1] 0.6616721 0.7187981 0.7277401 0.7350112 0.7379161 0.7406429 0.7444763 0.7465540 0.7477322
[10] 0.7484878 0.7489813 0.7490458 0.7490500

$adjr2
 [1] 0.6603188 0.7165395 0.7244466 0.7307199 0.7325892 0.7342913 0.7371457 0.7382101 0.7383504
[10] 0.7380516 0.7374763 0.7364456 0.7353426

$cp
 [1] 72.868837 20.690746 14.210205  9.314331  8.559272  7.973197  6.337654  6.367146  7.249744
[10]  8.533156 10.065111 12.003988 14.000000

$bic
 [1] -262.0435 -303.1197 -305.7338 -307.0259 -304.2743 -301.3805 -299.6035 -296.1315 -291.7763
[10] -287.0028 -281.9683 -276.5036 -270.9784
```

Las salidas en S-Plus para el método de eliminación hacia atrás, método de selección hacia delante, y método "stepwise" respectivamente son como sigue:

```
> grasa.y<-grasa[,1]
> grasa.x<-grasa[,2:14]

> breg<-stepwise(grasa.x,grasa.y,method="back")
> breg
$rss:
 [1]  4411.522  4412.655  4421.330  4434.613  4455.324  4491.849  4553.520  4619.874
4658.236  4786.054  4943.245  5947.463 17578.990

$size:
 [1] 12 11 10  9  8  7  6  5  4  3  2  1  0

$which:
      edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
12(- 9)    T    T      T      T     T       T      T     T       F       T      T         T      T
11(- 5)    T    T      T      T     F       T      T     T       F       T      T         T      T
10(- 3)    T    T      F      T     F       T      T     T       F       T      T         T      T
 9(-10)    T    T      F      T     F       T      T     T       F       F      T         T      T
 8(-11)    T    T      F      T     F       T      T     T       F       F      F         T      T
 7(- 7)    T    T      F      T     F       T      F     T       F       F      F         T      T
 6(- 4)    T    T      F      F     F       T      F     T       F       F      F         T      T
 5(- 8)    T    T      F      F     F       T      F     F       F       F      F         T      T
 4(- 1)    F    T      F      F     F       T      F     F       F       F      F         T      T
 3(-12)    F    T      F      F     F       T      F     F       F       F      F         F      T
 2(-13)    F    T      F      F     F       T      F     F       F       F      F         F      F
 1(- 2)    F    F      F      F     F       T      F     F       F       F      F         F      F
 0(- 6)    F    F      F      F     F       F      F     F       F       F      F         F      F

$f.stat:
 [1]  0.003988103  0.061378633  0.471848985  0.723998939  1.130246734  1.992092066
3.350023992  3.570128594  4.042716405
```

```
[10]   6.777492645   8.145201631  50.584224182 488.928083209

$method:
[1] "backward"

> freg<-stepwise(grasa.x,grasa.y,method="forw")
> freg
$rss:
 [1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849 4455.324 4434.613
4421.330 4412.655 4411.522 4411.448

$size:
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13

$which:
      edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
 1(+ 6)    F    F      F      F     F      T      F     F       F       F      F         F      F
 2(+ 2)    F    T      F      F     F      T      F     F       F       F      F         F      F
 3(+13)    F    T      F      F     F      T      F     F       F       F      F         F      T
 4(+12)    F    T      F      F     F      T      F     F       F       F      F         T      T
 5(+ 4)    F    T      F      T     F      T      F     F       F       F      F         T      T
 6(+ 1)    T    T      F      T     F      T      F     F       F       F      F         T      T
 7(+ 8)    T    T      F      T     F      T      F     T       F       F      F         T      T
 8(+ 7)    T    T      F      T     F      T      T     T       F       F      F         T      T
 9(+11)    T    T      F      T     F      T      T     T       F       F      T         T      T
10(+10)    T    T      F      T     F      T      T     T       F       T      T         T      T
11(+ 3)    T    T      T      T     F      T      T     T       F       T      T         T      T
12(+ 5)    T    T      T      T     T      T      T     T       F       T      T         T      T
13(+ 9)    T    T      T      T     T      T      T     T       T       T      T         T      T

$f.stat:
 [1] 488.928083209  50.584224182   8.145201631   6.777492645   2.726691325   2.575843689
3.660481076   1.992092066   1.130246734
[10]   0.723998939   0.471848985   0.061378633   0.003988103

$method:
[1] "forward"

> stepreg<-stepwise(grasa.x,grasa.y,method="efroymson")
> stepreg
$rss:
[1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849

$size:
[1] 1 2 3 4 5 6 7

$which:
      edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
1(+ 6)    F    F      F      F     F      T      F     F       F       F      F         F      F
2(+ 2)    F    T      F      F     F      T      F     F       F       F      F         F      F
3(+13)    F    T      F      F     F      T      F     F       F       F      F         F      T
```

4(+12)	F	T	F	F	F	T	F	F	F	F	F	T	T
5(+ 4)	F	T	F	T	F	T	F	F	F	F	F	T	T
6(+ 1)	T	T	F	T	F	T	F	F	F	F	F	T	T
7(+ 8)	T	T	F	T	F	T	F	T	F	F	F	T	T

$f.stat:
[1] 488.928083 50.584224 8.145202 6.777493 2.726691 2.575844 3.660481

$method:
[1] "efroymson"

Notar que R/S-Plus da toda la secuencia de como todas las variables son removidas en el método "backward" y de como son añadidas en el método "forward" pero no selecciona las mejores variables. Sin embargo en el método "stepwise" (Efroymson) si se reportan las mejores variables.

6.2 Método de los mejores subconjuntos

Para problemas con un número pequeño de variables predictoras (no más de 8), se podrían calcular uno o dos criterios de selección para las 2^k regresiones posibles, luego se escogerían unos cuantos de estos modelos para un análisis más detallado y decidir sobre el mejor modelo.
Lamentablemente hoy en día existen modelos con un gran número de variables predictoras, fácilmente se pueden encontrar problemas con más de 200 variables predictoras y ajustar 2^{200} modelos sería un trabajo computational bien pesado. Basándose en el algoritmo "Branch and Bound" (Ramificación y acotamiento) Hocking and Leslie (1967) propusieron un método para seleccionar solo los mejores subconjuntos de acuerdo a cierto criterio. Más tarde en 1974, Furnival and Wilson, propusieron un algoritmo llamado "Leaps and Bound" (Brincando y acotando) que permite elegir los mejores subconjuntos más eficientemente y este es el algoritmo adoptado por la mayoría de los programas estadísticos de computadoras.

6.3 Criterios para elegir el mejor modelo:

6.3.1 El coeficiente de Determinación R^2

La manera más básica de determinar el mejor modelo es eligiendo aquél que da un R^2 bastante alto con el menor número de variables predictoras posibles. Aparte del efecto de datos anormales que pueden afectar este criterio, hay otro problema pues un modelo con pocas variables siempre tendrá un R^2 menor o igual que un modelo que incluye un mayor número de variables, en consecuencia este criterio tendería a sugerirnos un modelo que contiene una buena cantidad de variables. Como una regla práctica se debería elegir un modelo con k variables si al incluir una variable adicional el R^2 no se incrementa sustancialmente, algo como un 5%, en términos relativos.

6.3.2 El R^2 ajustado

Para subsanar la tendencia del R^2 de elegir como mejor modelo aquel que tiene un gran número de variables predictoras, se ha definido un R^2 ajustado de la siguiente manera:

$$R^2_{ajus} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2) \qquad (1)$$

donde p es el número de parámetros en el modelo. El R^2 ajustado podría disminuir al incluirse una variable adicional en el modelo. Nuevamente, el modelo que se busca es quel que tiene un R^2-ajustado alto con pocas variables.

6.3.3 La varianza estimada del error (s^2).

El mejor modelo será aquel que tenga la varianza estimada (o desviación estándar) del error más pequeña.

6.3.4 El C_p de Mallows.

La idea de este criterio, introducido por Mallows en 1973, es que el mejor modelo es áquel que no tiene ni mucha falta de ajuste ("underfitting") ni mucho sobreajuste ("overfitting") al ajustar los datos. Cuando hay falta de ajuste el estimado del valor predicho de la variable de respuesta tiene mucho sesgo y poca varianza, mientras que cuando hay "overfitting" la varianza del estimado del valor predicho es bastante alta, pero el sesgo es bajo. El cuadrado medio del error para un valor predicho sumando sobre todas las observaciones está dado por

$$\sum_{i=1}^{n}\frac{MSE(\hat{y}(x_i))}{\sigma^2}=\sum_{i=1}^{n}\frac{E[\hat{y}(x_i)-y(x_i)]^2}{\sigma^2}=\sum_{i=1}^{n}\frac{Var(\hat{y}(x_i))+Sesgo^2(\hat{y}(x_i))}{\sigma^2} \qquad (2)$$

Puede ser demostrado que

$$\sum_{i=1}^{n}\frac{Var(\hat{y}(x_i))}{\sigma^2}=p \qquad (3)$$

y que

$$\sum_{i=1}^{n}\frac{Sesgo^2(\hat{y}(x_i))}{\sigma^2}=(n-p)(\frac{E(s_p^2)-\sigma^2}{\sigma^2}) \qquad (4)$$

El criterio de Mallows trata de encontrar un modelo donde tanto el sesgo como la varianza sean moderados. El estimado del lado derecho de la ecuación (2) es llamado el estadístico de Mallows y está dado por

$$C_p=p+(n-p)\frac{s_p^2}{s^2}-(n-p)=\frac{SSE_p}{s^2}-(n-2p) \qquad (5)$$

donde SSE_p es la suma de cuadrados del error del modelo que contiene p parámetros, incluyendo el intercepto, y s^2 es la varianza estimada con el modelo completo. Si un modelo con p parámetros es adecuado entonces $E(SSE_p)=(n-p)\sigma^2$. Luego, $E[SSE_p/s^2]$ es aproximadamente$(n-p)$ $\sigma^2/\sigma^2=(n-p)$. En consecuencia si el modelo fuera adecuado adecuado $E(C_p)=p$. Para decidir acerca del valor de p se acostumbra a plotear C_p versus p. Los valores p más adecuados serán aquellos cercanos a la intersección de la gráfica con la línea $C_p=p$

MINITAB permite seleccionar los mejores subconjuntos basados en los criterios anteriores. Se debe usar la secuencia **STAT ▸ Regression ▸ Best Subsets**.

Ejemplo 2. Elegir los mejores subconjuntos de variables predictoras para el conjunto de datos **grasa** usando los criterios anteriores.

Best Subsets Regression: grasa versus edad, peso, ...

```
Response is grasa

                                                                                  r t   a
                                                                                  o o b n
                                                                  a c   a c       d b i t m
                                                              e p t u p d a m u i l l e b u n
                                                              d e u l c o m e r l l l p a r e c
Vars   R-Sq    R-Sq(adj)        C-p         S                a s r a o n a o a o s z a
   1   66.2         66.0       72.9    4.8775                     X
   1   49.4         49.2      232.2    5.9668                   X
   2   71.9         71.7       20.7    4.4556                X  X
   2   70.2         70.0       36.6    4.5866                   X                    X
   3   72.8         72.4       14.2    4.3930                X  X                    X
   3   72.4         72.0       18.0    4.4251                X  X  X
   4   73.5         73.1        9.3    4.3427                X  X                 X  X
   4   73.3         72.8       11.4    4.3609                X  X              X     X
   5   73.8         73.3        8.6    4.3276                X  X  X           X  X
   5   73.7         73.2        9.2    4.3336             X  X  X              X  X
   6   74.1         73.5        7.7    4.3111             X  X  X     X        X  X
   6   74.1         73.4        8.0    4.3138             X  X  X  X  X        X  X
   7   74.4         73.7        6.3    4.2906             X  X  X  X  X  X     X  X
   7   74.3         73.6        7.4    4.2998             X  X  X  X  X        X  X  X
   8   74.7         73.8        6.4    4.2819             X  X  X  X  X  X     X  X
   8   74.6         73.8        7.0    4.2872             X  X  X  X  X  X  X  X  X  X
   9   74.8         73.8        7.2    4.2808             X  X  X  X  X  X  X  X  X  X
   9   74.7         73.8        7.7    4.2851             X  X  X  X  X  X  X     X  X
  10   74.8         73.8        8.5    4.2832             X  X  X  X  X  X  X  X  X  X  X
  10   74.8         73.8        8.7    4.2850             X  X  X  X  X  X  X     X  X  X
  11   74.9         73.7       10.1    4.2879             X  X  X  X  X  X  X  X  X  X  X
  11   74.8         73.7       10.5    4.2920             X  X  X  X  X  X  X  X  X  X  X
  12   74.9         73.6       12.0    4.2963             X  X  X  X  X  X  X  X  X  X  X
  12   74.9         73.6       12.1    4.2968             X  X  X  X  X  X  X  X  X  X  X
  13   74.9         73.5       14.0    4.3053             X  X  X  X  X  X  X  X  X  X  X
```

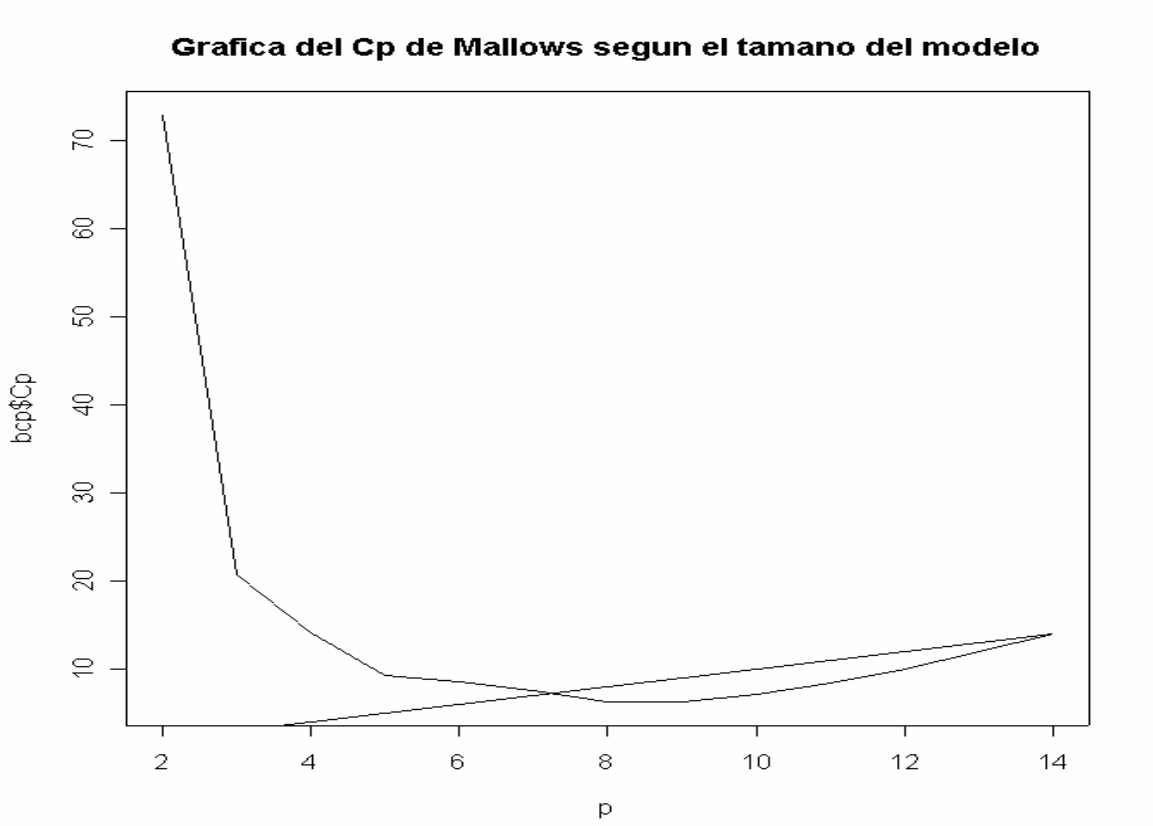
Lo que se muestra aquí son los dos mejores subcojuntos de variables para cada número de variables predictoras, excepto cuando se tiene el modelo que incluye todas las variables. De acuerdo al R^2 y R^2 ajustado el mejor modelo sería aquel que incluye solo dos variables predictoras: peso y abdomen. De acuerdo al C_p de Mallows se escogería el modelo que incluye 6 variables predictoras: Edad, peso, abdomen, muslo, antebrazo y muñeca. El C_p es de 7.7.

De acuerdo a la varianza estimada del error se escogería el modelo que incluye 4 variables predictoras: peso, abdomen, antebrazo y muñeca.

La library **leaps** de R selecciona los mejores subconjuntos usando los criterios de R^2, R^2 ajustado y el C_p de Mallows. Aquí solo mostramos los resultados para el criterio Cp.

```
># El numero maximo de variables a entrar sera igual al numero de
```

```
> # predictoras del conjunto original
> maxvar<-dim(grasa)[2]
> #Mejor modelo usando Cp de mallows
> bcp<-leaps(grasa.x,grasa.y,method="Cp",nbest=1,names=nombres)
> bcp
$which
   edad  peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps
1  FALSE FALSE  FALSE  FALSE FALSE    TRUE  FALSE FALSE   FALSE   FALSE  FALSE
2  FALSE  TRUE  FALSE  FALSE FALSE    TRUE  FALSE FALSE   FALSE   FALSE  FALSE
3  FALSE  TRUE  FALSE  FALSE FALSE    TRUE  FALSE FALSE   FALSE   FALSE  FALSE
4  FALSE  TRUE  FALSE  FALSE FALSE    TRUE  FALSE FALSE   FALSE   FALSE  FALSE
5  FALSE  TRUE  FALSE   TRUE FALSE    TRUE  FALSE FALSE   FALSE   FALSE  FALSE
6   TRUE  TRUE  FALSE  FALSE FALSE    TRUE  FALSE  TRUE   FALSE   FALSE  FALSE
7   TRUE  TRUE  FALSE   TRUE FALSE    TRUE  FALSE  TRUE   FALSE   FALSE  FALSE
8   TRUE  TRUE  FALSE   TRUE FALSE    TRUE   TRUE  TRUE   FALSE   FALSE  FALSE
9   TRUE  TRUE  FALSE   TRUE FALSE    TRUE   TRUE  TRUE   FALSE   FALSE   TRUE
10  TRUE  TRUE  FALSE   TRUE FALSE    TRUE   TRUE  TRUE   FALSE    TRUE   TRUE
11  TRUE  TRUE   TRUE   TRUE FALSE    TRUE   TRUE  TRUE   FALSE    TRUE   TRUE
12  TRUE  TRUE   TRUE   TRUE  TRUE    TRUE   TRUE  TRUE   FALSE    TRUE   TRUE
13  TRUE  TRUE   TRUE   TRUE  TRUE    TRUE   TRUE  TRUE    TRUE    TRUE   TRUE
   antebrazo muneca
1     FALSE  FALSE
2     FALSE  FALSE
3     FALSE   TRUE
4      TRUE   TRUE
5      TRUE   TRUE
6      TRUE   TRUE
7      TRUE   TRUE
8      TRUE   TRUE
9      TRUE   TRUE
10     TRUE   TRUE
11     TRUE   TRUE
12     TRUE   TRUE
13     TRUE   TRUE

$label
 [1] "(Intercept)" "edad"        "peso"        "altura"      "cuello"
 [6] "pecho"       "abdomen"     "cadera"      "muslo"       "rodilla"
[11] "tobillo"     "biceps"      "antebrazo"   "muneca"

$size
 [1]  2  3  4  5  6  7  8  9 10 11 12 13 14

$Cp
 [1] 72.868837 20.690746 14.210205  9.314331  8.559272  7.664855  6.337654
 [8]  6.367146  7.249744  8.533156 10.065111 12.003988 14.000000

> p<-2:maxvar
> plot(p,bcp$Cp,type="l")
> title("Grafica del Cp de Mallows segun el tamano del modelo")
> lines(2:maxvar,2:maxvar)
```



Notar que la curva y la línea se intersectan alrededor de p=7.

6.3.5 PRESS (Suma de cuadrados de Predicción)

El criterio suma de cuadrados de Predicción [PRESS], introducido por Allen en 1974, es una combinación de todas las regresiones posibles, análisis de residuales y "leave-one-out" (ver más adelante validación cruzada).
Supongamos que hay p parámetros en el modelo y que tenemos n observaciones disponibles para estimar los parámetros. En cada paso se deja de lado la i-ésima observacion del conjunto de datos y se calculan todas las regresiones posibles (más eficientemente se podrían calcular solamente los mejores subconjuntos de regresión que resultan de aplicar algún criterio, tal como el C_p de Mallows). Luego se calcula la predicción $\hat{y}_{(i)}$ para la observación que no fue incluida y se calcula el residual correpondiente $e_{(i)} = y_i - \hat{y}_{(i)}$, el cual es llamado el residual PRESS. Ya se vió en la sección 3.1.4 que la relación entre el residual PRESS y el residual usual $\hat{e}_i$ es

$$e_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \qquad (6)$$

donde los h_{ii} representan los elementos de la diagonal de la matriz $\mathbf{H=X(X'X)^{-1}X.}$
Si la diferencia entre el residual PRESS y el residual usual de una observación es bien grande entonces se considera que dicha observación es influencial.
La medida PRESS para el modelo de regresión que contiene p parámetros se define por:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \quad (7)$$

Otra forma, equivalente de cálculo sería

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2 \quad (8)$$

Según el criterio PRESS el mejor modelo será aquel que tenga el valor de PRESS más bajo. MINITAB calcula el PRESS cuando se hace regresión y cuando se hace el "stepwise". No existe una function en R que calcule el PRESS pero ésta puede ser rpogramada fácilmente.

Ejemplo 3. Calcular el PRESS para los mejores modelos según los criterios discutidos anteriormente del conjunto de datos grasa

Usando **Stat ▸ Regression ▸ Regression** y oprimiendo **Options** se marca luego que se desea calcular PRESS.

1)Regression Analysis: grasa versus peso, abdomen

```
The regression equation is
grasa = - 46.0 - 0.148 peso + 0.990 abdomen

Predictor        Coef     SE Coef          T        P
Constant      -45.952       2.605     -17.64    0.000
peso         -0.14800     0.02081      -7.11    0.000
abdomen       0.98950     0.05672      17.45    0.000

S = 4.456              R-Sq = 71.9%         R-Sq(adj) = 71.7%
PRESS = 5109.10        R-Sq(pred) = 70.94%
```

2)Regression Analysis: grasa versus peso, abdomen, antebraz, muneca

```
The regression equation is
grasa = - 34.9 - 0.136 peso + 0.996 abdomen + 0.473 antebraz - 1.51 muneca

Predictor        Coef     SE Coef          T        P
Constant      -34.854       7.245      -4.81    0.000
peso         -0.13563     0.02475      -5.48    0.000
abdomen       0.99575     0.05607      17.76    0.000
antebraz       0.4729      0.1817       2.60    0.010
muneca        -1.5056      0.4427      -3.40    0.001

S = 4.343              R-Sq = 73.5%         R-Sq(adj) = 73.1%
PRESS = 4908.05        R-Sq(pred) = 72.08%
```

3)Regression Analysis: grasa versus peso, abdomen, antebraz, muneca, edad, muslo

```
The regression equation is
grasa = - 38.3 - 0.136 peso + 0.912 abdomen + 0.489 antebraz - 1.78 muneca
           + 0.0629 edad + 0.220 muslo

Predictor        Coef     SE Coef          T        P
Constant      -38.322       8.612      -4.45    0.000
peso         -0.13648     0.03288      -4.15    0.000
abdomen       0.91179     0.06975      13.07    0.000
```

```
antebraz       0.4891      0.1823       2.68    0.008
muneca        -1.7788      0.4947      -3.60    0.000
edad          0.06290     0.03080       2.04    0.042
muslo          0.2202      0.1166       1.89    0.060

S = 4.311              R-Sq = 74.1%         R-Sq(adj) = 73.5%
PRESS = 4877.67        R-Sq(pred) = 72.25%
```

Si buscamos un modelo parsimonioso, sería mejor elegir aquel que incluye 4 variables ya que si bien tiene un PRESS mayor que el de 6 variables, la diferencia no es mucha. Esta selección coincide con la que da el método de eliminación hacia atrás al nivel de significación del 5%.

6.3.6 Validación Cruzada (CV)

Fue introducido por Stone en 1974. La idea aquí es estimar el error de predicción dividiendo al azar el conjunto de datos en varias partes. En cada paso una de las partes se convierte en una muestra de prueba que sirve para validar el modelo y las restantes partes constituyen lo que es llamado una muestra de entrenamiento que sirve para construir el modelo. Por lo general se usan 10 partes y eso es llamado una "10 fold cross-validation" , ó n partes y en ese caso es llamado el método "leave-one-out"(dejar uno afiera). Este ultimo se relaciona bastante con el PRESS. El cálculo del error por validación cruzada usando K partes estará dado por:

$$CV = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i}(y_j - \hat{y}_j^{(-i)})^2}{n}$$

donde $\hat{y}_j^{(-i)}$ representa el valor predicho para la j-ésima observación de la parte N_i usando una línea de regresión que ha sido estimada sin haber usado las observaciones de dicha parte.
La idea es escoger el mejor modelo como aquel que tiene el más error de validación cruzada promedio más pequeño. En el caso de ":leave-one-out" el error de predicción promedio es PRESS/n.
El cálculo de validación cruzada para regresión no está disponible en ninguno de los programas estadísticos usados en este texto. Se debe tiene que escribir un programa para obtenerlo, o usar solamente el método "leave-one-out". Nosotros hemos escrito la función **CV10reg** que estima el error promedio de predicción usando validación cruzada 10.

Ejemplo 4. Aplicar la función CV10reg al conjunto grasa.

```
> #Leyendo el conjunto de datos pero excluyendo los nombres de las columnas
> grasa<-read.table(file="c:/grasa.txt",header=F,skip=1)
> dim(grasa)
[1] 252  14
> #Estimando el error promedio de prediccion con todas las predictoras
> CV10reg(grasa,10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 20.51001 19.91260 20.20943 19.94021 20.48995 20.54480 20.29945 20.40238
 [9] 19.98185 20.60561
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 20.28963
> # Estimando el error promedio de prediccion usando: peso, abdomen
> CV10reg(grasa[,c(1,3,7)],10)
```

Los estimados del error promedio de prediccion en cada repeticion son
 [1] 20.21294 20.15848 20.43202 20.43359 20.31295 20.27680 20.15327 20.26616
 [9] 20.30443 20.24162
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 20.27923
> # Estimando el error promedio de prediccion usando: peso, abdomen, antebrazo y muneca.
> CV10reg(grasa[,c(1,3,7,13,14)],10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 19.26377 19.48586 19.63418 19.66251 19.63294 19.35995 19.40967 19.58761
 [9] 20.51080 19.54337
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 19.60907
> # Estimando el error promedio de prediccion usando: edad, peso, abdomen, muslo, antebrazo y muneca.
>CV10reg(grasa[,c(1:3,7,9,13,14)],10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 19.61473 19.29554 19.23239 19.69176 19.52733 19.36840 19.66784 19.21728
 [9] 19.24110 19.15240
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 19.40088
>
De los resultados previos el criterio CV favorece al modelo que incluye la última variable como el mejor modelo.

6.3.7 AIC

El criterio de información de Akaike (Akaike, 1973), tiene su origen en conceptos de teoría de información y está basado en la minimización de la distancia Kullback-Leibler entre la distribución de la variable de respuesta *y* bajo el modelo reducido y bajo el modelo completo. Se define como,

$$AIC= -2*\text{máximo de la log likelihhod} +2p \qquad (9)$$

Donde p es el número de parámetros dle modelo. En particular para el caso de regresión, asumiendo que la varianza de las y's es estimada por SSE/n, la fórmula anterior se reduce a:

$$AIC=n\log[SSE_p/n]+2p \qquad (10)$$

Existen otras variantes a la formula (10). Un buen modelo es aquel con bajo AIC.
MINITAB no da el AIC, pero si aparece en SAS y S-Plus (aunque la versión que calculan es $AIC=[SSE_p/s^2] +2p$.). Tanto en R como en S-Plus están disponibles la funciones step y stepAIC (de la librería MASS) que calcula el mejor modelo por el método "stepwise" basado en el criterio AIC.
Ejemplo 5. Seleccionar el mejor modelo de regresión para el conjunto grasa usando el criterio AIC usando los métodos "forward" y "backward".

```
># primero hay que hallar la regresion con todas las variables predictoras
> #Hallando el mejor subcojunto usando stepwise y el criterio AIC
> l1<-lm(grasa~.,data=grasa)
> step(l1,scope=~.,direction="backward")
Start:  AIC= 749.36
```

```
 grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
    muslo + rodilla + tobillo + biceps + antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- rodilla    1       0.1 4411.5  747.4
- pecho      1       1.1 4412.5  747.4
- altura     1       9.7 4421.2  747.9
- tobillo    1      11.4 4422.9  748.0
- biceps     1      20.9 4432.3  748.5
<none>                   4411.4  749.4
- cadera     1      37.5 4448.9  749.5
- muslo      1      49.6 4461.0  750.2
- peso       1      50.6 4462.1  750.2
- edad       1      68.3 4479.7  751.2
- cuello     1      76.0 4487.4  751.7
- antebrazo  1      95.5 4507.0  752.8
- muneca     1     170.1 4581.6  756.9
- abdomen    1    2261.0 6672.4  851.6

Step:  AIC= 747.36
 grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
    muslo + tobillo + biceps + antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- pecho      1       1.1 4412.7  745.4
- altura     1       9.7 4421.2  745.9
- tobillo    1      12.1 4423.6  746.1
- biceps     1      20.8 4432.3  746.5
<none>                   4411.5  747.4
- cadera     1      37.4 4448.9  747.5
- peso       1      53.1 4464.6  748.4
- muslo      1      54.9 4466.4  748.5
- edad       1      74.1 4485.6  749.6
- cuello     1      78.4 4490.0  749.8
- antebrazo  1      96.8 4508.3  750.8
- muneca     1     170.5 4582.1  754.9
- abdomen    1    2269.9 6681.4  850.0

Step:  AIC= 745.43
 grasa ~ edad + peso + altura + cuello + abdomen + cadera + muslo +
    tobillo + biceps + antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- altura     1       8.7 4421.3  743.9
- tobillo    1      12.4 4425.1  744.1
- biceps     1      20.1 4432.8  744.6
<none>                   4412.7  745.4
- cadera     1      36.3 4449.0  745.5
- muslo      1      60.1 4472.7  746.8
- peso       1      70.8 4483.5  747.4
- edad       1      73.8 4486.5  747.6
```

```
- cuello     1      79.5 4492.1  747.9
- antebrazo  1      95.6 4508.3  748.8
- muneca     1     170.0 4582.6  753.0
- abdomen    1    2879.4 7292.1  870.0

Step:  AIC= 743.92
 grasa ~ edad + peso + cuello + abdomen + cadera + muslo + tobillo +
    biceps + antebrazo + muneca

           Df Sum of Sq    RSS     AIC
- tobillo    1      13.3 4434.6  742.7
- biceps     1      22.4 4443.7  743.2
- cadera     1      30.4 4451.8  743.6
<none>                   4421.3  743.9
- muslo      1      68.8 4490.1  745.8
- cuello     1      77.1 4498.4  746.3
- edad       1      81.3 4502.6  746.5
- antebrazo  1      98.1 4519.4  747.5
- peso       1     119.6 4540.9  748.6
- muneca     1     181.3 4602.6  752.0
- abdomen    1    3178.5 7599.9  878.4

Step:  AIC= 742.68
 grasa ~ edad + peso + cuello + abdomen + cadera + muslo + biceps +
    antebrazo + muneca

           Df Sum of Sq    RSS     AIC
- biceps     1      20.7 4455.3  741.9
- cadera     1      31.7 4466.4  742.5
<none>                   4434.6  742.7
- muslo      1      72.3 4506.9  744.8
- edad       1      77.6 4512.2  745.1
- cuello     1      87.3 4521.9  745.6
- antebrazo  1      97.4 4532.0  746.2
- peso       1     107.2 4541.8  746.7
- muneca     1     168.0 4602.6  750.0
- abdomen    1    3182.0 7616.7  877.0

Step:  AIC= 741.85
 grasa ~ edad + peso + cuello + abdomen + cadera + muslo + antebrazo +
    muneca

           Df Sum of Sq    RSS     AIC
<none>                   4455.3  741.9
- cadera     1      36.5 4491.8  741.9
- cuello     1      79.1 4534.4  744.3
- edad       1      83.8 4539.1  744.5
- peso       1      93.0 4548.3  745.1
- muslo      1     100.7 4556.0  745.5
- antebrazo  1     140.5 4595.8  747.7
- muneca     1     166.8 4622.2  749.1
```

```
- abdomen    1    3163.0 7618.3  875.0

Call:
lm(formula = grasa ~ edad + peso + cuello + abdomen + cadera +     muslo + antebrazo +
muneca, data = grasa)

Coefficients:
(Intercept)         edad         peso       cuello      abdomen       cadera        muslo
  -22.65637      0.06578     -0.08985     -0.46656      0.94482     -0.19543      0.30239
  antebrazo       muneca
    0.51572     -1.53665

> #Hallando primero la regresion con la variable predictora mas correlacionada V7
> l2=lm(grasa~abdomen,data=grasa)
>
step(l2,scope=~.+edad+peso+altura+cuello+pecho+cadera+muslo+rodilla+tobillo+biceps+antebr
azo+muneca,direction="forward")
Start:  AIC= 800.65
 grasa ~ abdomen

           Df Sum of Sq    RSS    AIC
+ peso       1    1004.2 4943.2  756.0
+ muneca     1     709.2 5238.3  770.6
+ cuello     1     614.5 5332.9  775.2
+ cadera     1     548.2 5399.2  778.3
+ altura     1     458.8 5488.7  782.4
+ rodilla    1     318.7 5628.8  788.8
+ tobillo    1     233.3 5714.1  792.6
+ edad       1     200.9 5746.5  794.0
+ pecho      1     195.5 5752.0  794.2
+ muslo      1     174.6 5772.9  795.1
+ biceps     1     135.3 5812.2  796.8
+ antebrazo  1      54.3 5893.2  800.3
<none>                 5947.5  800.6

Step:  AIC= 756.04
 grasa ~ abdomen + peso

           Df Sum of Sq    RSS    AIC
+ muneca     1     157.2 4786.1  749.9
+ cuello     1      86.9 4856.3  753.6
+ muslo      1      81.4 4861.9  753.9
+ antebrazo  1      66.9 4876.4  754.6
+ biceps     1      63.8 4879.4  754.8
+ altura     1      40.3 4903.0  756.0
<none>                 4943.2  756.0
+ rodilla    1       9.7 4933.5  757.5
+ edad       1       1.9 4941.3  757.9
+ tobillo    1       1.5 4941.7  758.0
+ pecho      1   0.01017 4943.2  758.0
+ cadera     1   0.00529 4943.2  758.0
```

```
Step:  AIC= 749.9
 grasa ~ abdomen + peso + muneca

           Df Sum of Sq    RSS    AIC
+ antebrazo  1     127.8 4658.2  745.1
+ biceps     1      88.7 4697.3  747.2
+ muslo      1      40.5 4745.6  749.8
<none>                   4786.1  749.9
+ cuello     1      25.2 4760.9  750.6
+ altura     1      23.4 4762.6  750.7
+ edad       1      21.2 4764.9  750.8
+ rodilla    1      20.5 4765.5  750.8
+ tobillo    1      15.0 4771.1  751.1
+ cadera     1       9.2 4776.8  751.4
+ pecho      1       1.3 4784.8  751.8

Step:  AIC= 745.07
 grasa ~ abdomen + peso + muneca + antebrazo

         Df Sum of Sq    RSS    AIC
+ cuello   1      51.1 4607.2  744.3
+ edad     1      38.4 4619.9  745.0
<none>                 4658.2  745.1
+ biceps   1      33.9 4624.4  745.2
+ muslo    1      27.2 4631.0  745.6
+ rodilla  1      19.8 4638.4  746.0
+ tobillo  1      18.2 4640.1  746.1
+ altura   1      18.0 4640.2  746.1
+ cadera   1       3.5 4654.7  746.9
+ pecho    1       0.5 4657.7  747.0

Step:  AIC= 744.3
 grasa ~ abdomen + peso + muneca + antebrazo + cuello

         Df Sum of Sq    RSS    AIC
+ edad     1      47.9 4559.2  743.7
+ biceps   1      45.9 4561.2  743.8
<none>                 4607.2  744.3
+ muslo    1      25.1 4582.1  744.9
+ altura   1      18.9 4588.3  745.3
+ cadera   1      11.0 4596.2  745.7
+ tobillo  1      10.7 4596.5  745.7
+ rodilla  1      10.4 4596.8  745.7
+ pecho    1  0.009572 4607.2  746.3

Step:  AIC= 743.66
 grasa ~ abdomen + peso + muneca + antebrazo + cuello + edad

         Df Sum of Sq    RSS    AIC
+ muslo    1      67.4 4491.8  741.9
```

```
+ biceps   1      48.1 4511.1  743.0
<none>               4559.2  743.7
+ altura   1      19.0 4540.3  744.6
+ tobillo  1      14.8 4544.5  744.8
+ rodilla  1       6.6 4552.7  745.3
+ cadera   1       3.2 4556.0  745.5
+ pecho    1       0.8 4558.4  745.6

Step:  AIC= 741.91
 grasa ~ abdomen + peso + muneca + antebrazo + cuello + edad +
    muslo

         Df Sum of Sq    RSS    AIC
+ cadera   1      36.5 4455.3  741.9
<none>               4491.8  741.9
+ biceps   1      25.5 4466.4  742.5
+ tobillo  1      12.8 4479.1  743.2
+ altura   1       4.3 4487.5  743.7
+ pecho    1       0.8 4491.1  743.9
+ rodilla  1  0.002584 4491.8  743.9

Step:  AIC= 741.85
 grasa ~ abdomen + peso + muneca + antebrazo + cuello + edad +
    muslo + cadera

         Df Sum of Sq    RSS    AIC
<none>               4455.3  741.9
+ biceps   1      20.7 4434.6  742.7
+ altura   1      11.7 4443.6  743.2
+ tobillo  1      11.6 4443.7  743.2
+ rodilla  1 3.651e-02 4455.3  743.8
+ pecho    1 9.904e-05 4455.3  743.9

Call:
lm(formula = grasa ~ abdomen + peso + muneca + antebrazo + cuello +      edad + muslo +
cadera, data = grasa)

Coefficients:
(Intercept)      abdomen         peso       muneca     antebrazo       cuello         edad
  -22.65637      0.94482     -0.08985     -1.53665      0.51572     -0.46656      0.06578
      muslo      cadera
    0.30239     -0.19543
```

6.3.8 BIC

Fue introducido por Schwarz en 1978 y está basado en argumentos bayesianos. Se define por

$$BIC=nlog[SSE_p/n]+2plog(n)$$

Los criterios AIC y C_p de Mallows tienden a dar modelos óptimos más grandes que el criterio BIC. MINITAB no da el BIC, pero si aparece en SAS (ldonde es llamado SBC) y S-Plus (da una versión modificada).

6.3.9 Validación Cruzada Generalizada (CGV)

Fue introducido en 1979 por Golub, Heath and Whaba El cálculo de validación cruzada "leave-one out" es computacionalmente pesado y el GCV es una aproximación al "leave-one-out", que puede ser calculado más rápidamente.

Se define por

$$GCV = \frac{SSE_p}{[n-1-tr(H_p)]^2}$$

donde H_p es la matriz HAT para el modelo que incluye p variables. El modelo óptimo será aquel que incluye las p variables predictoras que hacen que GCV sea mínimo.

SAS da esta medida pero en Regresión Noparamétrica. El cálculo del GCV puede ser fácilmente programable en S-Plus o Matlab.

Ejemplo 6. A continuación se muestran los 15 mejores modelos para el conjunto de datos **grasa**, ordenados de acuerdo al Cp de Mallows, mostrando además los valores del AIC, BIC y SBC. Los resultados fueron obtenidos usando SAS version 8.

Number in Model	C(p)	R-Square	AIC	BIC	SBC	Variables in Model
**7	**6.3377**	**0.7445**	**741.9088**	**744.5436**	**770.14425**	**edad peso cuello abdomen muslo antebraz muneca**
8	6.3671	0.7466	741.8514	744.7178	773.61622	edad peso cuello abdomen cadera muslo antebraz muneca
8	6.9626	0.7459	742.4748	745.2951	774.23968	edad peso cuello abdomen muslo biceps antebraz muneca
9	7.2497	0.7477	742.6772	745.7373	777.97144	edad peso cuello abdomen cadera muslo biceps antebraz muneca
7	7.3761	0.7434	742.9864	745.5508	771.22181	edad peso cuello abdomen biceps antebraz muneca
8	7.6488	0.7452	743.1915	745.9589	774.95637	edad peso cuello abdomen muslo tobillo antebraz muneca
*6	**7.6649**	**0.7410**	**743.3451**	**745.7048**	**768.05114**	**edad peso abdomen muslo antebraz muneca**
9	7.7333	0.7472	743.1859	746.2041	778.48021	edad peso altura cuello abdomen cadera muslo antebraz muneca
9	7.7402	0.7472	743.1933	746.2108	778.48755	edad peso cuello abdomen cadera muslo tobillo antebraz muneca
6	7.9732	0.7406	743.6612	746.0031	768.36724	edad peso cuello abdomen antebraz muneca

6	8.0815	0.7405	743.7721	746.1077	768.47815	peso cuello abdomen biceps antebraz muneca
8	8.1043	0.7447	743.6660	746.3984	775.43091	edad peso altura cuello abdomen muslo antebraz muneca
9	8.1748	0.7468	743.6497	746.6295	778.94397	edad peso cuello abdomen muslo tobillo biceps antebraz muneca
8	8.2969	0.7445	743.8664	746.5841	775.63130	edad peso cuello pecho abdomen muslo antebraz muneca
8	8.3375	0.7445	743.9087	746.6232	775.67353	edad peso cuello abdomen muslo rodilla antebraz muneca

(*) sería el mejor modelo con el C_p de Mallows, porque aún cuando no es el modelo con el menor C_p (es un error bastante común elegir como el mejor modelo aquel con el menor C_p) su valor 7.66 está cerca a 6+1=7 (número de parámteros del modelo). También sería el mejor modelo si se usa el criterio SBC, el criterio Bayesiano de Schwartz. Algunos autores llaman a este el BIC.

(**) sería el mejor modelo de acuerdo al AIC, porque aunque es el modelo con el segundo AIC más pequeño, se está eligiendo solo 7 variables. También sería el mejor modelo si se usa el criterio BIC de Sawa.

6.3.10 Otros Criterios

Recientemente se han introducido muchos otros criterios para selección de variables en regresión entre los más conocidos están:

MDL: Longitud de descripción Mínima (Rissanen, 1978).
RIC: Criterio de Inflación del Riesgo (Foster y George, 1994)
CIC: Criterio de Inflación del Covarianza (Tibshirani and Knigth, 1999)
Bootstrapping (Efron, 1983)
El pequeño Boostrapping (Breiman, 1992)
La Garrote (Breiman, 1995)
El Lasso (Tibshirani, 1996)

Para más detalles acerca estos métodos veáse el texto “Subset selection in regression” por Alan Miller (2002).

6.3.11 Recomendaciones para elegir el mejor modelo

En cualquier problema las variables predictoras pueden ser clasificadas en 3 grupos:
a) Las que son importantes.
b) Las que uno no está seguro de su importancia.
c) Las que no son relevantes para explicar el comportamiento de la variable de respuesta.

Lo que se recomienda es eliminar las variables tipo c) eligiendo un buen subconjunto de variables predictoras usando para ello los criterios C_p, AIC o BIC y luego aplicar “stepwise” para descartar las variables tipo b) y quedarnos con las variables tipo a) que son las que nos interesantes. Aplicando esta metodología a nuestro conjunto de datos se obtiene:

Stepwise Regression: grasa versus edad, peso, abdomen, muslo, antebraz

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is  grasa   on  5 predictors, with N =  252


    Step          1        2        3        4
Constant     -39.28   -45.95   -52.96   -59.83

abdomen       0.631    0.990    0.992    1.008
T-Value       22.11    17.45    17.60    17.73
P-Value       0.000    0.000    0.000    0.000

peso                  -0.148   -0.183   -0.200
T-Value                -7.11    -6.82    -7.03
P-Value                0.000    0.000    0.000

muslo                           0.22     0.21
T-Value                         2.04     1.96
P-Value                        0.043    0.051

antebraz                                 0.32
T-Value                                  1.76
P-Value                                 0.080

S             4.88     4.46     4.43     4.41
R-Sq         66.17    71.88    72.34    72.68
R-Sq(adj)    66.03    71.65    72.01    72.24
C-p           57.2      7.7      5.5      4.4
```

Solo quedan: abdomen, peso, muslo y antebrazo como las variables predictoras más importantes.

6.4 Otros métodos de Selección de variables

Existen mucho otros métodos de selección de variables en regresión, solo menionaremos dos de ellos

6.4.1 Métodos Bayesianos

Es considerado en gran detalle por Mitchel y Beauchamp (JASA, 1988). Supongamos que ya se tiene un conjunto de buenos modelos. La idea se basa en asignar probabilidades a priori a los coeficientes de cada uno de estos modelos que incluyen solo un subconjunto de predictoras e igualmente se asignan probabilidades a priori a cada uno de los modelos. Finalmente se elige como mejor modelo aquel que tiene la probabilidad posterior más alta con respecto a la variable de respuesta.

6.4.2.Algoritmo Genéticos: En este caso el problema de selección de variables es considerado como un problema de optimización con respecto al número de variables predictoras que deben incluirse en el modelo. Luego el problema de optimización es resuelto usando algoritmos Genéticos.

Ejercicios

1. Supongamos que se desea omitir la variable predictora Xj de un modelo de regresión con p parámetros (incluyendo el intercepto) y n observaciones. Si F_j es el estadístico para probar la hipótesis H: βj=0 demostrar que:

$$C_{p-1} = \frac{F_j SSE_p}{s^2(n-p)} + C_p - 2$$

2. Hacer selección de variables predictoras usando el conjunto de datos
Berkeley: La variable de respuesta es SOMA y las predictoras son WT2, HT2, WT9, HT9, LG9, ST9. Disponible en la página de internet del texto.

a) Los metodos "stepwise". Explicar los pasos de los procesos y justificar la terminación del mismo.
b) Los mejores usando los mejores subcojuntos con por lo menos 6 criterios. Explicar los resultados
c) Comparar los resultados obtenidos en a y b. Dar su seleccion final.

3. Hacer selección de variables predictoras usando el conjunto de datos

Highway: La variable de respuesta es TASA y todas las otras son predictoras

a) Los metodos "stepwise". Explicar los pasos de los procesos y justificar la terminación del mismo.
b) Los mejores usando los mejores subcojuntos con por lo menos 6 criterios. Explicar los resultados
c) Comparar los resultados obtenidos en a y b. Dar su selección final.

4. Verificar las ecuaciones (3) y (4) de las sección 6.3.4 del texto.
5. Investigar la relación entre los criterios AIC, BIC, Cp de Mallows y R^2.

CAPÍTULO 7

MULTICOLINEALIDAD

7.1 Multicolinealidad.

Dos predictoras X_1 y X_2 son exactamente colineales si existe una relación lineal tal que $C_1X_1+C_2X_2=C_0$ para algunas constantes C_1, C_2 y C_0. Si la ecuación se cumple aproximadamente para los datos observados entonces se dice que hay colinealidad aproximada. Una medida comúnmente usada para detectar colinealidad es el coeficiente de determinación. Se dice que X_1 y X_2 son colineales si R_{12}^2 es bastante cercano a 1 (ó 100%). Sin embargo, cuando existen "outliers" esta medida no es completamente adecuada.
La definición se extiende al caso cuando hay más de dos variables predictoras. Un conjunto de predictoras X_1, X_2,….X_p son colineales si para constantes c_o,c_1,…..c_p, la ecuación

$$c_1X_1+c_2X_2+\ldots\ldots c_pX_p=c_o \qquad (1)$$

se cumple aproximadamente. De la ecuación anterior se desprende que cuando hay multicolinealidad una de las predictoras puede ser determinada de las otras. Es decir,

$$X_k = (c_o - \sum_{j\neq k} c_j X_j)/c_k \qquad (2)$$

si el coeficiente de determinación R_k^2 de la regresión de X_k con las otras variables predictoras es cercano a 1 se puede concluir tentativamente que hay multicolinealidad.

7.1.1 Efectos de multicolinealidad

Si consideramos el modelo de regresión lineal múltiple

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots.\beta_p X_p + e$$

entonces se puede mostrar que la varianza del j-ésimo coeficiente de regresión estimado es

$$\text{var}(\hat{\beta}_j) = \sigma^2 (\frac{1}{1-R_j^2})(\frac{1}{S_{X_jX_j}}) \qquad (3)$$

donde R_j^2 es el coeficiente de Determinación de la regresión lineal de X_j contra todas las demas predictoras. La cantidad $\frac{1}{1-R_j^2}$ es llamado el j-ésimo **Factor de inflación de la varianza**, or VIF_j (Marquardt, 1970). Si R_j^2 es cercano a 1 entonces la varianza de $\hat{\beta}_j$ aumentará grandemente. El VIF representa el incremento en la varianza debido a la presencia de multicolinealidad. Una variable predictora con un VIF mayor de 10 (esto es equivalente a un R^2=.90), puede causar multicolinealidad. La mayoría de los programas estadísticos da los valores

VIF. Los VIF son los elementos que están en la diagonal de la matriz C^{-1}, que es la inversa de la matriz de correlaciones C .

Ejemplo 1. Usar R para calcular los VIF para los datos del conjunto **millaje**.

```
> # Hallando la matriz de correlaciones de las variables predoctoras
> mcor<-cor(millaje[,2:5])
> mcor
             sp        wt         vol         hp
sp   1.00000000 0.6785339 -0.04306242 0.96654517
wt   0.67853388 1.0000000  0.38495423 0.83222021
vol -0.04306242 0.3849542  1.00000000 0.07647905
hp   0.96654517 0.8322202  0.07647905 1.00000000
>
```

```
># Hallando la inversa de la matriz de correlaciones
> invcorr=solve(mcor)
> invcorr
            sp         wt       vol        hp
sp   71.675777  30.997855 -1.584250 -94.95376
wt   30.997855  18.318201 -2.272839 -45.03178
vol  -1.584250  -2.272839  1.554038   3.30390
hp  -94.953756 -45.031780  3.303900 130.00077
>
```

Los elementos de la diagonal de la inversa de la matriz de correlaciones son los VIF's.

```
> #Hallando los VIFs
> vif<-diag(solve(mcor))
> cat("los VIF's son:\n")
los VIF's son:
> vif
        sp         wt        vol         hp
 71.675777  18.318201   1.554038 130.000772
```

Las variables HP y SP tiene un VIF bastante alto.

Eliminando la variable HP y recalculando los VIF's se obtiene

```
> #recalculando los VIF's sin la variable hp
> diag(solve(cor(millaje[,2:4])))
      sp       wt      vol
2.320683 2.719362 1.470071
```

Con lo cual parece resolverse el problema de multicolinealidad.

Comparando la regresión considerando todas las predictoras con la regresión que excluye la predictora hp se obtiene

```
> l1=lm(mpg~.,data=millaje)
> summary(l1)

Call:
lm(formula = mpg ~ ., data = millaje)

Residuals:
   Min     1Q  Median     3Q    Max
```

```
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
sp          -1.29482    0.24477  -5.290 1.11e-06 ***
wt          -1.85980    0.21336  -8.717 4.22e-13 ***
vol         -0.01565    0.02283  -0.685    0.495
hp           0.39221    0.08141   4.818 7.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733,     Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16

> l2=lm(mpg~sp+wt+vol,data=millaje)
> summary(l2)

Call:
lm(formula = mpg ~ sp + wt + vol, data = millaje)

Residuals:
    Min      1Q  Median      3Q     Max
-10.0003  -2.7013  -0.5674   1.2842  16.7766

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.18296    5.12341  15.845  < 2e-16 ***
sp          -0.13484    0.04992  -2.701  0.00847 **
wt          -0.91127    0.09318  -9.780 3.35e-15 ***
vol         -0.04121    0.02516  -1.638  0.10554
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.14 on 78 degrees of freedom
Multiple R-Squared: 0.8351,     Adjusted R-squared: 0.8287
F-statistic: 131.7 on 3 and 78 DF,  p-value: < 2.2e-16
```

Notar el gran cambio en los coeficientes de regresión y en la desviación estándar de los mismos. La multicolinealidad también afecta al valor predicho y a la varianza de las predicciones.

7.1.2 Diagnósticos de Multicolinealidad.

De acuerdo a Besley, et al. (1991) se pueden seguir los siguientes pasos para detectar multicolinealidad

1) Cotejar si hay coeficientes de regresión con valores bien grandes o de signo opuesto a lo que se esperaba que ocurriera.
2) Cotejar si las variables predictoras que se esperaban sean importantes tienen valores de t pequeños para las hipótesis de sus coeficientes.

3) Cotejar si la eliminación de una fila o columna de la matriz X produce grandes cambios en el modelo ajustado.
4) Cotejar las correlaciones entre todas las parejas de variables predictoras para detectar las que son bastante altas.
5) Examinar el VIF. Si el VIF es grande, mayor que 10, entonces puede haber multicolinealidad.
6) Usar el número condición de la matriz correlación X*'X*, la cual es de la forma

$$\begin{bmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & r_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1} & r_{p2} & . & . & 1 \end{bmatrix}$$

donde r_{ij} representa la correlación entre las variables X_i y X_j.. La matriz X* es obtenida restando a cada columna de X la media correspondiente y dividiendo luego entre la raíz de la suma de cuadrados corregida por la media de la misma columna.

Sea U una matriz tal que Z=XU y que Z'Z=U'X'XU=D donde D es una matriz diagonal con elementos positivos $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$. Los λ's son llamados los eigenvalues (valores propios de X'X y las columnas de U son los eigenvectors (vectores propios) de X'X. Se puede mostrar que U es ortogonal, es decir U'U=UU'=I. Las columnas de Z=XU son llamados **componentes principales** (ver sección 7.3). Algunos autores prefieren usar los "eigenvalues" de la matriz X'X, que se obtiene cuando cada columna de X es centrada solamente (matriz de covarianza), o cuando cada columna de X tiene norma 1, o sea se ha dividido cada columna de la matriz por la suma de sus cuadrados (matriz de correlación).
La presencia de "eigenvalues" cerca de 0 es también una indicación de multicolinealidad.

El número condición de la matriz X* está definido por

$$K=(\text{mayor "eigenvalue"} / \text{menor "eigenvalue"})^{1/2} \qquad (4)$$

Los "eigenvalues" que se usan son de la matriz de correlación. Notar que K≥1. Weisberg sugiere que un K>30 indica presencia de multicolinealidad.

```
> # Hallando los valores y vectores propios
> ev<-eigen(cor(millaje[,2:5]))
> ev
$values
[1] 2.689421048 1.100439312 0.205497158 0.004642483

$vectors
          [,1]       [,2]       [,3]       [,4]
[1,] -0.5616362  0.2921753  0.5229999  0.57067460
[2,] -0.5526927 -0.2365918 -0.7530524  0.26733338
[3,] -0.1387771 -0.9134537  0.3820731 -0.01906127
```

```
[4,] -0.5998637  0.1557567  0.1157728 -0.77620876

> evals<-ev$values
> evals
[1] 2.689421048 1.100439312 0.205497158 0.004642483
> #Hallando el numero condicion
> cond<-sqrt(evals[1]/ev$values[4])
> cond
[1] 24.06879
```

O sea que hay algo de multicolinealidad presente.

Otras alternativas para detectar multicolinealidad son:

a) Usar los ***índices de condición,*** definidos por $(\frac{\lambda_{\max}}{\lambda_j})^{1/2}$. Un indice condición grande indica presencia de multicolinealidad.
b) Descomponer la varianza de cada coeficiente en proporciones debido a cada una de las otras variables. Una proporción grande indica que dicha variable está produciendo la multicolinealidad.

R no da estos diagnósticos, pero SAS si los muestra.

Ejemplo 2. Usar SAS para hallar los diagnósticos de multicolinealidad para el conjunto de datos **millaje.**

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	192.43775	23.53161	8.18	<.0001	0
VOL	1	-0.01565	0.02283	-0.69	0.4951	1.55404
HP	1	0.39221	0.08141	4.82	<.0001	130.00077
SP	1	-1.29482	0.24477	-5.29	<.0001	71.67578
WT	1	-1.85980	0.21336	-8.72	<.0001	18.31820

Collinearity Diagnostics(intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation VOL	HP	SP	WT
1	2.68942	1.00000	0.00461	0.00103	0.00164	0.00620
2	1.10044	1.56331	0.48792	0.00016958	0.00108	0.00278
3	0.20550	3.61765	0.45711	0.00050172	0.01857	0.15065
4	0.00464	24.06879	0.05036	0.99830	0.97871	0.84038

Aquí se ha usado la matriz de correlación. Hay que darle a SAS la opción COLLINNOINT, de lo contrario salen otros resultados. Notar que la variable HP es la que da más problemas de multicolinealidad.

7.1.3 Medidas remediales al problema de multicolinealidad

Básicamente hay tres propuestas:

a) Regresión Ridge (Hoerl and Kennard, 1970)
b) Componentes principales (Hotelling, 1965)
c) Mínimos Cuadrados Parciales (H. Wold, 1975)

Sin embargo el problema de multicolinealidad también está relacionado con el proceso de selección de variables y esto puede ser considerado como una cuarta manera de resolver el problema de multicolinealidad. En las próximas dos secciones se discutirá regresión ridge y componentes principales. Mínimos cuadrados parciales, que es una técnica muy usada cuando hay pocas observaciones y muchas variables no será considerado en este texto.

7.2 Regresión Ridge

Consideremos la suma de las varianzas de los coeficientes estimados $\hat{\beta}$, dada por E($\hat{\beta}$ -β)'($\hat{\beta}$ -β). Hoerl and Kennard (1970) mostraron que

$$E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)=E[e'X^*(X^{*\prime}X^*)^{-2}X^{*\prime}e]=\sigma^2 Traza(X^{*\prime}X^*)^{-1}=\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i} \qquad (5)$$

Notar que si un valor propio (eigenvalue) es cercano a cero la suma de las varianzas se hace muy grande. Por otro lado, de la ecuación (5) se puede establecer que

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i} \qquad (6)$$

De donde,

$$E(\hat{\beta}'\hat{\beta}) \geq \beta'\beta + \sigma^2\frac{1}{\lambda_p} \qquad (7)$$

Es decir, que aún cuando $\hat{\beta}$ es insesgado, se tiene que $\| \hat{\beta} \|^2 = \hat{\beta}'\hat{\beta} = \sum_{i=1}^{p}\hat{\beta}_i$ es un estimador sesgado.

La idea en regresión Ridge es encontrar un estimador $\tilde{\beta}$ que aunque sea sesgado tenga menor longitud que $\hat{\beta}$. Es decir, $\tilde{\beta}$' $\tilde{\beta} < \hat{\beta}$' $\hat{\beta}$, que significa que el estimador mínimo cuadrático será encogido hacia el origen.

Hoerl y Kennard, en 1970, propusieron el siguiente estimador:

$$\tilde{\beta} = (X'X + kI)^{-1}X'Y \qquad (8)$$

donde el parámetro de encogimiento k (por lo general, 0<k<1) debe ser estimado de los datos tomados. Si k=0 se obtiene el estimador minimo cuadratico y a medida que k aumenta el estimador se aleja del estimador minimo cuadrático y se hace mas sesgado.

Se puede mostrar que el estimador ridge se obtiene al resolver

$$\text{Min}_{\mathbf{B}}\ (\mathbf{y-XB})'(\mathbf{y-XB})$$

$$\text{Sujeto a que}\quad ||B||^2<k^2 \qquad (9)$$

Cuando la restricción (9) se sustituye por |B|<k se obtiene el estimador ***Lasso*** (Tibshirani, 1996).

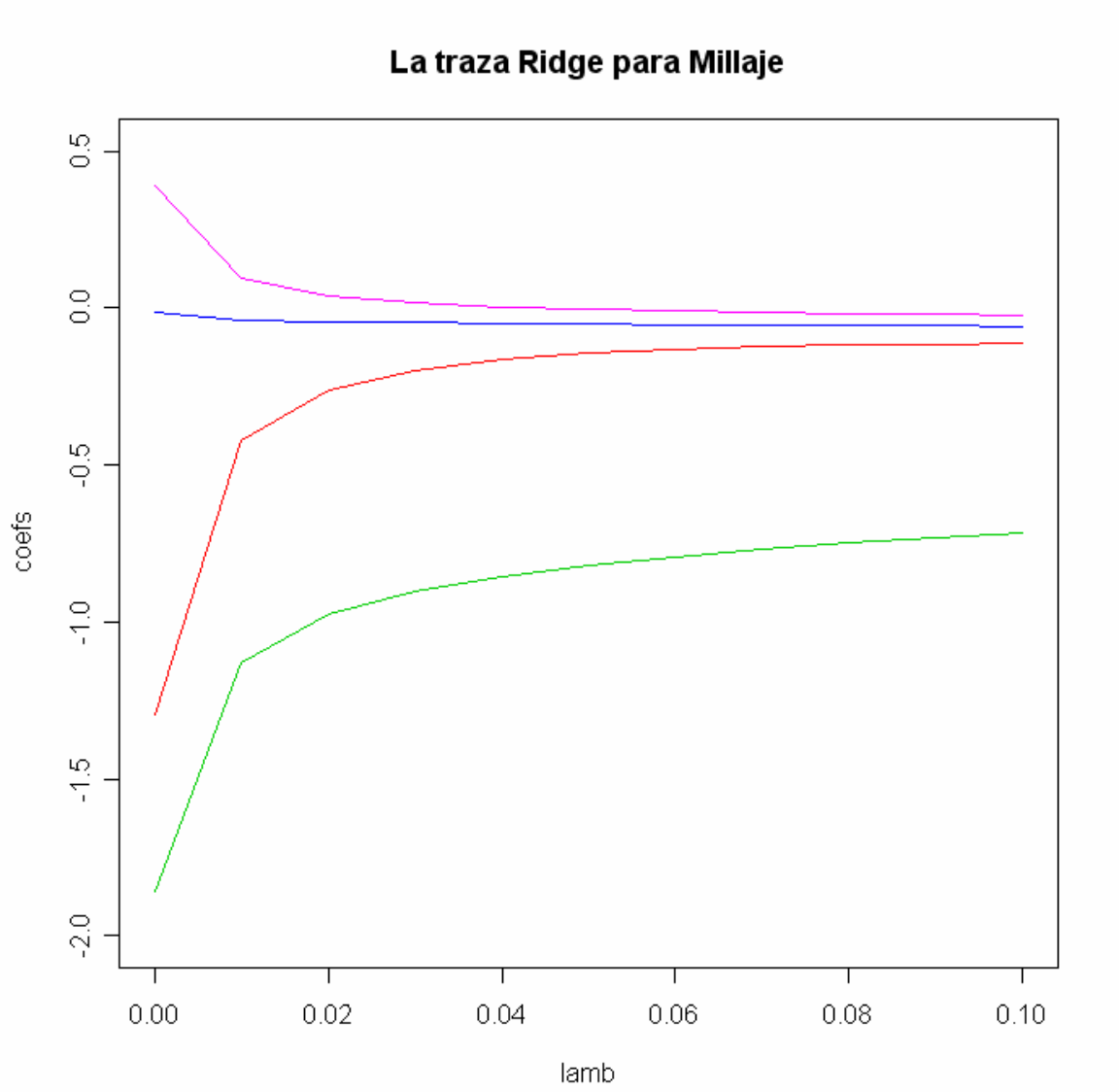
La fórmula anterior se puede usar con las variables predictoras y/o de respuesta en su forma original o en su forma estandarizadas Sin embargo, desde el punto de vista computacional se recomienda trabajar con variables estandarizadas. Existen diversos tipos de estandarizaciones, el que más se usa es centrar los datos (restando por la media) y luego rescalarlos (dividiendo entre la desviación estándar o la raíz cuadrada de las sumas de cuadrados). Tambien existen variantes de la formulación original de la regresión ridge, uno de ellos es considerar el parámetro k como un vector o una matriz, a estos métodos se le llama regresión ridge generalizada.
Hay varias propuestas acerca de la elección de k, pero lo que más se recomienda consiste en hacer un plot de los coeficientes del modelo para varios valores de k (generalmente entre 0 y 1) este plot es llamado la **Traza Ridge** . Para elegir k hay que considerar los siguientes aspectos

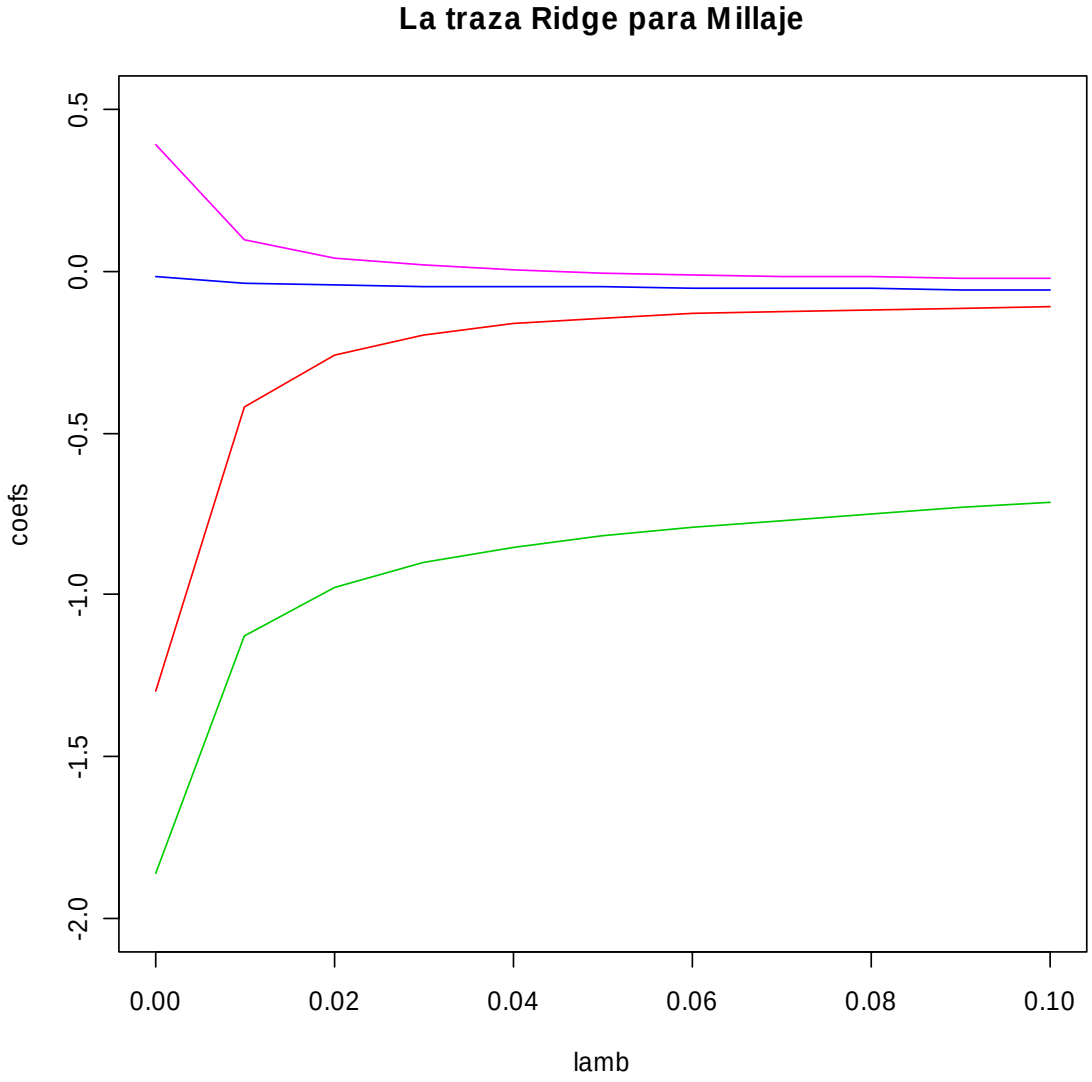
a) Que los valores de los coeficientes de regresión se estabilizen.
b) Que los coefcientes de regresión que tenían un valor demasiado grande comienzen a tener valores razonables.
c) Que los coeficientes de regresión que inicialmente tenían el signo equivocado cambien de signo.

MINITAB no hace regresión Ridge, pero se puede preparar un macro donde se calcule $\tilde{\beta}$ según la fórmula dada para varios valores de k. y luego se plotea la traza Ridge. Para esto hay que usar las operaciones matriciales en MINITAB que aparecen en el menu CALC> MATRICES. SAS si produce los coeficientes de la regresión ridge.
En R la librería MASS tiene la función **lm.ridge** que ejecuta la regresión ridge.aunque con otro tipo de estandarizacion. Nosotros hemos construido una función **acunaridge** que da los resultados iguales a los producdos por la mayoria de los otros programas estadísticos.
Para los datos de **millaje** se obtienen los siguientes resultados.

```
> rr1<-acunaridge(mpg~.,data=millaje,lambda=seq(0,0.1,.01))
          [,1]
[1,] -163.587914
[2,] -136.273017
[3,]   -3.121126
[4,]  200.643157
[1] 13.34262
> rr1
$coef
              0.00        0.01        0.02        0.03        0.04
Intercept 192.43775332 108.39612046 92.87359971 86.49190066 83.081558954
sp         -1.29481848  -0.42131888 -0.26186771 -0.19761416 -0.164264695
wt         -1.85980373  -1.12799710 -0.97515848 -0.90114610 -0.853794192
```

```
vol        -0.01564501  -0.03686107 -0.04266515 -0.04618844 -0.048824946
hp          0.39221231   0.09614332  0.04019377  0.01644930  0.003284967
                0.05         0.06         0.07         0.08         0.09
Intercept 80.992915643 79.59954428 78.61244211 77.88043814 77.31718428
sp         -0.144631624 -0.13219090 -0.12393456 -0.11829065 -0.11436237
wt         -0.819022227 -0.79139270 -0.76833247 -0.74844834 -0.73090915
vol        -0.050966877 -0.05277652 -0.05433824 -0.05570306 -0.05690530
hp         -0.005096282 -0.01090589 -0.01517129 -0.01843518 -0.02101149
                0.10
Intercept 76.87004259
sp         -0.11160387
wt         -0.71518116
vol        -0.05796981
hp         -0.02309457

$lambda
 [1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10

$kHKB
[1] 0.0006234958

> #Haciendo la Traza Ridge
> matridge<-cbind(rr1$lambda,t(rr1$coef[-1,]))
> win.graph()
> plot(matridge[,1],matridge[,2],ylim=c(-
2,0.5),type="l",xlab="lamb",ylab="coefs",col=2)
> title("La traza Ridge para Millaje")
> lines(matridge[,1],matridge[,3],col=3)
> lines(matridge[,1],matridge[,4],col=4)
> lines(matridge[,1],matridge[,5],col=6)
```


La traza Ridge para Millaje
coefs
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0.00
0.02
0.04
0.06
0.08
0.10
lamb



En este plot se nota que el k-óptimo parece ser menor que 0.05. Detallando más el plot hemos llegado a establecer un k-óptimo de alrededor de 0.015.Un segundo método de elegir el parámetro k propuesto por Hoerl, Kennard y Baldwin (1975) está basado en procedimientos bayesianos. En este caso, el k óptimo es un estimado de la razón entre la varianza poblacional σ^2 y la varianza del estimador ridge. Más especificamente

$$k_{opt} = \frac{ps^2}{\sum_{i=1}^{p} b_i^{*2}(0)} \qquad (9)$$

Donde p es el número de variables predictoras, s^2 es la estimación de la varianza de los errores del modelo de mínimos cuadrados trabajando con las variables originales y sin usar ningún tipo de estandarización. Finalmente, $b_i^{*2}(0)$, es el cuadrado del i-ésmo coeficiente de la regresión por mínimos cuadrados pero donde se ha usado la siguiente estandarización:

a) La variable de respuesta Y ha sido centrada, es decir se le ha restado la media a cada dato.

b) Las variables predictoras han sido centradas y escaladas. Es decir a cada dato se le ha restado la media de la variable y dividido entre la raíz cuadrada de la suma de cuadrados con respecto a la media de cada variable. Es decir, para cada variable predictora x se ha aplicado la siguiente transformación.

$$z = \frac{x - \bar{x}}{\sqrt{Sxx}} \qquad (10)$$

En el ejemplo anterior s^2=13.34, p=4 y $\sum_{i=0}^{p} b_i^{*2}(0) = 85598.8$ dando k_{opt}=0.000623.

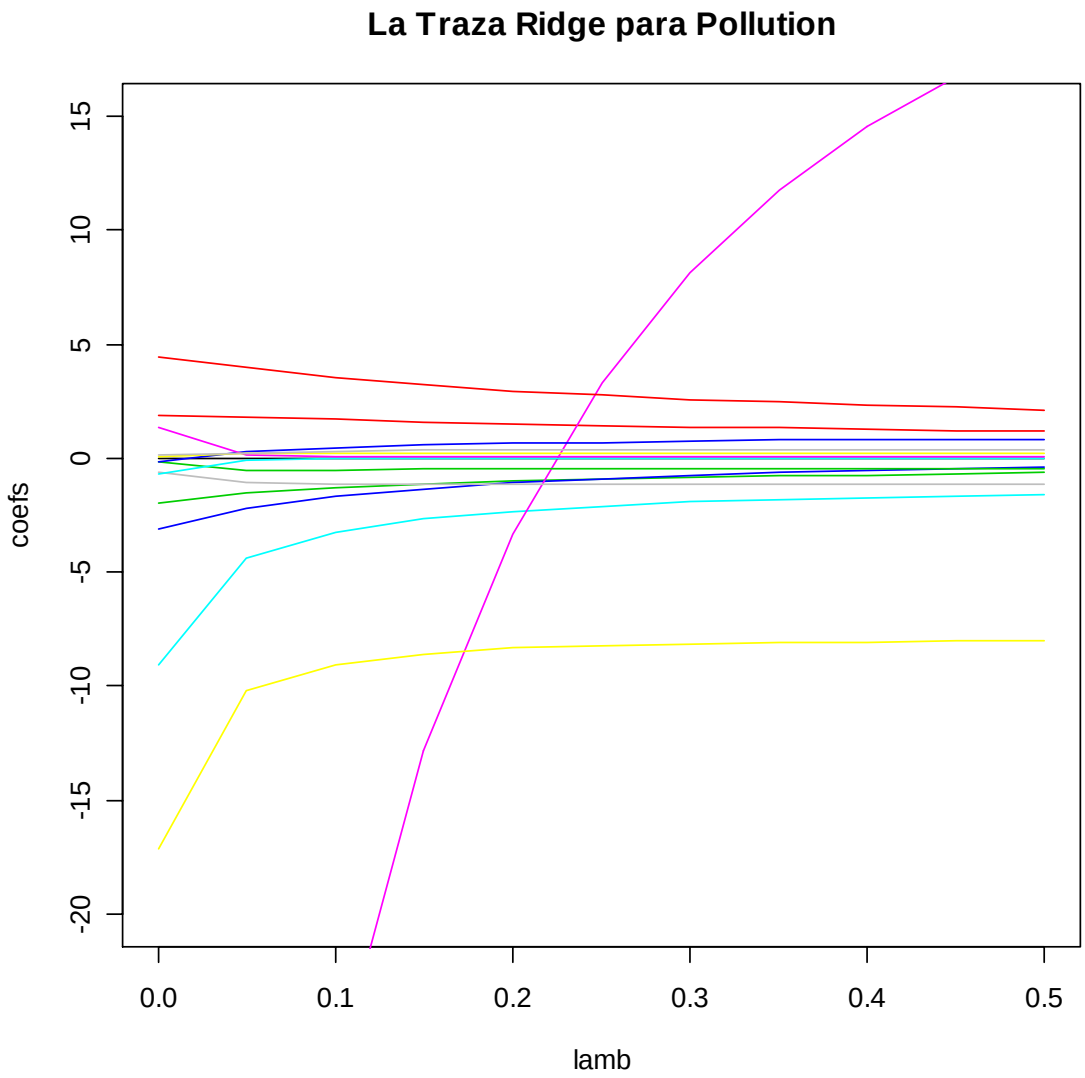
Ejemplo 3. El conjunto de datos **Pollution** es uno de los más usados en los trabajos de ridge regression. Aplicar el método anterior para elegir el k-óptimo.

Fuente: McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

Variables en orden:

PREC Average annual precipitation in inches
JANT Average January temperature in degrees F
JULT Same for July
OVR65 % of 1960 SMSA population aged 65 or older
POPN Average household size
EDUC Median school years completed by those over 22
HOUS % of housing units which are sound & with all facilities
DENS Population per sq. mile in urbanized areas, 1960
NONW % non-white population in urbanized areas, 1960
WWDRK % employed in white collar occupations
POOR % of families with income < $3000
HC Relative hydrocarbon pollution potential
NOX Same for nitric oxides
SOD Same for sulphur dioxide
HUMID Annual average % relative humidity at 1pm
MORT Total age-adjusted mortality rate per 100,000

En algunos textos se ha elegido el k-óptimo como 0.2 , simplemente observando a la traza ridge que es mostrado en la siguiemte figura



Usando MINITAB se obtienen los siguientes resultados

A) La regresión mínimo cuadrática con las variables originales

Regression Analysis: MORT versus PREC, JANT, ...

```
The regression equation is
MORT = 1764 + 1.91 PREC - 1.94 JANT - 3.10 JULT - 9.07 OVR65 - 107 POPN
           - 17.2 EDUC - 0.65 HOUS + 0.00360 DENS + 4.46 NONW - 0.19 WWDRK
           - 0.17 POOR - 0.672 HC + 1.34 NOX + 0.086 SOD + 0.11 HUMID

Predictor        Coef     SE Coef          T        P       VIF
Constant       1764.0       437.3       4.03    0.000
PREC           1.9054      0.9237       2.06    0.045       4.1
JANT           -1.938       1.108      -1.75    0.087       6.1
JULT           -3.100       1.902      -1.63    0.110       4.0
OVR65          -9.065       8.486      -1.07    0.291       7.5
POPN         -106.83       69.78      -1.53    0.133       4.3
```

```
EDUC           -17.16       11.86      -1.45    0.155       4.9
HOUS           -0.651       1.768      -0.37    0.714       4.0
DENS         0.003601    0.004027       0.89    0.376       1.7
NONW            4.460       1.327       3.36    0.002       6.8
WWDRK          -0.187       1.662      -0.11    0.911       2.8
POOR           -0.167       3.227      -0.05    0.959       8.7
HC            -0.6722      0.4910      -1.37    0.178      98.6
NOX             1.340       1.006       1.33    0.190     105.0
SOD            0.0863      0.1475       0.58    0.562       4.2
HUMID           0.107       1.169       0.09    0.928       1.9

S = 34.93       R-Sq = 76.5%     R-Sq(adj) = 68.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression        15      174630       11642      9.54    0.000
Residual Error    44       53681        1220
Total             59      228311
```

De aquí se obtiene s^2=1220

B) La siguiente es la salida con la Y centrada y las predictoras centradas y escaladas

Regression Analysis: ycenter versus x1esy, x2est, ...

```
The regression equation is
ycenter = 146 x1esy - 151 x2est - 113 x3est - 102 x4est - 111 x5est - 111 x6est
           - 25.7 x7est + 40.2 x8est + 306 x9est - 6.6 x10est - 5 x11est
           - 475 x12est + 477 x13est + 42.0 x14est + 4.4 x15est

Predictor        Coef     SE Coef          T        P
Noconstant
x1esy          146.13       70.05       2.09    0.043
x2est         -151.35       85.61      -1.77    0.084
x3est         -113.43       68.80      -1.65    0.106
x4est         -101.98       94.40      -1.08    0.286
x5est         -110.99       71.68      -1.55    0.129
x6est         -111.40       76.15      -1.46    0.150
x7est          -25.71       69.03      -0.37    0.711
x8est           40.22       44.48       0.90    0.371
x9est          305.59       89.93       3.40    0.001
x10est          -6.63       58.22      -0.11    0.910
x11est           -5.3       102.0      -0.05    0.958
x12est         -474.9       343.0      -1.38    0.173
x13est          476.9       353.9       1.35    0.184
x14est          42.00       71.03       0.59    0.557
x15est           4.40       47.70       0.09    0.927

S = 34.54

Analysis of Variance

Source            DF          SS          MS         F        P
Regression        15      174630       11642      9.76    0.000
Residual Error    45       53681        1193
Total             60      228311
```

El vector de coeficientes $b_i^*(0)$

```
146.129  -151.345  -113.433  -101.978  -110.986  -111.397   -25.714
    40.216   305.591    -6.631    -5.349  -474.876   476.933    42.002
     4.403
```

Obteniendose, $\sum_{i=0}^{p} b_i^{*2}(0) =$ 642746

Finalmente, el k-óptimo será k=15*1220/622746=0.0284716. Bastante más pequeño de lo que dice la literatura.

Existen otros métodos para elegir la traza Ridge, que están relacionanados con los criterios usados para seleccionar los mejores subconjuntos tales como el C_p de Mallows, Validación Cruzada, PRESS, Validación cruzada generalizada, BIC, AIC, etc.

7.2.1 Aplicación de Regresión Ridge a Selección de variables

Según Hoerl y Kennard la regresión ridge puede usarse para seleccionar variables de la siguiente manera:

a) Eliminar las variables cuyos coeficientes sean estables pero de poco valor. Si se trabaja con variables previamente estandarizadas, se pueden comparar directamente los coeficientes.
b) Eliminar las variables con coeficientes inestables que tienden a cero.
c) Eliminar las variables con coeficientes inestables.

Ejemplo 4. Usar el conjunto de datos **Pollution** los datos y aplicar regresión ridge para seleccionar variables predictoras.

Solución: Usando SAS se obtiene la siguiente salida

OBS	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	INTERCEPT	PRECIP	JANTEMP	JULTEMP	OVR65	POPN	EDUC
1	MODEL1	PARMS	MORT	.	.	34.9285	1763.98	1.90542	-1.93762	-3.10043	-9.06540	-106.826	-17.1572
2	MODEL1	RIDGE	MORT	0.1	.	36.5458	1280.23	1.68862	-1.29827	-1.67510	-3.26643	-27.078	-9.0576
3	MODEL1	RIDGE	MORT	0.2	.	37.8288	1139.49	1.51214	-1.02409	-1.10190	-2.33137	-3.346	-8.3319
4	MODEL1	RIDGE	MORT	0.3	.	38.9837	1072.60	1.37928	-0.85077	-0.77000	-1.93227	8.117	-8.1336
5	MODEL1	RIDGE	MORT	0.4	.	40.0250	1035.23	1.27425	-0.72849	-0.54990	-1.71974	14.528	-8.0491
6	MODEL1	RIDGE	MORT	0.5	.	40.9739	1012.20	1.18855	-0.63669	-0.39235	-1.58968	18.414	-7.9841
7	MODEL1	RIDGE	MORT	0.6	.	41.8464	997.01	1.11693	-0.56489	-0.27394	-1.50157	20.884	-7.9140
8	MODEL1	RIDGE	MORT	0.7	.	42.6543	986.48	1.05591	-0.50704	-0.18188	-1.43699	22.492	-7.8337
9	MODEL1	RIDGE	MORT	0.8	.	43.4067	978.89	1.00312	-0.45936	-0.10852	-1.38661	23.548	-7.7435
10	MODEL1	RIDGE	MORT	0.9	.	44.1107	973.24	0.95685	-0.41935	-0.04897	-1.34534	24.233	-7.6454
11	MODEL1	RIDGE	MORT	1.0	.	44.7720	968.92	0.91584	-0.38529	0.00008	-1.31022	24.660	-7.5415

OBS	HOUSS	DENSS	WNOONNW	WDRKR	POOR	HC	NOX	SOD	HUMIDT	MORT	
1	-0.65112	.003600485	4.45960	-0.18706	-0.16764	-0.67214	1.34010	0.08625	0.10680	-1	
2	-1.14408	.004991729	3.51348	-0.50802	0.43558	-0.04714	0.08419	0.23997	0.31582	-1	
3	-1.16451	.005272350	2.94416	-0.46296	0.63735	-0.03115	0.05770	0.23816	0.38659	-1	
4	-1.15283	.005282467	2.58080	-0.45762	0.74431	-0.02556	0.04698	0.23013	0.38872	-1	
5	-1.13166	.005192915	2.32186	-0.46673	0.80776	-0.02270	0.04026	0.22064	0.36752	-1	
6	-1.10759	.005061951	2.12452	-0.47991	0.84746	-0.02094	0.03531	0.21105	0.33845	-1	
7	-1.08304	.004914159	1.96723	-0.49307	0.87269	-0.01972	0.03139	0.20185	0.30763	-1	
8	-1.05894	.004761212	1.83780	-0.50467	0.88847	-0.01881	0.02816	0.19319	0.27760	-1	
9	-1.03568	.004609020	1.72873	-0.51424	0.89776	-0.01810	0.02541	0.18511	0.24942	-1	
10	-1.01338	.004460646	1.63508	-0.52176	0.90243	-0.01752	0.02304	0.17761	0.22346	-1	
11	-0.99208	.004317651	1.55347	-0.52737	0.90374	-0.01703	0.02097	0.17063	0.19977	-1	

Las variables 4, 7, 10 11 y 15 tienen coeficientes estables pero pequeños. Las variables 12 y 13 tienen coeficientes inestables y que tienden a 0 y las variables 3 y 5 tienen coeficientes inestables. Así que solo deberían quedar en el modelo las variables 1, 2, 6, 8, 9 y 14.
Usando los métodos stepwise se eligen: 1, 2, 3, 6, 9 y 14 y usando los mejores subconjuntos se eligen: 1, 2, 6, 9, y 14.

7.3. Componentes principales para Regresión

El objetivo del análisis por componentes principales ((Hotelling, 1933) es hacer una reducción de la información disponible. Es decir, la información contenida en p variables predictoras $\mathbf{X}=(X_1,\ldots,X_p)$ puede ser reducida a $\mathbf{Z}=(Z_1,\ldots Z_{p'})$, con p'<p y donde las nuevas variables Z_i's llamadas las ***componentes principales*** no están correlacionadas. Los componentes principales de un vector aleatorio **X** son los elementos de una transformación lineal ortogonal de **X**
Geométricamente hablando la aplicación de componentes principales equivale a hacer una rotación de los ejes coordenados.
Consideremos el modelo de regresión lineal múltiple

$$Y = \beta_o + \beta_1 X_1 + \ldots\ldots + \beta_p X_p + e$$

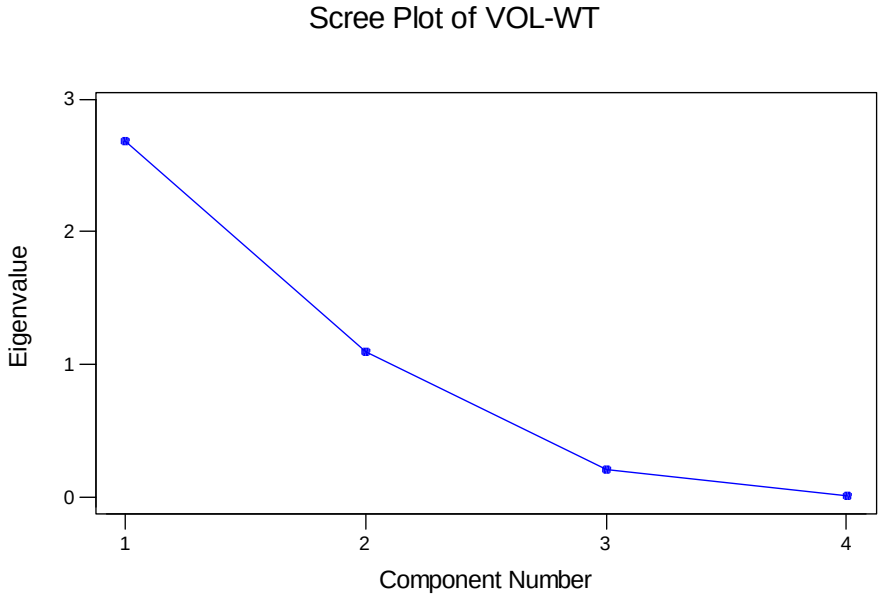
Estandarizemos todas las variables predictoras X_j usando $X_j^* = \dfrac{x_j - \bar{x}_j}{\sqrt{SXX_j}}$. Sea X* la matriz obtenida usando las X_j^* como columnas. Luego, X*'X* viene a ser la matriz de correlación de las variables predictoras X_j .

Para determinar los componentes principales hay que hallar una matriz ortogonal V tal que Z=X*V y para la cual Z'Z=(X*V)'(X*V)=V'X*'X*V=diag(λ_1,….,λ_p), y VV'=V"V=I., donde los λ_j son los valores propios de la matriz de correlación X*'X*. Luego, la j-ésima componente principal Z_j tiene desviación estándar igual a $\sqrt{\lambda_j}$ y puede ser escrita como

$$Z_j = v_{j1}X_1^* + v_{j2}X_2^* + ... + v_{jp}X_p^* \qquad (12)$$

donde v_{j1},v_{j2},…..v_{jp} son los elementos de la j-ésima fila de V. La matriz V es llamada la matriz de cargas ("loadings"), y contiene los coeficientes de las variables en cada componente principal. Los valores calculados de las componentes principales Z_j son llamados los valores rotados o simplemente "scores".

El número máximo de componentes principales que se puede construir es igual a p el número de variables predictoras. Sin embargo usando solo algunas de ellas se consiguen buenos resultados. Decidir acerca del número de componentes principales que se deben usar es un gran problema.



Elección del número de componentes principales:
Por lo general se usan las siguientes dos alternativas:

a) Elegir el número de compnentes hasta donde se ha acumulado por lo menos 75% de la proporción de los valores propios.
b) Elegir hasta la componente cuyo valor propio sea mayor que 1. Para esto se puede ayudar del "Scree Plot". (Ver figura)

MINITAB calcula los componentes principales. Hay que usar la secuencia **STAT ▸ Multivariate ▸ Principal components.** Elegir **correlation** en **Type of Matrix**

Ejemplo 5. Hallar los componentes principales del conjunto de datos de **millaje**

Principal Component Analysis: VOL, HP, SP, WT

```
Eigenanalysis of the Correlation Matrix

Eigenvalue    2.6894    1.1004    0.2055    0.0046
Proportion     0.672     0.275     0.051     0.001
Cumulative     0.672     0.947     0.999     1.000

Variable         PC1       PC2       PC3       PC4
VOL           -0.139    -0.913    -0.382    -0.019
HP            -0.600     0.156    -0.116    -0.776
SP            -0.562     0.292    -0.523     0.571
WT            -0.553    -0.237     0.753     0.267
```

Las funciones **prcomp** y **princomp** de la librería **mva** de R calcula los componentes principales. Para el conjunto millaje se obtiene los siguientes resultados

```
> pc<-prcomp(millaje[,2:5],scale=T,retx=T)
> pc
Standard deviations:
[1] 1.63994544 1.04901826 0.45331794 0.06813577

Rotation:
        PC1       PC2       PC3        PC4
sp  0.5616362 -0.2921753  0.5229999  0.57067460
wt  0.5526927  0.2365918 -0.7530524  0.26733338
vol 0.1387771  0.9134537  0.3820731 -0.01906127
hp  0.5998637 -0.1557567  0.1157728 -0.77620876
> summary(pc)
Importance of components:
                        PC1   PC2    PC3     PC4
Standard deviation     1.640 1.049 0.4533 0.06814
Proportion of Variance 0.672 0.275 0.0514 0.00116
Cumulative Proportion  0.672 0.947 0.9988 1.00000
```

De acuerdo al screeplot de la figura habria que usar solamente dos componentes principales. Componentes principales se puede aplicar en regresión aunque con reservas ya que ellos no han usado para nada la variable de respuesta en sus cálculos. Hay dos opciones

a) La opción clásica: Hacer la regresión solamente con las componentes principales que no han sido eliminadas, o
b) La opción moderna: Hacer la regresión con el número de componentes principales determinadas por un proceso de selección de variables para elegir el mejor modelo.

Haciendo la regresión de Y* versus todas las componentes principales se obtiene el modelo estimado

$$\hat{Y}^* = \hat{\alpha}_1 Z_1 + ... + \hat{\alpha}_p Z_p \qquad (13)$$

En forma matricial, el modelo Y=XB+e se transforma en el modelo Y=ZA+e' donde Z=XV y A=V'B. Se puede mostrar que la varianza del vector de coeficientes A está dado por $\Lambda^{-1}\sigma^2$, donde Λ= es una matriz diagonal cuyos elementos son los valores propios. Es decir, la varianza estimada del j-ésimo coeficiente de regresión con los componentes principales viene dada por $Var(\alpha_j)=s^2/\lambda_j$.

MINITAB produce los siguientes resultados:

A) Regresión con las variables originales:

Regression Analysis: MPG versus VOL, HP, SP, WT

```
The regression equation is
MPG = 192 - 0.0156 VOL + 0.392 HP - 1.29 SP - 1.86 WT

Predictor        Coef     SE Coef          T        P       VIF
Constant       192.44       23.53       8.18    0.000
VOL          -0.01565     0.02283      -0.69    0.495       1.6
HP            0.39221     0.08141       4.82    0.000     130.0
SP            -1.2948      0.2448      -5.29    0.000      71.7
WT            -1.8598      0.2134      -8.72    0.000      18.3

S = 3.653       R-Sq = 87.3%     R-Sq(adj) = 86.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4      7080.1      1770.0    132.66    0.000
Residual Error    77      1027.4        13.3
Total             81      8107.5

Source       DF      Seq SS
VOL           1      1101.6
HP            1      4731.1
SP            1       233.6
```

B) Regresión con las variables originales estandarizadas:

Regression Analysis: mpgest versus volest, hpest, spest, wtest

```
The regression equation is
mpgest = - 0.0347 volest + 2.23 hpest - 1.82 spest - 1.51 wtest

Predictor        Coef     SE Coef          T        P
Noconstant
volest       -0.03466     0.05025      -0.69    0.492
hpest          2.2283      0.4596       4.85    0.000
spest         -1.8168      0.3412      -5.32    0.000
wtest         -1.5134      0.1725      -8.77    0.000

S = 0.3628
```

Importante: Cuando se estandarizan las variables predictoras y la variable de respuesta. Los coeficientes estandarizados son iguales a $\frac{s_j}{s}\hat{\beta}$, donde s es la desviación estándar de las y's, y s_j es la desviación estándar de la j-ésima variable. Los valores de la prueba t y F no cambian y la varianza estimada s^2 cambia a s^2/Syy. En este ejemplo, $s^2=13.3/8107.5=0.00164046$

c) Regresión con todas las componentes principales sin estandarizar la respuesta.

Regression Analysis: MPG versus PC1, PC2, PC3, PC4

```
The regression equation is
MPG = 33.8 + 5.25 PC1 + 2.06 PC2 - 4.34 PC3 - 31.7 PC4

Predictor        Coef     SE Coef          T        P       VIF
Constant      33.7817      0.4034      83.75    0.000
PC1            5.2521      0.2475      21.22    0.000       1.0
PC2            2.0608      0.3869       5.33    0.000       1.0
PC3           -4.3445      0.8953      -4.85    0.000       1.0
PC4           -31.719       5.957      -5.32    0.000       1.0

S = 3.653       R-Sq = 87.3%     R-Sq(adj) = 86.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4      7080.1      1770.0    132.66    0.000
Residual Error    77      1027.4        13.3
Total             81      8107.5

Source       DF      Seq SS
PC1           1      6009.0
PC2           1       378.6
PC3           1       314.2
PC4           1       378.3
```

Observese que los VIF son iguales a 1. Lo que indica que ya no hay multicolinealidad. El R^2 de este modelo y del modelo con las variables originales también es el mismo, ya que el coeficiente de determinación es invariante a transformaciones lineales.
Si sustituimos los compomentes principales PC1-PC4 por sus respectivas ecuaciones como en la ecuación (12) se obtendrá la ecuación original. Así por ejemplo, el coeficiente –1.29 de la variable SP resulta de la siguiente ecuación

```
[5.252*(-0.562)+2.06*(0.292)-4.344*(-0.53)-31.719*(0.571)]/14.038

aqui 14.038 es la desviación estándar de la variable SP.
```

d) Regresión con solo las dos primeras componentes principales, sin estandarizar la respuesta.

Regression Analysis: MPG versus pc1, pc2

```
The regression equation is
MPG = 33.8 + 5.25 pc1 + 2.06 pc2

Predictor        Coef     SE Coef          T        P
Constant      33.7817      0.5153      65.56    0.000
pc1            5.2521      0.3161      16.61    0.000
pc2            2.0608      0.4942       4.17    0.000

S = 4.666       R-Sq = 78.8%     R-Sq(adj) = 78.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      6387.6      3193.8    146.70    0.000
Residual Error    79      1719.9        21.8
Total             81      8107.5
```

Aqui si no se obtiene la ecuación original al substituir los componentes principales. Por ejemplo el coeficiente de SP será

[5.252*(-0.562)+2.06*(0.292)]/14.038=-0.167

Similarmente para VOL, HP y WT se obtienen:
[5.252*(-0.139)+2.06*(-.913)]/22.166=-0.117
[5.252*(-0.600)+2.06*(0.156)]/56.841=-0.049
[5.252*(-0.553)+2.06*(-.237)]/8.1414=-0.416

Y el intercepto dará

```
33.8+5.25*((0.562*112.41)/14.038+(0.139*98.805)/22.166+(0.6*117.13)/56.
841+(0.553*30.915)/8.1414)+2.06*((-
0.292*112.41)/14.038+(0.913*98.805)/22.166+(-
0.156*117.13)/56.841+(0.237*30.915)/8.1414)=82.953
```

Aquí los coeficientes de regresión son estimaciones sesgadas del los coeficientes de regresión poblacional.
Los coeficientes de la regresión con los componentes principales no son fáciles de interpretar porque en este caso las predictoras son combinaciones lineales de las predictoras originales. Lo que hay que hacer es tratar de expresar la regresión en términos de las variables originales para luego hacer la interpretación.

Aplicando el procedimiento de selección hacia adelante (“Forward”) se obtiene los siguientes resultados

Stepwise Regression: mpg versus pc1, pc2, pc3, pc4

```
Forward selection.  Alpha-to-Enter: 0.15

 Response is   mpg    on  4 predictors, with N =   82


    Step          1        2        3        4
Constant      33.78    33.78    33.78    33.78
```

pc1	5.25	5.25	5.25	5.25
T-Value	15.14	16.61	18.69	21.22
P-Value	0.000	0.000	0.000	0.000
pc2		-2.06	-2.06	-2.06
T-Value		-4.17	-4.69	-5.33
P-Value		0.000	0.000	0.000
pc4			31.7	31.7
T-Value			4.69	5.32
P-Value			0.000	0.000
pc3				4.34
T-Value				4.85
P-Value				0.000
S	5.12	4.67	4.15	3.65
R-Sq	74.12	78.79	83.45	87.33
R-Sq(adj)	73.79	78.25	82.82	86.67
C-p	79.3	52.9	26.5	5.0

El mejor modelo según el método “forward” es aquel que incluye todas las componentes principales. Esto no debe ser tomado como regla general, Notar sin embargo que el método clásico recomendada las regresión con solamente dos componentes principales.

Ejercicios

1. a) Probar que el estimador Ridge $\hat{\boldsymbol{\beta}}_{\mathbf{R}}$ es un estimador sesgado del vector de coeficientes de regresión $\boldsymbol{\beta}$.

   b) Descomponer el error cuadrático medio total $\sum_{j=1} E(\hat{\beta}_{jR} - \beta_j)^2$ en la suma de la varianza total de los estimados ridge mas su sesgo al cuadrado. Encontrar explícitamente ambos términos.

2. Hallar la varianza del estimador de regresión Ridge.
3. Probar que el estimador ridge es la solución del problema de optimización

   Min (y-XB)'(y-XB)
   Sujeto a que: $B'B \leq k^2$

4. Aplicar 5 diagnósticos de multicolinealidad al conjunto de datos BERKELEY disponible en la página de internet del curso, y aplicar regresión ridge para resolver el problema creado . Evaluar el efecto

5 . Considerando el conjunto de datos CRIMEN, disponible en la página de internet del curso

a) Hallar los componentes principales para su conjunto de predictoras
b) Hallar la regresión usando los mas importantes componentes principales
c) Hallar la regresión usando "forward". Comparar sus resultados con el resultado anterior.

CAPÍTULO 8

REGRESIÓN ROBUSTA

8.1 Introducción

El método de mínimos cuadrados de ajustar una tendencia lineal es una de las herramientas más usadas desde que fue introducida en los inicios de los 1800. Sin embargo es un hecho reconocido por muchos que la presencia de datos "outliers" (verticales u horizontals) tiene un gran influencia en el ajuste por mínimos cuadrados. Un "outlier" mientras más exagerado sea hará que el ajuste lineal tienda a pasar cerca de él y el análisis de los residuales no sería muy confiable ya que estos darían la impression de que nada malo estuviera pasando. Cuando hay presente datos influenciales la alternativas a mínimos cuadrados son las siguientes:

0. Seguir usando mínimos cuadrados como si nada hubiese pasado.
1. Hacer una análisis exploratorio de datos antes de ajustar el modelo. Esto incluye el uso de diagnósticos, transformaciones, gráficas dinámicas, etc.
2. Usar modelos lineales generalizados.
3. Usar métodos de regresión robusta, las cuales son modificaciones de los mínimos cuadrados y tiene como objetivo ajustar un modelo que resista la influencia de los "outliers".

Los origenes de la regresión robusta se remontan a 1973, cuando se publicó en el *Annals of Statistics* el artículo de Peter Huber. Existen una variedad de métodos robustos de regresión que son agrupados en tres:

a) M-Regresión (M es por Máxima verosimilitud),
b) R-Regresión (R es por Rangos) y
c) L-Regresión (L es por combinación lineal de estadísticos de orden).

Existen también varias modificaciones de éstos. Los que han alcanzado más popularidad han sido los estimadores M de Regresión Robusta. En este texto solamente se discutirán los estimadores M y sus variantes.

Actualizando los comentarios hechos por Hogg en 1979. con la gran cantidad de métodos robustos existentes uno podría pensar que el uso de regresión robusta debería ser bastante amplio, pero este no ha sido el caso. Los principales programas estadísticos, excepto por S-Plus prácticamente ignoran regresión robusta. Nada más en la regresión M existen tantas maneras de elegir la función peso, tantas maneras de elegir las constantes de afinamiento y tantas posibilidades de elegir la constante de escala (variabilidad) que se le hace bien difícil al simple usuario elegir la estimación más adecuada. Para complicar más la cosa el cálculo de los estimadores se hace iterativamente y la elección del punto inicial y del algoritmo iterativo es crucial para alcanzar rapidamente convergencia. Finalmente, se ha avanzado muy poco en la parte inferencial

Después de los textos de Hampel, et al. (1986) y de Rousseeuw y Leroy (1987) no ha aparecido ningún otro texto importante en regresión robusta. Muchos de los investigadores pioneros de la regresión robusta se han movido a otras areas y desde los mediados de los 90's la investigación en esta área ha disminuido. Con el desarrollo de las computadoras la investigación se ha enfocado más ahora en regresión noparamétrica, que se verá en el próximo capítulo.

8.2 Regresión L_1

Consideremos el modelo de regresión lineal múltiple y=**Xβ**+e. En mínimos cuadrados para estimar el vector de parámetros **β** se debe minimizar $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x'}_i\boldsymbol{\beta})^2$ donde $\mathbf{x'}_i$ es una fila de la matriz **X**.

En este caso si el residual es pequeño se le da poco peso y si es grande se le da mayor peso siguiendo una ecuación cuadrática. En 1887, Edgeworth propuso reemplazar los errores cuadráticos por errores en valor absoluto. Más especificamente, la ecuación de regresión se obtendría minimizando $\sum_{i=1}^{n} |e_i|$. Esta regresión es llamada ***regresión L_1 o regresión de suma absoluta mínima.*** Sin embargo es usada muy restringidamente por las siguientes razones.

a) El vector de coeficientes estimados no es único.
b) La regresión L_1 resiste la presencia de valores anormales en la dirección vertical. Pero es poco efectivo para valores anormales en la dirección x.
c) La eficiencia del estimador disminuye a medida que aumenta el número de casos.
d) Para obtener las estimaciones del coeficiente de regresión hay que resolver un problema de programación lineal, el cual es muy lento computacionalmente.

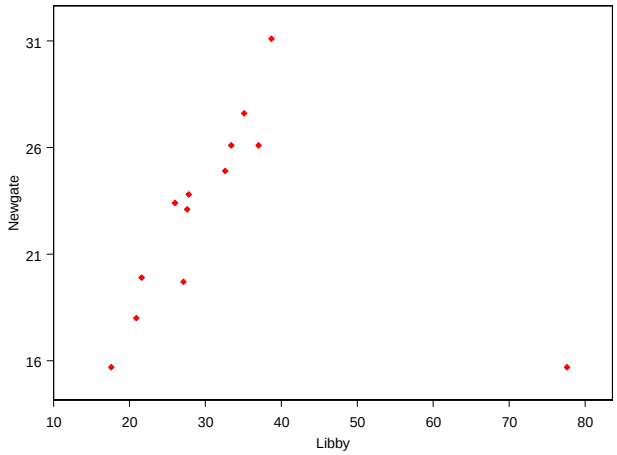
S-Plus tiene una función **l1fit** que calcula la regresión L_1.

Ejemplo 1. El siguiente conjunto de datos llamado **Kootenay**, es clásico en los libros de regresión robusta. Aquí se mide el caudal del rio Kootenay entre 1971 y 1973 en dos lugares uno llamado: Libby y otro que viene más adelante llamado Newgate. La idea es predecir el caudal en Newgate una vez que se observa lo que pasa en Libby. Ajustaremos una línea por mínimos cuadrados y una línea de regresión L_1 usando S-Plus.

	Libby	Newgate
1	27.1	19.7
2	20.9	18.0
3	33.4	26.1
4	77.6	15.7
5	37.0	26.1
6	21.6	19.9
7	17.6	15.7
8	35.1	27.6
9	32.6	24.9
10	26.0	23.4
11	27.6	23.1
12	38.7	31.1
13	27.8	23.8

En la siguiente gráfica se muestra el diagrama de puntos de los datos

Notar el punto en la parte derecha inferior qqe está bien alejado de la mayor parte de los datos.



Haciendo la regresión por minimos cuadrados (**lsfit**) y la regresión L_1 (l**1fit**) se obtiene

```
S-PLUS : Copyright (c) 1988, 2001 Insightful Corp.
S : Copyright Lucent Technologies, Inc.
Professional Edition Version 6.0.3 Release 2 for Microsoft Windows : 2001
Working data will be in C:\Program Files\Insightful\splus6\users\Administrator
> Kootenay
   Libby Newgate
 1  27.1    19.7
 2  20.9    18.0
 3  33.4    26.1
 4  77.6    15.7
 5  37.0    26.1
 6  21.6    19.9
 7  17.6    15.7
 8  35.1    27.6
 9  32.6    24.9
10  26.0    23.4
11  27.6    23.1
12  38.7    31.1
13  27.8    23.8
> koo.ls<-lsfit(Kootenay$Libby,Kootenay$Newgate)
> koo.ls
$coef:
 Intercept          X
  23.16443 -0.01427324

$residuals:
 [1] -3.0776245 -4.8661186  3.4122969 -6.3568256  3.4636806
 [6] -2.9561273 -7.2132203  4.9365615  2.2008784  0.6066749
[11]  0.3295121  8.4879451  1.0323668

$intercept:
[1] T
```

```
$qr:
$qr$qt:
 [1] -81.8460140  -0.7407183   3.9430373  -8.8018133
 [5]   3.7520540  -1.6309619  -5.6187582   5.3528507
 [9]   2.7854780   1.6356141   1.2507326   8.6618674
[13]   1.9401224

$qr$qr:
      Intercept           X
 [1,] -3.6055513 -117.31909150
 [2,]  0.2773501   51.89557562
 [3,]  0.2773501   -0.03935572
 [4,]  0.2773501   -0.89106609
 [5,]  0.2773501   -0.10872579
 [6,]  0.2773501    0.18802397
 [7,]  0.2773501    0.26510184
 [8,]  0.2773501   -0.07211381
 [9,]  0.2773501   -0.02394015
[10,]  0.2773501    0.10323833
[11,]  0.2773501    0.07240718
[12,]  0.2773501   -0.14148389
[13,]  0.2773501    0.06855329

$qr$qraux:
[1] 1.277350 1.201513

$qr$rank:
[1] 2

$qr$pivot:
[1] 1 2

$qr$tol:
[1] 1e-007


> koo.l1<-l1fit(Kootenay$Libby,Kootenay$Newgate)
> koo.l1
$coefficients:
 Intercept       X
  14.29149 0.319149

$residuals:
 [1]  -3.2404251  -2.9617021   1.1489357 -23.3574467
 [5]   0.0000000  -1.2851073  -4.2085114   2.1063848
 [9]   0.2042551   0.8106378   0.0000000   4.4574466
[13]   0.6361693
># Los siguientes commandos sirven para mostrar el diagama de puntos y las lineas de
># regresion por minimos cuadrados y L1 simultaneamente
> plot(Kootenay$Libby,Kootenay$Newgate)
> abline(koo.ls)
```

```
> abline(koo.l1)
```

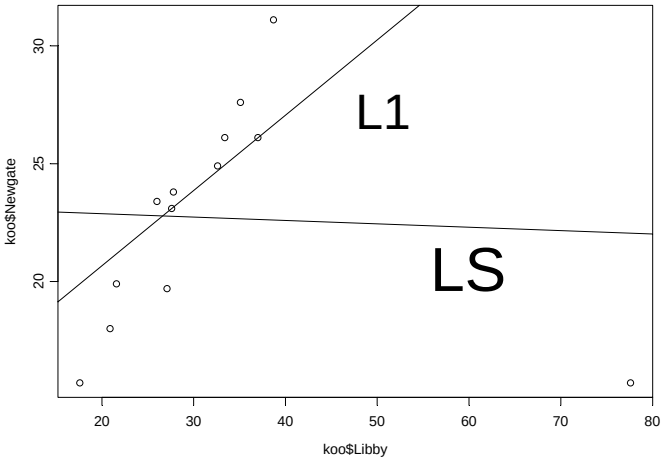
De la salida se obtiene que la regresión por mínimos cuadrados es

Newgate= 23.16443 -0.01427324 Libby

En tanto que la regresión L1 será

Newgate= 14.29149 +0.319149 Libby

A coninuación se muestra el diagrama de puntos y las dos líneas de regresión.



En la gráfica se puede ver el gran efecto del punto leverage sobre la línea de mínimos cuadrados.

8.3. Regresión M

En 1973, Huber propuso un nuevo método de regresión que era como una combinación de la regresión L1 y la regresión por mínimos cuadrados. La idea se basaba en que para residuales pequeños se le da un peso cuadrático y para residuales grandes se le da un peso lineal. Más específicamente, la propuesta de Huber era minimizar con respecto a B

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x'_i}\,\boldsymbol{\beta}) \qquad (1)$$

donde la función peso ρ se define por

$$\rho(t) = \begin{cases} \dfrac{t^2}{2} & |t| \le c \\ c|t| - \dfrac{c^2}{2} & |t| > c \end{cases} \qquad (2)$$

c es una constante de afinamiento que depende del nivel de eficiencia. Para una eficiencia del 95% se escoge c=1.345. Como se verá más adelante, la función peso de Huber ρ es la más usada entre las muchas funciones pesos existentes. Por otro lado, debido a que la varianza σ^2 de la variable de respuesta afecta la estimación de **β,** la función a minimizar es la que sigue

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x'_i}\,\boldsymbol{\beta}}{\sigma}\right) \qquad (3)$$

ρ es una función simétrica con mínimo en 0 y σ es desconocido. Derivando con respecto a **β** e igualando a cero se obtiene el sistema de p ecuaciones no lineales.

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - x'_i\, B}{\sigma}\right) x'_i = \mathbf{0} \qquad (4)$$

donde ψ=ρ'. En el caso de mínimos cuadrados $\rho(t)=t^2/2$ y ψ(t)=t.
Se puede mostrar que también se llega a la ecuación (4) a través del método de máxima versimilitud. En particular la propuesta de Huber es obtenida si se asume que los errores siguen una distribución normal contaminada con una doble exponencial, es decir, una mezcla de una normal y una doble exponencial.
El parámetro de escala (o de variabilidad) σ puede ser estimado de antemano y sustituido en (4). La MAD, ***desviación absoluta con respecto a la mediana,*** de los residuales es la medida que más se usa para estimar σ. Se define por

$$\text{MAD}=1.4825*\text{mediana}|e_i-\text{mediana}(e_i)| \qquad (5)$$

Otra alternativa más complicada es estimar σ simultaneamente con **β**.

Existen varias propuestas para las función peso ψ, entre las principales están

Andrews $\psi(u)= c*\sin(u/c)$ si $|u| \le \pi*c$ y 0 en otro caso. El valor default de c=1.339.

Bisquare (o Biweight) $\psi(u)=u*(1 - (u/c)^2)^2$ if $|u| \le c$ y 0 en otro caso. El valor default de c=4.685.

Cauchy $\psi(u)=u/(1 + (u/c)^2)$ con c=2.385 como valor default.

Fair $\psi(u)=u/(1 + |u|/c)$ y c=1.4 es el valor default.

Hampel Tiene 3 constantes de afinamiento, a, b y c. Se define por ψ(u) =u si |u|<=a; a*sign(u) si a< |u|<=b; a*(c-|u|)/(c-b)*sign(u) si b<|u|<=c; y 0 en otro caso. Los valores defaults para a, b y c son 2, 4 y 8 respectivamente.

Huber ψ(u)= u si |u| < c y c*sign(u)| en otro caso. El valor default de c=1.345.

Logistic ψ(u)= c*tanh(u/c) con valor default c=1.205.

Talworth ψ(u)=u si |u| ≤c y 0 en otro caso. El default es c=2.795.
Welsch $\psi(u)=u*\exp(-(u/c)^2)$. El valor default es c =2.985.

La función **rreg** de S-Plus y la función **rlm** del package **MASS** de **R** permiten calcular los estimadores m de regresión

Ejemplo 2. Calcular la regresión de Huber para el conjunto de datos **kootenay**.

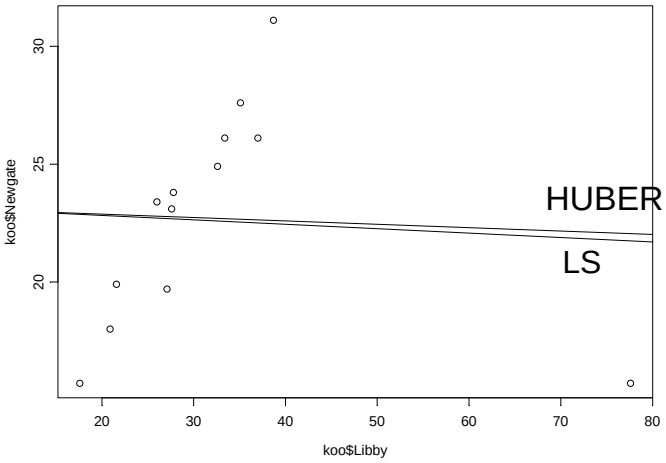
Usando S-Plus se obtiene los siguientes resultados

```
> koo.rob<-rreg(koo$Libby,koo$Newgate,method=wt.huber)
> koo.rob
$coefficients:
 (Intercept)          x
    23.19752 -0.01872314

$residuals:
 [1]  -2.990126  -4.806209   3.527830  -6.044607   3.595234  -2.893103  -7.167995   5.059660
2.312852  0.689279
[11]  0.419236  8.627063  1.122981

$fitted.values:
 [1]  22.69013  22.80621  22.57217  21.74461  22.50477  22.79310  22.86800  22.54034  22.58715
22.71072 22.68076
[12] 22.47294 22.67702

$w:
 [1]  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9799193  1.0000000
1.0000000 1.0000000
[11] 1.0000000 0.8149970 1.0000000

$int:
[1] T

$conv:
[1] 0.029915705 0.005059456 0.001065885

$status:
[1] "converged"

> plot(koo$Libby,koo$Newgate)
> abline(koo.ls)
> abline(koo.rob)
> plot(koo$Libby,koo$Newgate)
> abline(koo.ls)
> abline(koo.rob)
```

la ecuación de regresión Huber será

Newgate=23.19752 -0.01872314Libby

Ploteando la regresión minimocuadrática y la Huber se obtiene la siguiente gráfica



Notar que la regresión Huber también es afectada por el valor influencial.

En R todo lo anterior se ejecuta de la siguiente manera

R : Copyright 2001, The R Development Core Team
Version 1.3.1 (2001-08-31)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()' or `licence()' for distribution details.

```
> # asumiendo que los datos estan en el archivo koo.txt en el directorio de R
> uno<-read.table("koo.txt",header=T)
> uno
   Libby Newgate
1   27.1    19.7
2   20.9    18.0
3   33.4    26.1
4   77.6    15.7
5   37.0    26.1
6   21.6    19.9
7   17.6    15.7
8   35.1    27.6
9   32.6    24.9
10  26.0    23.4
11  27.6    23.1
12  38.7    31.1
13  27.8    23.8
> lsfit(uno$Libby,uno$Newgate)
$coefficients
  Intercept          X
23.16442942 -0.01427324
```

```
$residuals
 [1] -3.0776245 -4.8661186  3.4122969 -6.3568256  3.4636806 -2.9561273
 [7] -7.2132203  4.9365615  2.2008784  0.6066749  0.3295121  8.4879451
[13]  1.0323668

$intercept
[1] TRUE

$qr
$qr$qt
 [1] -81.8460140  -0.7407183   3.9430373  -8.8018133   3.7520540  -1.6309619
 [7]  -5.6187582   5.3528507   2.7854780   1.6356141   1.2507326   8.6618674
[13]   1.9401224

$qr$qr
      Intercept           X
 [1,] -3.6055513 -117.31909150
 [2,]  0.2773501   51.89557562
 [3,]  0.2773501   -0.03935572
 [4,]  0.2773501   -0.89106609
 [5,]  0.2773501   -0.10872579
 [6,]  0.2773501    0.18802397
 [7,]  0.2773501    0.26510184
 [8,]  0.2773501   -0.07211381
 [9,]  0.2773501   -0.02394015
[10,]  0.2773501    0.10323833
[11,]  0.2773501    0.07240718
[12,]  0.2773501   -0.14148389
[13,]  0.2773501    0.06855329

$qr$qraux
[1] 1.277350 1.201513

$qr$rank
[1] 2

$qr$pivot
[1] 1 2

$qr$tol
[1] 1e-07
> rlm(Newgate~.,data=uno,psi=psi.huber)
Call:
rlm.formula(formula = Newgate ~ ., data = uno)
Converged in 5 iterations

Coefficients:
(Intercept)       Libby
23.19413503 -0.01863951

Degrees of freedom: 13 total; 11 residual
```

```
Scale estimate: 5.23
> r1<-rlm(Newgate~.,data=uno)
> summary(r1)

Call: rlm.formula(formula = Newgate ~ ., data = uno)
Residuals:
    Min      1Q  Median      3Q     Max
-7.1661 -2.9890  0.6905  3.5284  8.6272

Coefficients:
            Value   Std. Error t value
(Intercept) 23.1941  3.8521     6.0212
Libby       -0.0186  0.1083    -0.1722

Residual standard error: 5.231 on 11 degrees of freedom

Correlation of Coefficients:

       (Intercept)
  Libby -0.9145
> attributes(r1)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "w"             "s"             "psi"           "k2"
[13] "conv"          "converged"     "x"             "call"
[17] "terms"         "xlevels"       "model"

$class
[1] "rlm" "lm"

> r1$residuals
         1          2          3          4          5          6          7
-2.9890044 -4.8045694  3.5284245 -6.0477094  3.5955267 -2.8915217 -7.1660797
         8          9         10         11         12         13
 5.0601116  2.3135129  0.6904921  0.4203153  8.6272139  1.1240432
```

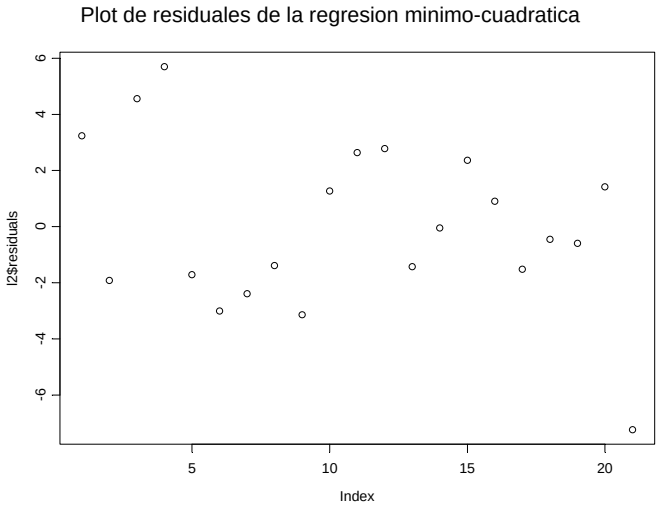
Ejemplo 3. El siguiente conjunto de datos llamado **Stackloss**, es otro también bien usado en textos de regresión robusta, consiste de 21 observaciones de 4 variables; Stackloss(y), rate(flujo del aire) temp(temperatura del agua) y Acid (concentración de acido). A continuación se muestran los datos, la regresión robusta usando la función bisquare y el plot de residuales de la regresión minimo cuadrática y de la regresión bisquare.

```
stackloss
  stackloss rate Temp Acid
1        42   80   27   89
2        37   80   27   88
3        37   75   25   90
4        28   62   24   87
5        18   62   22   87
6        18   62   23   87
```

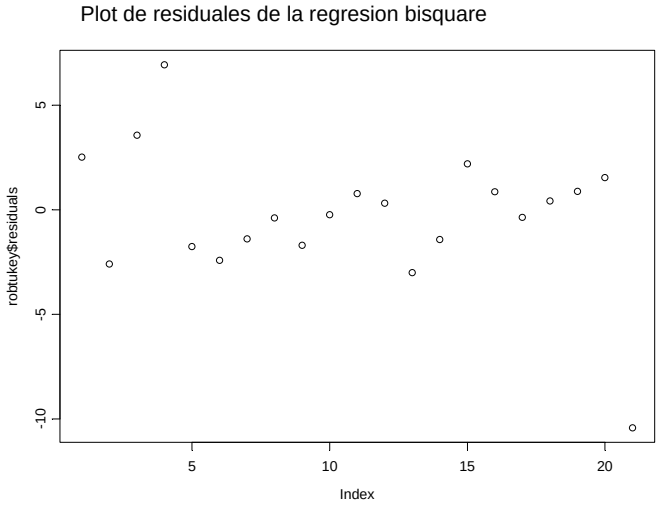
```
 7       19   62   24   93
 8       20   62   24   93
 9       15   58   23   87
10       14   58   18   80
11       14   58   18   89
12       13   58   17   88
13       11   58   18   82
14       12   58   19   93
15        8   50   18   89
16        7   50   18   86
17        8   50   19   72
18        8   50   19   79
19        9   50   20   80
20       15   56   20   82
21       15   70   20   91
>stackloss< -as.matrix(stackloss)
>stackloss1<-stackloss[ ,2:4]
>stackloss2<-stackloss[,1]
> robtukey<-rreg(stackloss1, stackloss2, method=wt.bisquare)
> robtukey
$coefficients:
 (Intercept)      rate      Temp      Acid
   -42.28459 0.9273997 0.6510081 -0.1123065

$residuals:
 [1]  2.5106733 -2.6016332  3.5619944  6.9322787 -1.7657050 -2.4167132 -1.3938823 -
0.3938823
 [9] -1.7071145 -0.2382194  0.7725391  0.3112407 -3.0136064 -1.4292430  2.1917365
0.8548170
[17] -0.3684821  0.4176634  0.8789617  1.5391767 -10.4336602

$fitted.values:
 [1] 39.489327 39.601633 33.438006 21.067721 19.765705 20.416713 20.393882 20.393882
16.707114 14.238219
[11] 13.227461 12.688759 14.013606 13.429243  5.808264  6.145183  8.368482  7.582337
8.121038 13.460823
[21] 25.433660

$w:
 [1] 0.892416394 0.885432076 0.789930362 0.336110641 0.946231809 0.900393417
0.966212323 0.997263673
 [9] 0.949531055 0.999032961 0.989517299 0.998266923 0.847699690 0.964641536
0.917808001 0.987286682
[17] 0.997611460 0.996972889 0.986607555 0.958974835 0.002377249

$int:
[1] T

$conv:
[1] 0.1623974286 0.0569754561 0.0822263512 0.0455994131 0.0419303121 0.0411232225
0.0213469351 0.0036786480
```

[9] 0.0008243514

$status:
[1] "converged"



Se notan como 4 outliers



Los 4 "outliers" se destacan un poco más en el plot de residuales correspondiente a la regresión bisquare, especialmente el que aparece en la parte inferior.

8.3.1 Cálculo de los estimadores M de regresión.

Los estimadores de regresión M se obtienen resolviendo el sistema de ecuaciones no lineales (4). La solución es obtenida mediante un proceso iterativo. Existen 3 propuestas para hacer esto:

a) ***El método de Newton Raphson:*** Cuya fórmula de iteración es:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \sigma\left[\sum_i \mathbf{x}_i^{'}\psi'(\frac{y_i - \mathbf{x_i}\hat{\boldsymbol{\beta}}^{(m)}}{\tilde{\sigma}})\mathbf{x}_i\right]^{-1}\mathbf{X'}\psi(\frac{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(m)}}{\tilde{\sigma}}) \qquad (6)$$

b) ***El método H de Huber:*** Teniendo en cuenta que el estimador mínimo cuadrático $\hat{\boldsymbol{\beta}}$ puede ser escrito como $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{(X'X)^{-1}X'e}$ entonces se puede deducir el siguiente proceso iterativo

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathbf{(X'X)^{-1}X'}\hat{\mathbf{e}}^{(m)} \qquad (7)$$

Huber (1977) sugirió el siguiente proceso iterativo

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \tilde{\sigma}(\mathbf{X'X})^{-1}\mathbf{X'}\psi(\frac{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(m)}}{\tilde{\sigma}}) \qquad (8)$$

La idea aquí es reemplazar en cada iteración el residual $\hat{\mathbf{e}} = \mathbf{y_i - x_i}\hat{\boldsymbol{\beta}}$ por el residual modificado $\psi(r_i/\hat{\sigma})\hat{\sigma}$.

c) ***El método de Minimos Cuadrados reponderados (IRLS)***

Beaton y Tukey(1974) se basaron en los mínimos cuadrados para proponer la siguiente fórmula iterativa

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \tilde{\sigma}\ (\mathbf{X'}W^{(m)}\mathbf{X})^{-1}\mathbf{X'}W^{(m)}(\mathbf{y} - \mathbf{X'}\boldsymbol{\beta}^{(\mathbf{m})}) \qquad (9)$$

donde $W^{(m)}$ es una matriz diagonal cuyos elementos son:

$$w_i^{(m)} = \frac{\psi((y_i - x_i\beta^{(m)})/\tilde{\sigma}}{(y_i - x_i\beta^{(m)})/\tilde{\sigma}} \qquad (10)$$

Este último procedimiento es el que implementa S-Plus y también puede programarse usando el procedimiento IML de SAS.

El valor inicial del proceso iterativo depende de la forma de la función ψ, si ésta es monótona se podria comenzar con los coeficientes mínimos cuadráticos, pero si no lo es se recomienda comenzar con los coeficientes de la regresión L_1 o de otra regresión robusta. Cuando se usa como valores iniciales los coeficientes de otra regresión robusta M entonces al estimador final se le llama estimador MM, nombre introducido por Yohai y Zammar (1990), propiamente el estimador inicial de escala tambien es estimado robustamente y en ese caso el estimador inicial M es llamdo un estimador robusto S.

A continuación se muestran algunos resultados en R de estimadores robustos para varias funciones pesos y parametros iniciales

```
> data(stackloss)
```

```
> r1<-rlm(stack.loss~.,data=stackloss)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss)
Converged in 9 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-41.0265311   0.8293739   0.9261082  -0.1278492

Degrees of freedom: 21 total; 17 residual
Scale estimate: 2.44
> r1<-rlm(stack.loss~.,data=stackloss,psi=psi.huber)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.huber)
Converged in 9 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-41.0265311   0.8293739   0.9261082  -0.1278492

Degrees of freedom: 21 total; 17 residual
Scale estimate: 2.44
> r1<-rlm(stack.loss~.,data=stackloss,psi=psi.hampel)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.hampel)
Converged in 4 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-40.4747324   0.7410830   1.2250793  -0.1455251

Degrees of freedom: 21 total; 17 residual
Scale estimate: 3.09
> library(lqs)
> r1<-rlm(stack.loss~.,data=stackloss,method="MM",psi=psi.huber)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.huber,
    method = "MM")
Converged in 11 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-41.5230431   0.9388404   0.5794524  -0.1129150

Degrees of freedom: 21 total; 17 residual
Scale estimate: 1.91
> r1<-rlm(stack.loss~.,data=stackloss,method="MM")
```

```
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, method = "MM")
Converged in 12 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-41.7072712   0.9372710   0.5940633  -0.1129477

Degrees of freedom: 21 total; 17 residual
Scale estimate: 1.98
> r1<-rlm(stack.loss~.,data=stackloss,method="MM",psi=psi.bisquare)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.bisquare,
   method = "MM")
Converged in 11 iterations

Coefficients:
(Intercept)    Air.Flow  Water.Temp  Acid.Conc.
-41.5230726   0.9388402   0.5794546  -0.1129150

Degrees of freedom: 21 total; 17 residual
Scale estimate: 1.91
> summary(r1)

Call: rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.bisquare,
   method = "MM")
Residuals:
     Min       1Q   Median       3Q      Max
-10.50957  -1.43820  -0.09084   1.02517   7.23167

Coefficients:
            Value    Std. Error t value
(Intercept) -41.5231   9.3070    -4.4615
Air.Flow      0.9388   0.1055     8.8983
Water.Temp    0.5795   0.2879     2.0125
Acid.Conc.   -0.1129   0.1223    -0.9234

Residual standard error: 1.912 on 17 degrees of freedom

Correlation of Coefficients:

            (Intercept) Air.Flow Water.Temp
  Air.Flow    0.1793
  Water.Temp -0.1489     -0.7356
  Acid.Conc. -0.9016     -0.3389   0.0002
> attributes(r1)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
```

```
 [9] "w"           "s"           "psi"         "k2"
[13] "conv"        "converged"   "x"           "call"
[17] "terms"       "xlevels"     "model"

$class
[1] "rlm" "lm"
```

Algoritmo para calcular la regresión robusta usando IRLS:

a) Calcular la MAD.
b) Hacer la regresión por mínimos cuadrados y guardar los coeficientes como $\hat{\boldsymbol{\beta}}^{(0)}$ y los residuales como $\hat{\mathbf{e}}_o$.
c) Estandarizar los residuales como $\mathbf{e}_o^* = \dfrac{\hat{\mathbf{e}}_o}{MAD}$
d) Calcular los pesos $\mathbf{w}_o = \psi(\mathbf{e}_o^*)/\mathbf{e}_o^*$.
e) Calcular la regresión por mínimos cuadrados ponderados usando la matriz diagonal de pesos W_o cuyos elementos en la diagonal son w_o, Guardar los coeficientes de regresión como $\hat{\boldsymbol{\beta}}^{(1)}$ y los residuales como $\hat{\mathbf{e}}^{(1)}$. Aquí.

$$\hat{\boldsymbol{\beta}}^{(1)} = (\mathbf{X}'\mathbf{W}_o\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_o\mathbf{y}$$

Sustituir el índice 0 por 1 y repetir los pasos c, d y e. Continuar el proceso hasta que la diferencia entre $\hat{\boldsymbol{\beta}}^{(n+1)}$ y $\hat{\boldsymbol{\beta}}^{(n)}$ sea despreciable.

De los tres algoritmos, el que converge más rápidamente es el de Newton-Raphson, pero tiene el problema que hay que calcular la derivada de ψ. El método de Huber es el que converge más lentamente pero tiene la ventaja de que $\mathbf{(X''X)^{-1}X'}$ se calcula una sola vez.

8.4 Regresión GM o Regresión de Influencia acotada.

La ***función influencia*** introducida por Hampel en 1974, mide el efecto de una observación en el cálculo de un estimador, asumiendo que los datos siguen una distribucion conocida F . En el caso de regresión la influencia (IF) puede ser escrita como el producto de la influencia en la dirección vertical (IR) por la diferencia en la dirección horizontal (IP). Por ejemplo, para mínimos cuadrados se obtiene que la influencia del estimador $\hat{\boldsymbol{\beta}}$ en el punto (y,**x**), asumiendo que la distribución F es la Normal está dada por

$$\text{IF}(y,\mathbf{x},F,\hat{\boldsymbol{\beta}}) = (y - \mathbf{x}\boldsymbol{\beta})\mathbf{(X'X)^{-1}x'} \qquad (11)$$

Además IR=$(y-\mathbf{x}\boldsymbol{\beta})$ y IP=$\mathbf{(X'X)^{-1}x'}$. Tanto IP como IR no son acotados por lo tanto el estimador minimo cuadrático no es robusto.
En el caso de los estimadores M se pude mostrar que IR es acotado pero no IP. El problema con la regresión M es que protege contra observaciones que son "outliers" verticales pero no contra aquellas observaciones que tienen un leverage alto. Para subsanar esta deficiencia

se han propuesto modificaciones a los estimadores M, los cuales son llamados ***estimadores M generalizados (GM)***. Básicamente hay 3 propuestas.

Estimador de Mallows (1975): Se obtiene modificando la ecuación (4) de la siguiente manera

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi(r_i/\tilde{\sigma})\mathbf{x}_i = \mathbf{0} \qquad (12)$$

donde w representa una función peso

Estimador de Schweppe (1975): En este caso la ecuación (4) se modifica a:

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi(r_i/(w(\mathbf{x}_i)\tilde{\sigma}))\mathbf{x}_i = \mathbf{0} \qquad (13)$$

donde $w(\mathbf{x}_i) = \sqrt{1-h_{ii}}$. Para calcular este estimador se puede usar mínimos cuadrados reponderados con w_i=min(1,2/|t_i|).

Estinador de Welsh (1982): Usa la misma ecuación anterior pero con $w(\mathbf{x}_i) = (1-h_{ii})/h_{ii}^{1/2}$. Para calcular este estimador se puede usar mínimos cuadrados reponderados con w_i=min(1,2/|DFFITS|), con $c=2(p/n)^{1/2}$.

Se ha establecido que cuando el número de variables predictoras se incrementa, la resistencia de los estimadores GM a los puntos leverages se va deteriorando.

8.5 Regresión de Mediana de Cuadrados Mínima (LMS)

Fue introducida por Rousseeuw en 1984. Es bien conocido que la mediana, es una medida más resistente que la media cuando hay presente "outliers". En efecto su "Breakdown Point" es 50% , mientras que el de la media es (1/n)*100%. Los minimos cuadrados están relacionados al uso de la media (Si se minimiza suma de cuadrados con respecto a un punto se puede mostrar que ese punto es la media). La propuesta de Rousseeuw, consiste en minimizar la mediana de los cuadrados de los residuales. Es decir

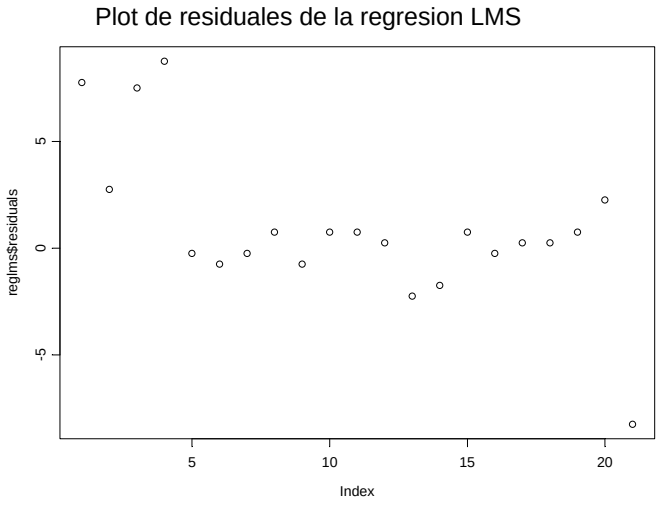
$$\text{Min Mediana}(y_i - \mathbf{x'}_i\boldsymbol{\beta})^2 \qquad (14)$$

Resulta ser que la regresión LMS resiste bastante bien los "outliers" verticales y horizontales. Esta regresión ha sido implementada en S-Plus, usando la función **lmsreg.**
Pero la desventaja de la regresión LMS es que no es eficiente aparte de ser computacionalmente pesada. Para resolver este problema, Rousseeuw (1984) introdujo la regresión por Sumas de Cuadrados podada mínima (LTS), en este caso se cuadran los residuales, luego se los ordena y se suma solamente los que no son muy grandes, se minimiza y el resultado dará la regresión LTS. Esta regresión ha sido implementada en S-Plus, usando la función **ltsreg**

Ejemplo: Calcular las regresiones LMS y LTS para los datos de Stackloss

```
> reglms<-lmsreg(stackloss1, stackloss2)
> reglms
$coefficients:
 Intercept rate Temp          Acid
```

```
   -39.25 0.75  0.5 -2.586466e-017

$scale:
       Y
 1.207615

$residuals:
 [1]  7.75  2.75  7.50  8.75 -0.25 -0.75 -0.25  0.75 -0.75  0.75  0.75  0.25 -2.25 -1.75  0.75 -0.25
0.25
[18]  0.25  0.75  2.25 -8.25

$intercept:
[1] T

$method:
[1] "Least Median of Squares with  3143 samples of size 4 , 143 were singular."
```

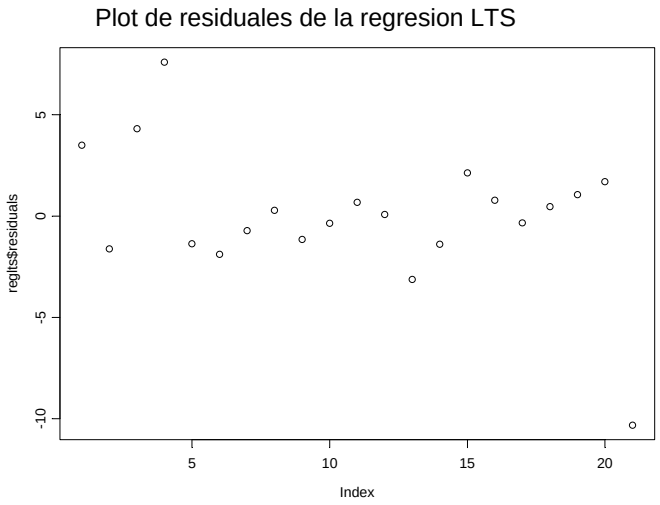


Notar que los 4 “outliers” se destacan bastante.

```
reglts<-ltsreg(stackloss1, stackloss2)
> reglts
Method:
Least Trimmed Squares Robust Regression.

Call:
ltsreg.default(stackloss1, stackloss2)

Coefficients:
 Intercept     rate     Temp     Acid
 -39.8935    0.9317   0.5205  -0.1145

Scale estimate of residuals:  2.041
```

Total number of observations: 21

Number of observations that determine the LTS estimate: 18



Los “outliers” no se destacan tanto como en el plot anterior.

EJERCICIOS

1. Usar el conjunto de datos Gessell, disponible en la página de internet del curso y realizar los siguientes analisis.
    a) Calcular la regresiones L1, M con pesos Hampel o Bisquare, LMS y LTS. Hallar las sumas de cuadrados de residuales y compararlas con la suma de cuadrados residuales del ajuste mínimo cuadrático. Plotear todas las lineas halladas junto con la de mínimos cudrados.
    b) Plotear sus residuales y compararlos con los de la regresión mínimo cuadrática..
2. Usar el conjunto de datos Highway, disponible en la página de internet del curso y realizar los siguientes análisis.
    a) Elegir una variable predictora adecuada (deberia tener “outliers”) y calcular la regresiones L1, M con pesos Hampel o Bisquare, LMS y LTS. Hallar las sumas de cuadrados de residuales y compararlas con la suma de cuadrados residuales del ajuste mínimo cuadrático. Plotear todas las lineas halladas junto con la de mínimos cudrados.
    b) Hallar las regresiones L1, M de Hampel y la regresión LMS usando todas las variables predictoras. Plotear sus residuales y compararlos con los de la regresión mínimo cuadrática..

CAPÍTULO 9

REGRESIÓN NOPARAMÉTRICA

9.1 Introducción

La suavización de un conjunto de n datos $\{X_i,Y_i\}$ para i=1,2,…..n. consiste en aproximar la función g en la siguiente relación de regresión.

$$Y_i=g(X_i)+\varepsilon_I \qquad (1)$$

Donde g es la curva de respuesta media, llamada también "signal" y el error aleatorio ε es llamado "noise". En regresión lineal $g(X_i)= \alpha+\beta X_i$ y para efectos de hacer inferencia se asume que ε se distribuye normalmente. Más precisamente, g(x)=E(Y/X) es una media condicional, es decir el promedio de todas las y's para un valor dado de X, donde (X,Y) no necesariamente aparece en la muestra.
En regresión noparamétrica, la forma de la función g y la distribución de los errores es determinada usando los datos que se han tomado.

En general hay dos maneras de atacar el problema:

a) Ajustar los datos localmente (o sea haciendo uso de vecindades o "WINDOWS") a través de modelos bien sencillos. Por ejemplo, aproximar la media condicional por un promedio de los valores observados de y que están en la vecindad del valor de x. La suavización por kernel es un ejemplo de este caso
b) Ajustar un modelo que incluye una parte paramétrica (tal como un modelo polinomial) y otra parte noparamétrica sujeta a una penalidad por complejidad para prevenir el "overfitting". Cuando ocurre "overfitting" es porque se ha tratado de ajustar el modelo más al "noise" que al "signal". La suavización por splines es un ejemplo de este caso.

Ambos métodos requieren la selección de un parámetro de suavización

a) Cuando se usa modelos locales se requiere estimar el ancho de banda ("bandwidth") o ancho de ventana.
b) Cuando se usa estimación penalizada hay que tratar de estimar la penalidad por complejidad, tratando de balancear la bondad de ajuste del modelo y la complejidad del mismo (la complejidad está relacionada al número de paramétros que hay que estimar en el modelo).

Primero consideraremos el caso de regresión noparamétrica cuando hay una sola variable predictora y una sola variable de respuesta y luego el caso de regresión noparamétrica multidimensional donde hay varias variables predictoras y una sola de respuesta. También existe el caso donde hay varias variables de respuesta y varias predictoras, siendo los más conocidos regresión por "projection pursuit" y MARS.

9.2 Suavización bivariada o Suavizadores de diagramas de dispersión (Scatterplot Smoothers)

Entre los métodos más usados están:
i) El Regresorgrama (Tukey, 1961),
ii) "Running means"(Promedios móviles),
iii) "Running line",
iv) Suavización usando los k vecinos más cercanos, K-nn smoothing .
v) Suavizacion por kernels, Nadaraya-Watson (1964)
vi) Regresión local ponderada, LOESS (Cleveland, 1979)
vii) Regresión polynomial,
viii) Suavización por splines, (Wabba, 1975)
ix) Regresión por splines, (Stone and Koo, 1985)

A continuacion se describiran cada uno de estos métodos usando el conjunto de datos **air**. Este es un conjunto de datos de 111 observaciones y 4 variables tomados de un estudio del medio ambiente en donde se midió 4 variables (columnas): ozone, solar radiation, temperature, y wind speed por 111 dias seguidos.

Ozone: surface concentration of ozone in New York, in parts per million.
Radiation: solar radiation
Temperature: observed temperature, in degrees Fahrenheit.
Wind: wind speed, in miles per hour.

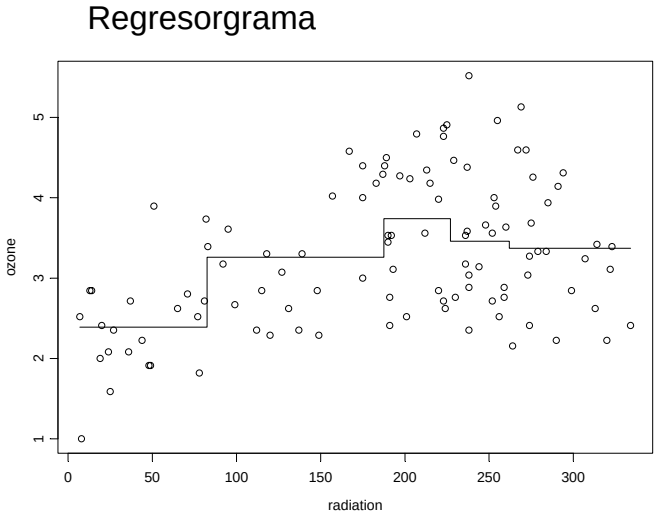
Fuente: John M. Chambers and Trevor J. Hastie, (eds.) Statistical Models in S, Wadsworth and Brooks, Pacific Grove, CA 1992, pg. 348.

9.2.1 El regresorgrama. Aquí se divide el intervalo de los valores de la variable predictora en varios subintervalos (usualmente 5). La amplitud de los subintervalos se elige de tal manera que haya aproximadamente igual numero de datos en cada uno de ellos. Luego se promedia los valores de la variable de respuesta en cada subintervalo. Esto determina varios segmentos de línea que al unirselos forma el regresorgrama. Lo malo de este estimador es que no es suave porque hay saltos en cada punto de corte. A continuación se muestra una función **regresorg** en R que calcula el regresorgrama.

```
regresorg<-function (x,y,k)
{# ************************************************************
# Funcion que calcula el regresorgrama
# Input : El vector x de valores de la variable predoctora
#         El vector y de valores de la variable de respuesta
#         k, el número de subintervalos a usar
#Output:  Los valores de las medias de y en cada  subintervalo y el plot
#         del regresorgrama superpuesto sobre el diagram de puntos
# Edgar Acuna, Mayo 2003
# ************************************************************
n<-length(x)
x<-sort(x)
y<-y[order(x)]
xpoints<-x[1]
nint<-floor(n/k)
```

```
ymeans<-rep(0,k)
for(j in 1:k)
{
ind<-((j-1)*nint+1):(j*nint)
if(j<k)
xpoints<-c(xpoints,x[j*nint])
if(j==k)
{ind<-((j-1)*nint+1):n
xpoints<-c(xpoints,x[n])
}
ymeans[j]<-mean(y[ind])
}
xpoints<-c(xpoints,xpoints[2:k])
xpoints<-sort(xpoints)
ymeans1<-rep(ymeans,each=2)
plot(x,y)
lines(xpoints,ymeans1)
cat("\las medias de y en cada subintervalo son:\n")
ymeans
}
```

En la siguiente figura se muestra el regresograma para los datos del ejemplo1 usando ozone como respuesta y radiation como predictora



9.2.2 "Running Means", "running Medians" y "Running Lines" Aquí para cada valor x_i se define una vecindad simétrico $N(x_i)$ que contenga a dicho punto. La simetría es el sentido que tiene igual número de puntos k tanto a la derecha como izquierda del punto dado, en los extremos esto no se puede lograr, pero se trata de estar lo mas cerca posible. El conjunto de indices de la vecindad simétrica para la observación x_i, varia entre max(i-k,1) hasta min(i+k,n).
Luego se calcula el suavizador por "running means" en el punto x_i de la siguiente manera:

$$s(x_i)=\text{promedio de las } y\text{'s en } N(x_i)$$

El suavizador por "running medians" en el punto x_i está definido de la siguiente manera:

$s(x_i)$=median de las y's en $N(x_i)$

En tanto que el suavizador por "running lines" se calcula por

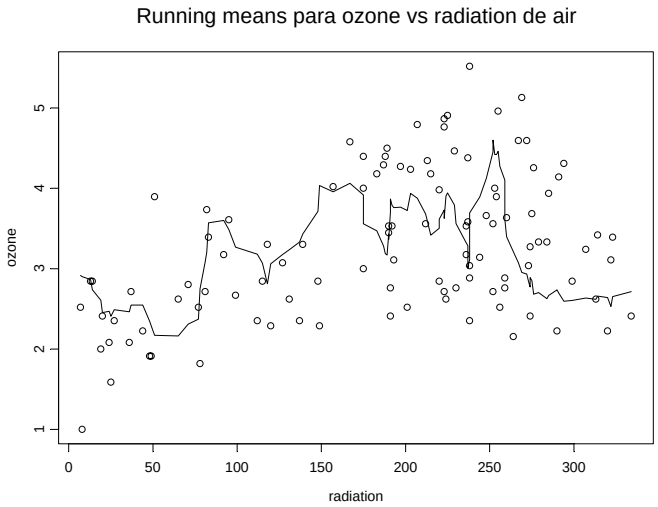
$s(x_i)$=valor estimado de la regresión mínimo cuadrática para $x=x_i$ que se obtiene usando los puntos (x_i,y_i) con x_i que cae en $N(x_i)$.

MINITAB calcula la suavización por "running medians" para datos de series de tiempos igualmente espaciadas usando la secuencia **STAT ▸ ▸ EDA ▸ RSMOOTH**.

A continuación se muestra la función **runmeans** escrita en R que calcula la suavización por "running means" , usando vecindades con k observaciones a cada lado

```
runmeans<-function (x,y,k)
{
#************************************************************
#Funcion que calcula el suavizador por running means
#Inputs: la variable de respuesta y, la variable predictora x y
#       k, el numero de vecinos a cada lado de una observacion
#Output: El plot del suavizador superpuesto en el diagrama de dispersion
#Edgar Acuna, mayo 2003
#************************************************************
n<-length(x)
rm<-rep(0,n)
for(i in 1:n)
{ind1<-max(i-k,1)
ind2<-min(i+k,n)
tempo<-y[ind1:ind2]
rm[i]<-mean(tempo)
}
plot(x,y)

lines(sort(x),rm,type="l")
title("Running means")
}
```

La siguiente grafica muestra el suavizador running means para ozone versus radiation en el conjunto air, usando vecindades con k=3 obervaciones a cada lado del centro.



9.2.3 Suavizador por los k vecinos más cercanos. Aquí para cada valor de x_i se define una vecindad $N_k(x_i)$ que contiene los k valores de x que están más cercanos a x_i. La cercania se determina usando una función distancia (por ejemplo la euclideana). El valor de k generalmente es impar. Luego el suavizador se calcula por

$s(x_i)$=promedio de las y's en $N(x_i)$

La mayoría de los programados estadísticos no tienen incluido este tipo de suavización en su menú. Pero pueden ser programados sin mucha dificultad.

9.2.4 Suavización por kernels.

Considerando que tanto x como y son aleatorias se puede escribir g(x)=E(y/x)= $\int yf(y/x)dy$ donde f(y/x) representa la función de densidad condicional de y dado x. Usando la definición de densidad condicional lo anterior se puede re-escribir como

$$g(x) = \frac{\int yf(x,y)dy}{f(x)} \qquad (2)$$

En la suavización por kernel la función de densidad de x y la función de densidad conjunta de (x,y) son estimadas usando los datos (x_i,y_i) de la muestra. Más específicamente,

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) \qquad (3)$$

y

$$\hat{f}(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{x-x_i}{h})K(\frac{y-y_i}{h}) \qquad (4)$$

Aquí K(t) es llamado el kernel y es una función nonegativa, simétrica con respecto a 0 y con valor máximo en dicho punto. Además, $\int_{-\infty}^{\infty} K(t)dt = 1$. El kernel actúa como una función de peso, que otorga peso grande a los puntos cercanos al punto donde se va a suavizar y bajo peso a los puntos que están alejados del mismo. Hay bastantes funciones que se pueden considerar como Kernel, pero el más usado es el kernel Gaussiano. definido por

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

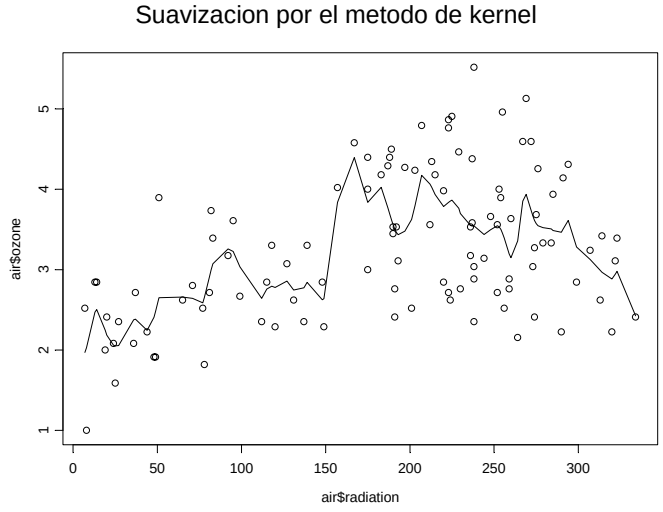
El parámetro h es llamado ancho de banda "bandwidth" y es estimado usando los datos. Hay bastantes propuestas para estimar h. Sustituyendo (3) y (4) en (2) se obtiene la estimación por el método de kernel para g

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K(\frac{x - x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x - x_i}{h})} \qquad (5)$$

Este estimador fue introducido independientemente en 1964 por Nadaraya y Watson.

La función ***ksmooth*** de la librería **modreg** de R halla la suavización basada en kernels. En SAS también se puede obtener este suavizador usando el modulo SAS/INSIGTH. Para estimar el ancho de banda se usa GCV (validación cruzada generalizada). A contnuación se muestran los comandos y la gráfica que da S-Plus.

```
> plot(radiation,ozone)
> lines(ksmooth(radiation,ozone,kernel="normal",bandwidth=5))
>
```



9.2.5 Regresión local ponderada, LOWESS

En este método, si x_o es un punto donde se desea hallar la suavización, entonces primero se halla una vecindad usando los k vecinos más cercanos y luego se halla una regresión ponderada en dicha vecindad el valor ajustado de y en x_o será el valor del suavizador. Más detalladamente el método trabaja así:

i) Se identifican los k vecinos mas cercanos de x_o y se denota la vecindad por $N(x_o)$

ii) Se calcula la distancia a x_o del punto más alejado que está dentro de la vecindad $N(x_o)$ y se lo representa por $\Delta(x_o)$.

iii) Para cada punto x_i en la vecindad $N(x_o)$ se calcula los pesos w_i usando la función peso tri-cúbica definida por:

$$W(t, xo) = [1 - (\frac{|t - xo|}{\Delta(xo)})^3]^3 \text{ siempre que } |t-x_o|<\Delta(x_o)$$

iv) Se define el suvizador s en x_o por:

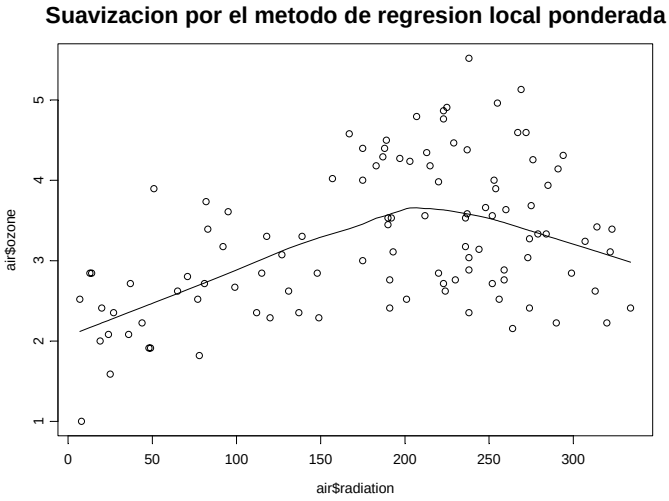
$s(x_o)$=valor ajustado en x_o de la regresión ponderada de y versus x en la vecindad $N(x_o)$, usando los pesos definidos en iii).

Cleveland, también propuso que se podría usar las funciones pesos de la regresión robusta para protegerse de la presencia de outliers.

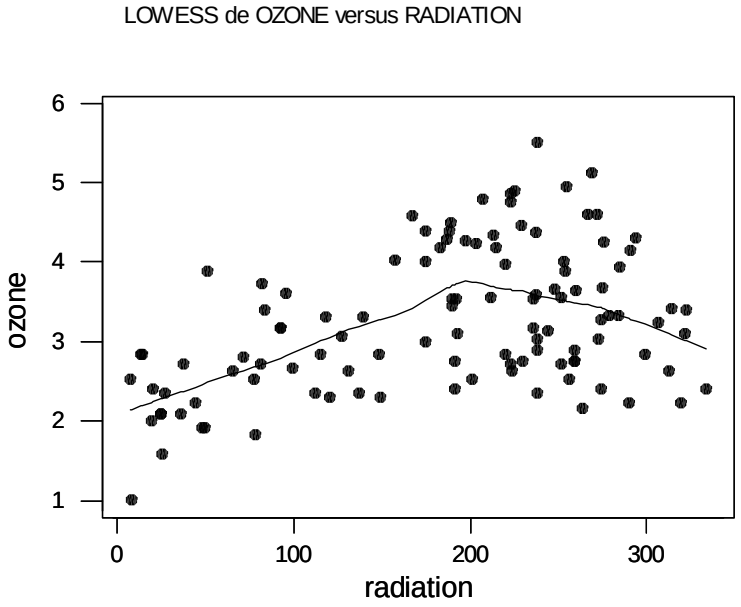
Este suavizador es calculado por la función ***lowess*** (local weighted scatterplot smoother) de S-plus. SAS lo calcula en SAS/INSIGHT y en versión hay el procedimiento LOESS que también lo hace. LOESS es la generalización de LOWESS y permite usar mas de una variable predictora.
MINITAB también calcula la suavización LOWESS en **GRAPH ▸ PLOT ▸ Display ▸ lowess**
A continuación se muestra LOWESS en MINITAB y S-PLUS para el ejemplo 1.

S-PLUS:

```
loess(ozone~radiation)
Call:
loess(formula = ozone ~ radiation)

 Number of Observations:         111
 Equivalent Number of Parameters: 4.5
 Residual Standard Error:        0.7447
 Multiple R-squared:             0.33
 Residuals:
    min   1st Q   median  3rd Q   max
 -1.382 -0.5808 -0.05205 0.5523 1.851

> plot(radiation,ozone)
> lines(lowess(radiation,ozone))
>
```



MINITAB:

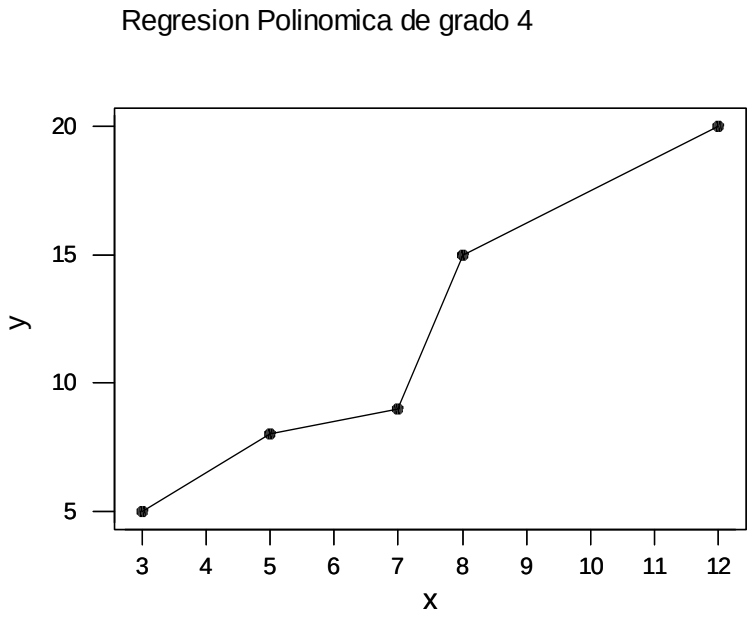


9.2.6 Regresión Polinomial.

Aquí se ajustan los datos (x_i,y_i) para i=1,…n, a un polinomio de la forma

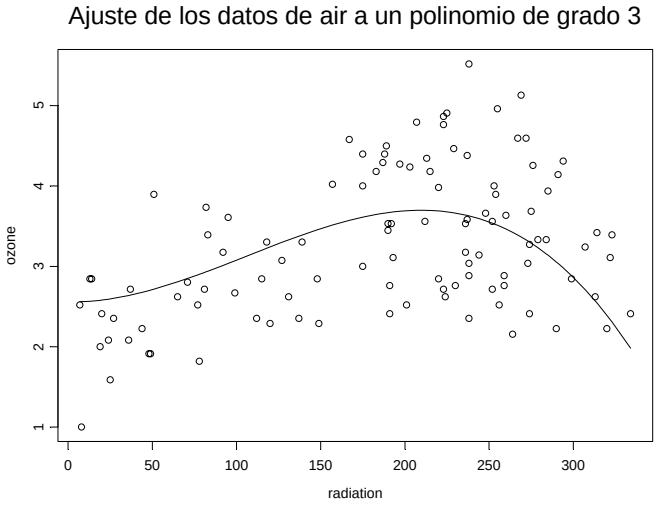
$$y = \beta_o + \beta_1 x + \beta_2 x^2 + ...\beta_k x^k$$

n debe ser mayor que k+1 de lo contrario se tendría un "overfitting" total como lo muestra la siguiente figura



Para obtener una suavización polinomial de grado 3 en S-Plus se ejecuta los siguientes comandos

```
plot(radiation,ozone)
lines(airsort[,1],fitted(lm(airsort[,2]~poly(airsort[,1],3))))
```



9.2.7 Regresión por Splines

Un spline (Schoemberg, 1964) de orden p con k nudos, $t_1,\ldots t_K$ en el intervalo [a,b] es una función que se se obtiene dividiendo primero el intervalo [a,b] en los subintervalos $[x_o,x_1)$, $\ldots\ldots,[x_k,x_{k+1}]$, con x_o=a y x_{k+1}=b y usando luego un polinomio de grado menor o igual que p en cada uno de los subintervalos, además estos pedazos polinomiales deben unirse suavemente en cada uno de los nudos. Más formalmente, el spline s(x) está definido por

$$s(x) = \beta_o + \beta_1 x + ...\beta_p x^p + \sum_{j=1}^{K} \beta_{p+j}(x-t_j)_+^p + \varepsilon \qquad (6)$$

donde β_0, $\beta_1,\ldots\ldots\beta_{P+1},\ldots,\beta_K$ son constantes a determinar, y

$$(t-x)_+^p = \begin{cases} (t-x)^p & t > x \\ 0 & t \le x \end{cases}$$

es llamada la función potencia truncada de orden p

En particular el spline cúbico está dado por

$$s(x) = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{K} \beta_{2+j}(x-t_j)_+^3 + \varepsilon \qquad (7)$$

Un spline es llamado un natural natural de orden 2m con nudos en $x_1,\ldots x_k$ si además de lo arriba mencionado el suavizador es un polinomio de grado m fuera del intervalo $[x_1,x_k]$

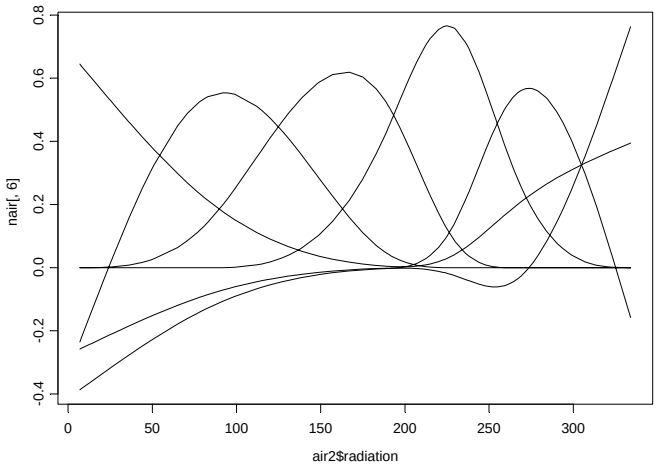
Las funciones 1, $x,x^2,\ldots,x^p,(x-t_1)_+^p,\ldots.(x-t_k)_+^P$ forman una base de funciones del spline. Lamentablemente esta base tiende a crear problemas de multicolinealidad, por la que se recomienda explorar otras bases. Una alternativa son los B-splines cuya base de funciones son calculadas recursivamente (ver Boor, 1978).

Las funciones *ns* y *bs* de S-Plus calculan las bases de funciones par el spline natural y el B-spline respectivamente. Los nudos son tomados como los quantiles. Es decir, si hay un solo nudo éste sería la mediana. Si hay dos nudos entonces estos son los percentiles del 33% y 66%. Si hay 3 nudos estos son los cuartiles y así sucesivamente. El grado de la parte polinomial es tomado por "default" como igual a 3. Si se usa natural spline la relación entre nudos y grados de libertad está dada por nudos=gl-1-intercepto, donde intrecepto es igual a 1 si se considera el intercepto y a cero si no se considera el intercepto. Para los B-splines la relación es nudos=gl-grado-1, donde grado=3 por "default".En la siguientes gráficas se muestran los 6 natural splines y los 6 B-Splines para ajustar ozone versus radiation

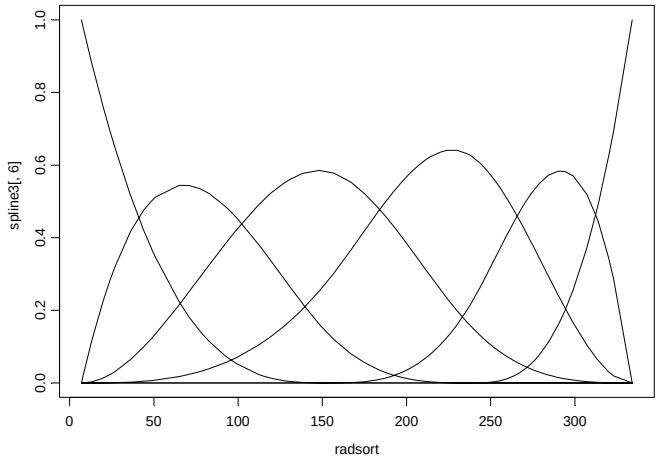
```
>air2<-sort.col(air1,c(1,2),2)
>nair<-ns(air2$radiation,df=6,intercept=T)
>plot(air2$radiation,nair[,6],type="l")
> lines(air2$radiation,nair[,5])
> lines(air2$radiation,nair[,4])
> lines(air2$radiation,nair[,3])
> lines(air2$radiation,nair[,2])
> lines(air2$radiation,nair[,1])

>bair<-bs(air2$radiation,df=6,intercept=T)
```

```
>plot(air2$radiation,bair[,6],type="l")
> lines(air2$radiation,bair[,5])
> lines(air2$radiation,bair[,4])
> lines(air2$radiation,nair[,3])
> lines(air2$radiation,bair[,2])
>lines(air2$radiation,bair[,1])
```

Las 6 funciones bases para N splines usando radiation



Las 6 funciones base para B-splines de radiation

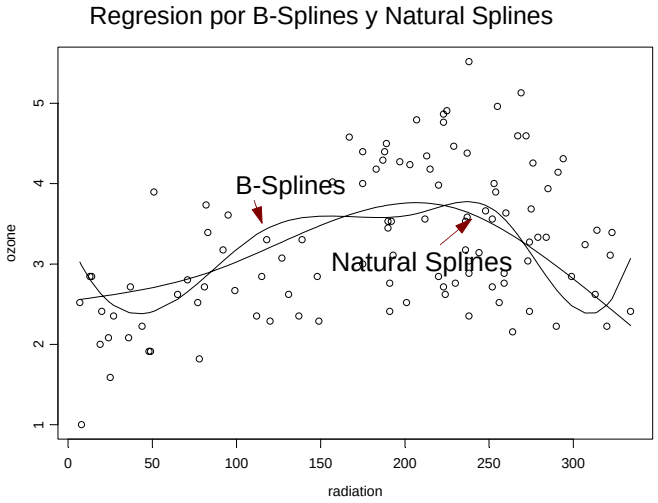


La regresión por spline usando k nudos t_j se define por:

$$y = \beta_o + \beta_1 x + ...\beta_p x^p + \sum_{j=1}^{K} \beta_{p+j} (x - t_j)_+^p + \varepsilon \qquad (8)$$

El modelo puede ser linealizado mediante transformaciones y hay que estimar p+K+1 parámetros. El problema es determinar el número de nodos K. La idea básica es añadir el máximo número de nudos posibles y luego ir eliminado uno por uno tratando de maximizar la bondad de predicción del modelo y minimizando su complejidad.
Una vez determinada la base de los splines se puede hacer la regresión usando las funcione *lsfit* o *lm* de S-Plus.
A continuación se muestran las regresiones usando spline natural y B-Spline.

```
>plot(radiation,ozone)
> lines(airsort[,1],fitted(lm(airsort[,2]~ns(airsort[,1],df=5))) )
> lines(airsort[,1],fitted(lm(airsort[,2]~bs(airsort[,1],df=6))))
```



9.2.8 Suavización por Splines

El suavizador por splines se obtiene minimizando

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int[f''(t)]^2\,dt \qquad (9)$$

El primer término es una media de la bondad de ajuste del modelo y el segundo término es una medida del grado de suavidad. El parámetro de suavidad λ es positivo y gobierna el intercambio entre la suavidad y la bondad de ajuste del suavizador. Cuando $\lambda=\infty$ se obtiene una aproximación polinomial y cuando λ=0 se obtiene una regresión por spline.

Considerando que $X_i^t = \{1, X_i, ..., X_i^p, (X_i - t_1)_+^P ....., (X_i - t_k)_+^p\}$

$$\mathbf{X} = \begin{pmatrix} X_1^t \\ . \\ X_n^t \end{pmatrix} \qquad \text{y} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_o \\ . \\ \beta_{k+p} \end{pmatrix}$$

Entonces la ecuación anterior se puede escribir como

$$(\mathbf{y} - \mathbf{X\beta})^t (\mathbf{y} - \mathbf{X\beta\beta} + \lambda \boldsymbol{\beta}^{\mathbf{t}} \boldsymbol{\Omega} \boldsymbol{\beta} \qquad (10)$$

donde $\boldsymbol{\Omega}$ es una matriz tal que $\{\Omega\}_{jk} = \int X_j^{''}(t) X_k^{''}(t) dt$

Reinsch (1967) mostró que existe existe un único mínimo de (9) , y que éste es un spline cúbico natural con knots en los únicos valores de x_i.

Minimizando la expresión (10) con respecto a β se obtiene que

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X'X} + \lambda \boldsymbol{\Omega})^{-1} \mathbf{X'y}$$

que es un resultado bien similar a Regresión Ridge.

Recordando que $f = \mathbf{XB}$ se tendría que

$$\hat{f} = \mathbf{X'(X'X} + \lambda \boldsymbol{\Omega})^{-1} \mathbf{X'y}$$

Aqui la matriz $\mathbf{H}(\lambda) = \mathbf{X'(X'X} + \lambda \boldsymbol{\Omega})^{-1} \mathbf{X'}$ es llamada la matriz "HAT".

Los grados de libertad de la suavización es igual a la traza de **H**(λ). Esto es bastante similar al número de variables predictoras en un modelo de regresión.

Elección del parámetro λ

a) Usando validación cruzada

Sea $s(x; \hat{\boldsymbol{\beta}}(\lambda))$ el spline ajustado con parámetro de suavización λ

Sea $s_{-i}(x; \hat{\boldsymbol{\beta}}(\lambda))$ el spline ajustado con parámetro de suavización λ pero sin usar la observación (x_i, y_i) entonces se define la función de validación cruzada como

$$CV(\lambda) = \sum_{i=1}^{n} \{s_i - s_{-i}(x_i, \hat{\beta}(\lambda)\}^2$$

el valor λ que minimiza CV(λ) es el valor que se escoge como parámetro de suavización.

El problema con CV es que es computacionalmente caro calcularlo. Una major alternativa es usar GCV

b)Usando validación cruzada generalizada (GCV)

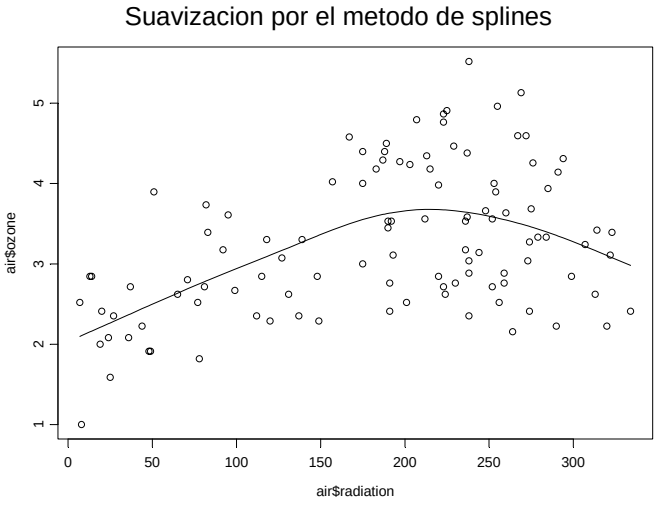
El GCV en realidad no es una generalización del CV sino por el contrario una aproximación. Se define por

$$GCV(\lambda) = \frac{\sum_{i=1}^{n} \{y_i - s(x_i, \hat{\beta}(\lambda))\}^2}{[1 - tr(H(\lambda)/n]^2} \qquad (11)$$

el valor λ que minimiza GCV(λ) es el valor que se escoge como parámetro de suavización. Este es el procedimiento que usan SAS y S-Plus para estimar el parámetro de suavización.

A continuación se muestran los resultados en S-plus

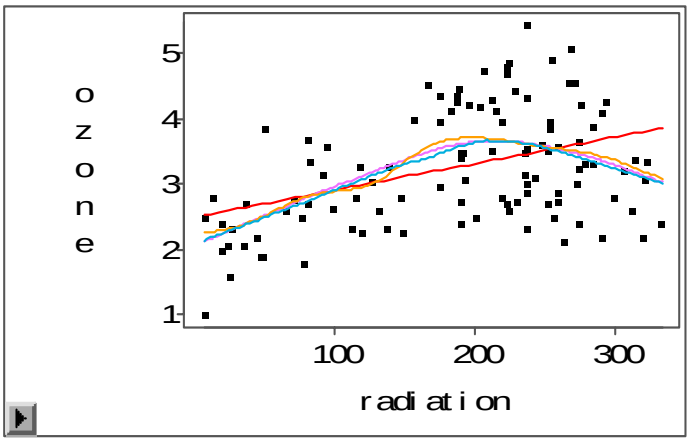
```
>smooth.spline(air$radiation,air$ozone)
Call:
smooth.spline(x = air$radiation, y = air$ozone)

Smoothing Parameter (Spar): 0.01844406
Equivalent Degrees of Freedom (Df): 4.065246
Penalized Criterion: 48.79781
GCV: 0.5747841

> plot(radiation,ozone)
> lines(smooth.spline(radiation,ozone))
```



Ejemplo 1. A continuación se muestran las suavizaciones usando LOESS, kernel, y splines usando SAS para ajustar la relación entre las variables ozone y radiation del conjunto de datos air

Model Equation

ozone = 2.4860 + 0.0041 radiation



Parametric Regression Fit

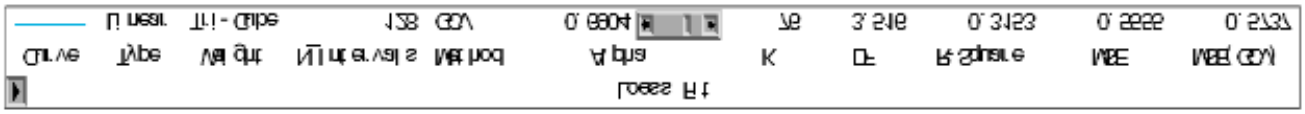
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
	1	1	15.5314	109	0.6576	0.1781	23.62	<.0001

Spline Fit

Curve	Method	C Value	Smoothing Parameter	DF	R-Square	MSE	MSE(GCV)
	GCV	5.2559	15694.0788	3.847	0.3185	0.5546	0.5745

Kernel Fit

Curve	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
	Normal	GCV	0.4846	27.2045	5.158	0.3324	0.5501	0.5769



9.3 Suavización multidimensional

i) Modelos Aditivos generalizados, GAM (Hastie y Tibshirani, 1985)
ii) Regresión por Projection Pursuit, PPR (Friedman, Stuelze, 1981)
iii) Regresión por arboles, CART (Breiman, Friedman, Olsen y Stone, 1984)
iv) Regresión multivariada adaptativa usando Splines, MARS (Friedman, 1991)
v) Esperados Condicionales Alternantes, ACE (Breiman y Friedman, 1985)
vi) Neural Networks (Barron
vii) Wavelets smoothing (Donoho y Johnstone, 1995)

9.3.1 Modelos Aditivos generalizados (GAM)

Un modelo aditivo generalizado es de la forma

$$y=f_1(x_1)+f_2(x_2)+\ldots\ldots f_p(x_p)+e$$

aqui las f_j son estimadas usando cualquiera de los suavizadores bivariados.
El modelo es ajustado usando el algoritmo "local scoring", el cual iterativamente ajusta modelos aditivos ponderados usando "backfitting". El algoritmo "backfitting" es un método de Gauss-Seidel para ajustar modelos aditivos usando residuales parciales de suavización iterativamente

Algoritmo "Backfitting"

1. En el paso inicial se define las funciones $f_j^{(o)} \equiv 1$

2. En la i-ésima iteration, se estima $f_j^{(i+1)}$ por

$$f_j^{(i+1)} = s(y - \sum_{k \neq j} f_k^i(x_k)) \text{ para j=1,….p}$$

3. Cotejar si $| f_j^{(i+1)} - f_j^i | < \delta$ para todo j=1,..,p, donde δ es una constante de tolerancia. Si no se cumple la condición volver al paso 2. En caso contrario parar y usar $f_j^{(i)}$ como f_j en el modelo aditivo.

Si bien terminos de suavización tales como lowess , bs , ns, kernel o k-nn pueden ir mezclados en una fórmula, es más conveiniente usar el mismo suavizador para ahorrar memoria del computador.

S-Plus y R tiene la función **gam** para estimar un modelo aditivo genralizado. Aqui se aplica a los datos del ejemplo 1, usando ozone como variable de respuesta, y radiation y temperature como variables predictoras con regresion splines. La funcion gam de S-Plus es mas general y permite usar otros suavizadores.

```
> gam1<-gam(ozone ~ s(radiation) + s(temperature), data = air)
> gam1
Call:
gam(formula = ozone ~ lo(radiation) + bs(temperature),
          data = air)

Degrees of Freedom: 111 total; 103.433 Residual
Residual Deviance: 29.23443
> attributes(gam1)
$names:
 [1] "coefficients"       "residuals"
 [3] "fitted.values"      "R"
 [5] "rank"               "smooth"
 [7] "nl.df"              "df.residual"
 [9] "var"                "assign"
[11] "terms"              "call"
[13] "formula"            "family"
[15] "nl.chisq"           "y"
[17] "weights"            "iter"
[19] "additive.predictors" "deviance"
[21] "null.deviance"      "contrasts"

$class:
[1] "gam" "glm" "lm"

> gam1$fitted.values
     1      2      3      4      5      6      7      8      9     10     11     12
 2.65671 2.54582 2.773776 2.597299 2.611294 2.393684 2.045584 2.808312 2.645998 2.736994 2.345038
2.512691
     13     14     15     16     17     18     19     20     21     22     23     24
 2.596751  2.458358  2.596977  2.10186  2.094763  2.824071  2.065725  2.284439  1.997066  3.595331
3.432517 3.137973
```

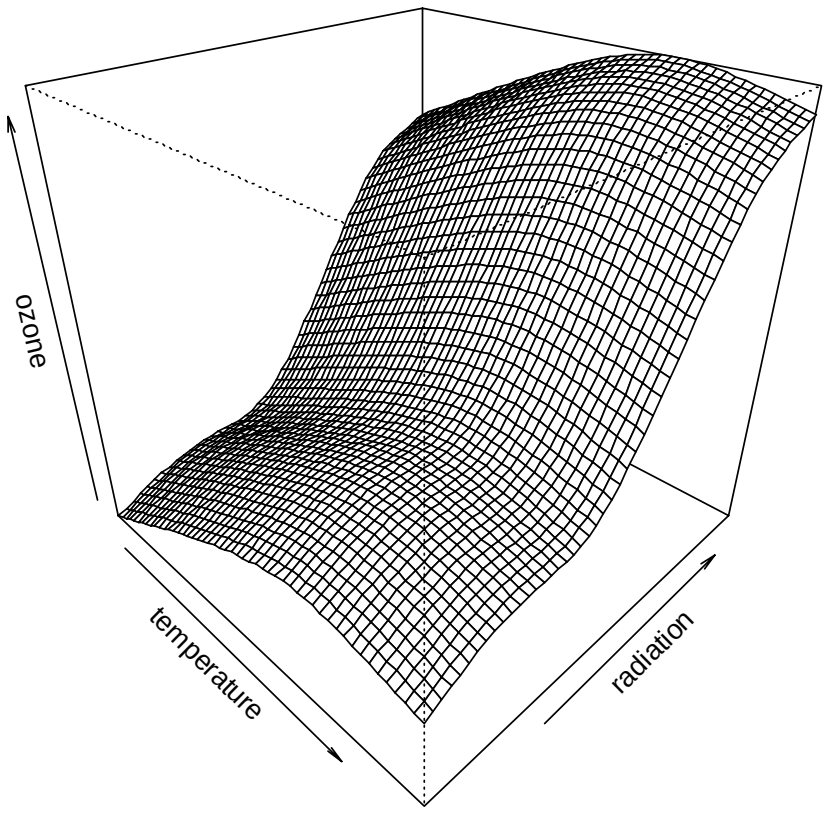
```
      25       26       27       28       29       30       31       32       33       34       35       36
 3.311191  4.253792  3.917691  3.382748  3.177898  2.868866  2.052431  2.610131  2.868028  3.814197
3.947865 3.606233
      37       38       39       40       41       42       43     44     45       46       47       48
 3.586038 3.599523 4.138319 4.472126 4.461953 4.094639 2.979593 3.4159 2.86658 3.578421 3.632034
3.767283
      49       50       51       52       53       54       55       56       57       58       59       60
 3.988882 3.933699 2.283529 3.914839 3.939381 3.141197 3.487908 4.17284 3.972782 3.76244 3.590106
3.062994
      61       62       63       64       65       66       67       68       69       70       71       72
 2.869332 3.128421 4.267184 4.362023 4.321021 3.93223 3.634388 3.25018 2.70473 2.795085 2.785039
3.365681
      73       74       75       76       77       78       79       80       81       82       83       84
 3.248156  3.261354  2.265019  3.409377  3.607601  4.00857  4.772059  4.647696  4.774663  4.528619
4.213954 4.44216
      85       86       87       88       89       90       91       92       93       94       95       96
 4.436223  4.46699  3.613728  3.347621  3.512145  3.347788  3.141048  2.992318  3.606233  3.188219
3.287716 2.173835
      97       98       99      100      101      102      103      104      105      106      107      108
 2.469837  3.363355  2.76363  2.486444  2.797151  3.630543  2.746346  2.140027  3.267639  2.174337
2.003213 2.770275
     109      110      111
 3.035249 2.847853 2.78659
>
gam1$coefficients
 (Intercept) lo(radiation) bs(temperature)1 bs(temperature)2 bs(temperature)3
    2.777342       2.448357        -1.033315         1.147127          1.8088
```

También se puede usar la función predict.gam para predecir nuevos valores de la variable de respuesta.

A continuación se presenta un programa para hacer un plot de la superficie estimada por el modelo aditivo generalizado.

```
gtemp<-seq(min(air$temperature),max(air$temperature),length=50)
  gradiation<-seq(min(air$radiation),max(air$radiation),length=50)
  grid1<-list(radiation=gradiation,temperature=gtemp)
  grid1<-expand.grid(grid1)
estimado1<-predict.gam(gam1,grid1)
  matest1<-matrix(estimado1,50,50)
  persp(gtemp,gradiation,matest1, theta=45, phi=30, xlab="temperature", ylab="radiation",
zlab="ozone")
```

9.3.2 Regresión usando árboles de decisión (CART)



En este caso la superfice de regresión es estimada usando el siguiente modelo aditivo

$$s(\mathbf{x}) = \sum_{i=1}^{n} c_i I_{N_i}(\mathbf{x})$$

las c_i son constantes y $I_{Ni}(\mathbf{x})=1$ si $\mathbf{x} \in N_i$ y es igual 0 en otro caso. Los N_i son hiperrectangulos disjuntos con lados paralelos a los ejes coordenados. Los hiperrectangulos son construidos por partición recursiva y pueden ser representados como un árbol.

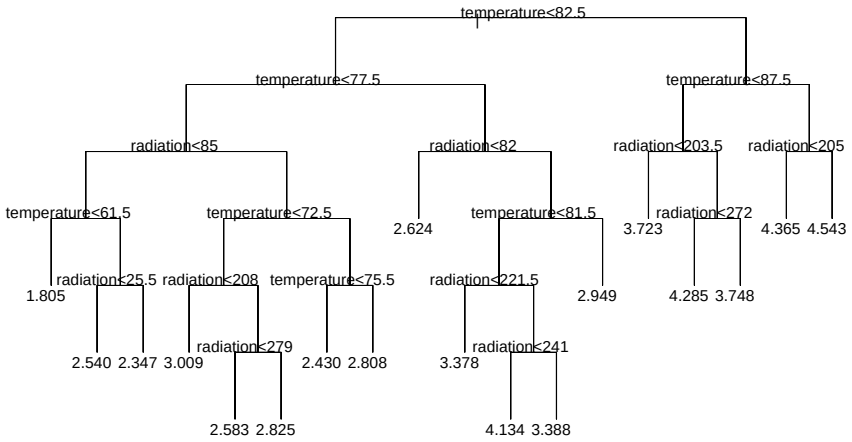
Ejemplo: Obtener la regresión usando arboles de decision para el conjunto de datos *air*

```
> arbol<-tree(ozone~radiation+temperature,data=air)
> arbol
node), split, n, deviance, yval
      * denotes terminal node

 1) root 111 87.2100 3.248
```

```
  2) temperature<82.5 77 35.3700 2.828
    4) temperature<77.5 50 11.0600 2.572
      8) radiation<85 16  3.8930 2.238
       16) temperature<61.5 5  1.3990 1.805 *
       17) temperature>61.5 11  1.1310 2.435
         34) radiation<25.5 5  0.4135 2.540 *
         35) radiation>25.5 6  0.6169 2.347 *
      9) radiation>85 34  4.5340 2.730
       18) temperature<72.5 19  2.6920 2.807
         36) radiation<208 6  0.8644 3.009 *
         37) radiation>208 13  1.4700 2.714
           74) radiation<279 6  0.1971 2.583 *
           75) radiation>279 7  1.0840 2.825 *
       19) temperature>72.5 15  1.5840 2.631
         38) temperature<75.5 7  0.4731 2.430 *
         39) temperature>75.5 8  0.5792 2.808 *
    5) temperature>77.5 27 15.0100 3.301
     10) radiation<82 5  2.4310 2.624 *
     11) radiation>82 22  9.7670 3.454
       22) temperature<81.5 17  7.7900 3.603
         44) radiation<221.5 7  0.9025 3.378 *
         45) radiation>221.5 10  6.2840 3.761
           90) radiation<241 5  4.0340 4.134 *
           91) radiation>241 5  0.8587 3.388 *
       23) temperature>81.5 5  0.3218 2.949 *
  3) temperature>82.5 34  7.4700 4.199
    6) temperature<87.5 17  4.0370 3.929
     12) radiation<203.5 6  0.7499 3.723 *
     13) radiation>203.5 11  2.8940 4.041
       26) radiation<272 6  1.5190 4.285 *
       27) radiation>272 5  0.5891 3.748 *
    7) temperature>87.5 17  0.9438 4.470
     14) radiation<205 7  0.1251 4.365 *
     15) radiation>205 10  0.6890 4.543 *
> summary(arbol)

Regression tree:
tree(formula = ozone ~ radiation + temperature, data = air)
Number of terminal nodes:  18
Residual mean deviance:  0.1919 = 17.85 / 93
Distribution of residuals:
   Min. 1st Qu.   Median       Mean 3rd Qu.  Max.
 -0.959 -0.2394 -0.01998 -7.001e-017  0.2474 1.384
>
```

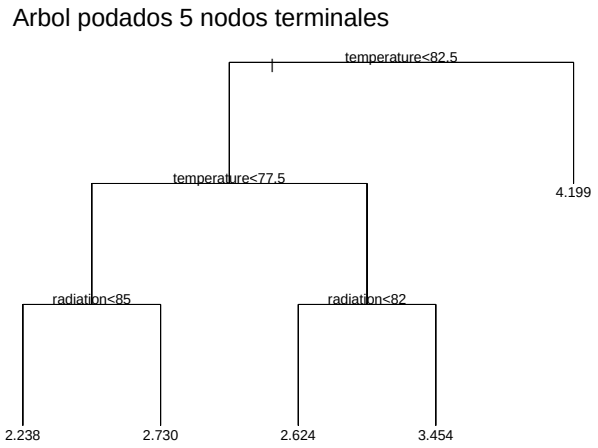
El siguiente es un ejemplo de árbol obtenido con la función plot.tree de S-Plus

```
>plot.tree(arbol, type="u")
> text(arbol)
```



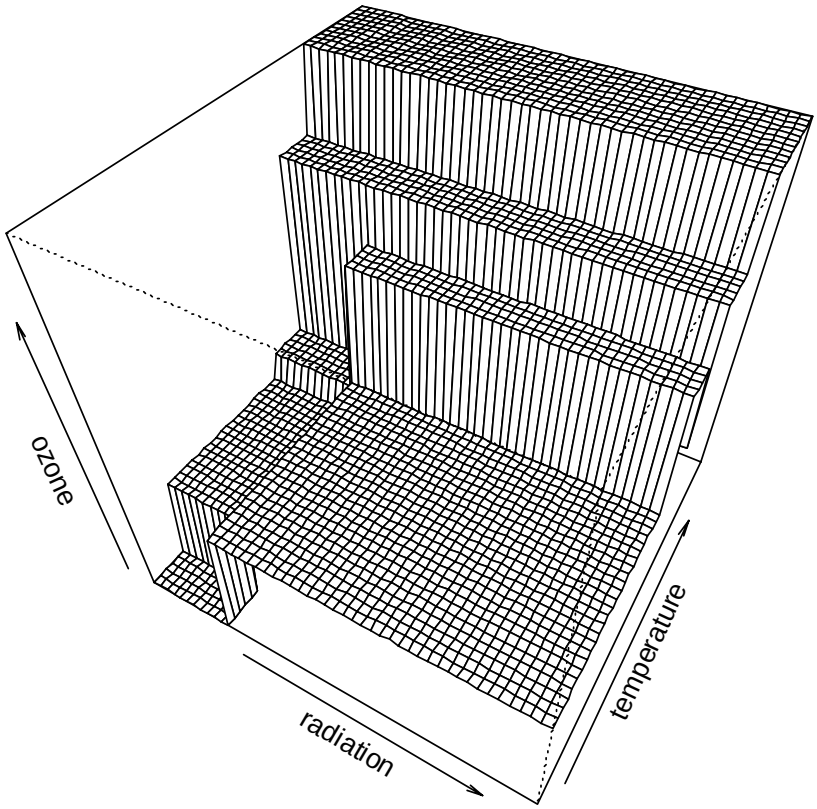
Se puede recortar el arbol en forma similar a hacer seleccion de variables usando la función prune.tree.

```
 mejorarbol<-prune.tree(arbol,best=5)
> mejorarbol
node), split, n, deviance, yval
      * denotes terminal node

 1) root 111 87.210 3.248
   2) temperature<82.5 77 35.370 2.828
     4) temperature<77.5 50 11.060 2.572
       8) radiation<85 16  3.893 2.238 *
       9) radiation>85 34  4.534 2.730 *
     5) temperature>77.5 27 15.010 3.301
      10) radiation<82 5  2.431 2.624 *
      11) radiation>82 22  9.767 3.454 *
   3) temperature>82.5 34  7.470 4.199 *
> plot.tree(mejorarbol, type="u")
> text(mejorarbol)
```



Para obtener una superficie de la suavización por árboles se ejecuta los siguientes comandos

```
> gtemp<-seq(min(air$temperature),max(air$temperature),length=50)
>  gradiation<-seq(min(air$radiation),max(air$radiation),length=50)
>  grid<-cbind(gtemp,gradiation)
>  grid1<-list(radiation=gradiation,temperature=gtemp)
>  grid1<-expand.grid(grid1)
>  estimado<-predict.tree(arbol,grid1)
>   matest<-matrix(estimado,50,50)
>   persp(gradiation,gtemp,matest, theta=30, phi=45, xlab="radiation", ylab="temperature",
zlab="ozone")
```



Para hacer predicciones para nuevos datos se usa la función *predict.tree.*

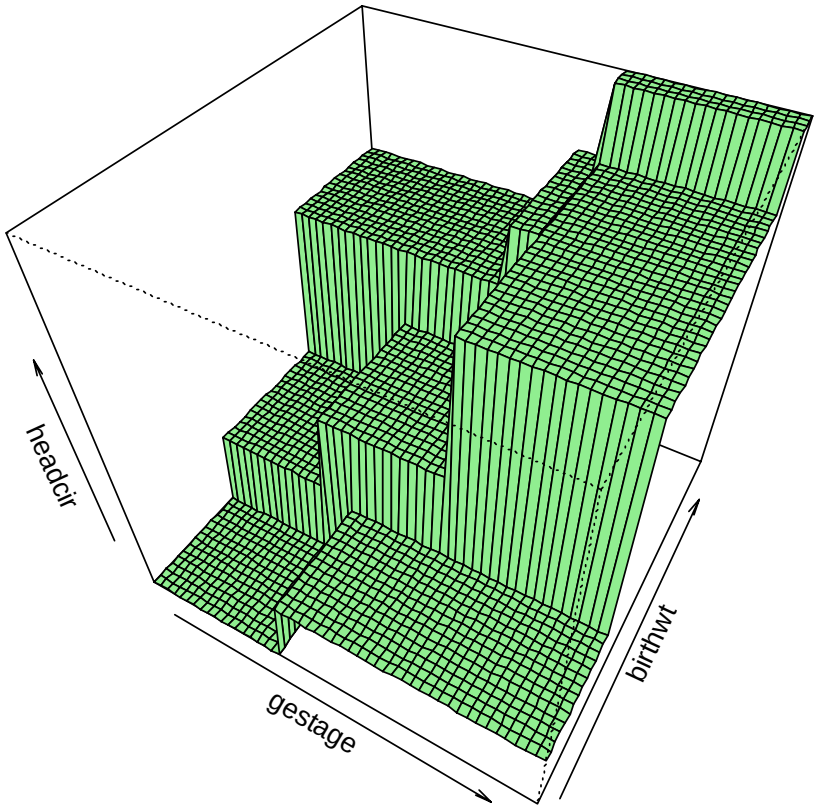
Ejemplo: Obtener la regresión usando arboles de decision para el conjunto de datos *headcirc.*

Usando R se obtiene:

```
> library(tree)
>   arbol<-tree(headcirc~birthwt+gestage,data=headcirc)
>   arbol
node), split, n, deviance, yval
      * denotes terminal node

 1) root 100 634.800 26.45
   2) birthwt < 840 21  25.810 23.10
     4) gestage < 27.5 15   9.333 22.67 *
     5) gestage > 27.5 6   6.833 24.17 *
```

```
   3) birthwt > 840 79 309.800 27.34
     6) gestage < 30.5 54  97.500 26.50
      12) birthwt < 1180 30  27.200 25.60
        24) gestage < 26.5 5   2.800 24.20 *
        25) gestage > 26.5 25  12.640 25.88 *
      13) birthwt > 1180 24  15.630 27.63
        26) gestage < 29.5 17   7.529 27.29 *
        27) gestage > 29.5 7   1.714 28.43 *
     7) gestage > 30.5 25  91.360 29.16
      14) birthwt < 1430 15  63.330 28.67 *
      15) birthwt > 1430 10  18.900 29.90 *
>  summary(arbol)

Regression tree:
tree(formula = headcirc ~ birthwt + gestage, data = headcirc)
Number of terminal nodes:  8
Residual mean deviance:  1.338 = 123.1 / 92
Distribution of residuals:
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.667e+00 -6.667e-01 -3.333e-02  2.842e-16  5.714e-01  6.333e+00
>   win.graph()
>  plot.tree(arbol, type="u")
>   text(arbol)

Obteniendo un mejor árbol de tamaño 5.

> mejorarbol<-prune.tree(arbol,best=5)
>   mejorarbol
node), split, n, deviance, yval
      * denotes terminal node

 1) root 100 634.80 26.45
   2) birthwt < 840 21  25.81 23.10 *
   3) birthwt > 840 79 309.80 27.34
     6) gestage < 30.5 54  97.50 26.50
      12) birthwt < 1180 30  27.20 25.60
        24) gestage < 26.5 5   2.80 24.20 *
        25) gestage > 26.5 25  12.64 25.88 *
      13) birthwt > 1180 24  15.63 27.63 *
     7) gestage > 30.5 25  91.36 29.16 *
> win.graph()
>   plot.tree(mejorarbol, type="u")
>   text(mejorarbol)

> ggest<-seq(min(headcirc$gestage),max(headcirc$gestage),length=50)
>  gbw<-seq(min(headcirc$birthwt),max(headcirc$birthwt),length=50)
>  grid<-cbind(ggest,gbw)
>  grid1<-list(gestage=ggest,birthwt=gbw)
>  grid1<-expand.grid(grid1)
>  estimado<-predict.tree(arbol,grid1)
>   #grid2<-as.data.frame(grid1)
```

```
> matest<-matrix(estimado,50,50)
> persp(ggest, gbw, matest, theta=30, phi=45, xlab="gestage", ylab="birthwt",
zlab="headcir",col="lightgreen")
```



EJERCICIOS

1. Considerar el conjunto Berkeley, disponible en la página de internet del curso. Elegir una variable predictora y hallar las suavizaciones por kernel, lowess y splines. Calcular en cada caso la suma de cuadrados de los residuales y plotear las curves suavizadas

2. Hacer un programa de preferencia en R que haga la suavización por "running lines" (k=5) y aplicarlo al conjunto de datos Highway.

3. Hacer un programa de preferencia en R que haga la suavización por los k vecinos más cercanos (k=3) y aplicarlo al conjunto de datos Fuel.

Apéndice A. Repaso de Matrices

1.-Definición: Una matriz es una arreglo rectangular de números reales dispuestos en filas y columnas. Una matriz com m filas y n columnas se dice que es de orden m x n de la matriz. Cuando m=n la matriz es llamada matriz cuadrada. Una matriz usualmente es denotada por una letra mayúscula y un elemento cualquiera de ella es llamado una entrada de la matriz. Así $A=(a_{ij})$ representa a la matriz A y a_{ij} es la entrada en la fila i columna j.

Ejemplo 1: La matriz

$$A=\begin{bmatrix} 1 & 4 & 9 \\ 0.5 & 5 & 7 \\ 2 & 6 & 11 \end{bmatrix}$$

La entrada $a_{21}=0.5$ y la entrada $a_{32}=6$.

Cuando la matriz tiene una sola columna es llamado un **vector columna** y si la matriz tiene una sola fila es llamado un **vector fila**. Así una matriz de orden mxn se puede descomponer en m vectores filas o n vectores columnas. El número de elementos del vector es llamada la **dimensión** del vector.

Ejemplo 2. Escribir un vector fila y un vector columna de la matriz A

Solución:

$$\mathbf{a}=[2\ 6\ 11] \quad \text{y} \quad \mathbf{b}=\begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

Los elementos de un vector solo tienen un subíndice el cual indica su posición.

2.- Operaciones con matrices:

Se pueden sumar y restar matrices siempre que éstas sean del mismo orden. Para obtener la matriz suma (resta) simplemente se suman y restan las entradas correspondientes.
O sea, $A+B=(a_{ij}+b_{ij})$.

Multiplicación de dos vectores. El producto (interno o escalar) de dos vectores de igual dimension se obtiene sumando los productos de sus correspondientes elementos. Más especificamente, si $\mathbf{a}=(a_1,\ldots..a_n)$ y $\mathbf{b}=(b_1\ldots.b_n)$ entonces $\mathbf{ab}=a_1b_1+\ldots..+a_nbn$

Notar que la multiplicación de dos vectores produce un número y no un vector.

Ejemplo 3. Hallar el producto de los vectores **a** y **b** del ejemplo 2.

Solución: ab=(2)(4)+(6)(5)+(11)(6)=8+30+66=104.

Multiplicación de matrices. Para que dos matrices se puedan multiplicar el número de columnas de la primera debe coincidir con el numero de filas de la segunda. Así una matriz de orden 5x3 se puede multiplicar con una matriz de orden 3x4, pero no con una matriz 4x4.
Sea A de orden mxn y B de orden nxq entonces el producto AB=C en donde C es una matriz mxq cuya entrada en la posición (i,j) se obtiene multiplicando la i-ésima fila de A con la j-ésima columna de B.
Se debe notar que AB≠BA

Ejemplo 4. Calcular AB si A es la matriz del ejemplo 1 y $B=\begin{bmatrix} 1 & 0 \\ 2 & 3 \\ 5 & 4 \end{bmatrix}$

Solución:

$$C=AB=\begin{bmatrix} 54 & 48 \\ 45.5 & 43 \\ 69 & 62 \end{bmatrix}$$

Por ejemplo la entrada c_{21}=45.5 se obtuvo multiplicado el vector $\mathbf{a_2}$=[0.5 5 7] con $\mathbf{b_1}=\begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$.

Nota: La división por matrices no está definida

Transpuesta de una matriz: La transpuesta de una matriz se obtiene intercambiando sus filas por sus columnas. La transpuesta de la matriz A se representa por A'.

Ejemplo 5. Hallar la transpuesta de la matriz B

Solución

$$B'=\begin{bmatrix} 1 & 2 & 5 \\ 0 & 3 & 4 \end{bmatrix}$$

La transpuesta de un vector columna da un vector fila. Para ser compatible cpn el producto de matrices, el producto de dos vectores (columnas) **a** y **b** de igual dimension es representado más adecuadamente por **a'b.**

Propiedades.

i) (A+B)'=A'+B'
ii) (AB)'=B'A'

Ejemplo 6. Verificar la propiedad (ii) usando R y las matrices

$$A=\begin{bmatrix} 3 & 8 \\ 4 & 9 \\ 5 & 12 \end{bmatrix} \text{ y } C=\begin{bmatrix} 5 & 1 \\ 3 & 7 \end{bmatrix}$$

Solución:

```
> A<-c(3,4,5,8,9,12)
> A<-matrix(A,3,2)
> A
     [,1] [,2]
[1,]    3    8
[2,]    4    9
[3,]    5   12
> # Hallando la transpuesta de A
> B<-t(A)
> B
     [,1] [,2] [,3]
[1,]    3    4    5
[2,]    8    9   12
> C<-c(5,3,1,7)
> C<-matrix(C,2,2)
> C
     [,1] [,2]
[1,]    5    1
[2,]    3    7
> #Multiplicando A por C
> A%*%C
     [,1] [,2]
[1,]   39   59
[2,]   47   67
[3,]   61   89
> # Calculando transpuesta de A por C
> t(A%*%C)
     [,1] [,2] [,3]
[1,]   39   47   61
[2,]   59   67   89
> #Multiplicando C' por A'
> t(C)%*%t(A)
     [,1] [,2] [,3]
[1,]   39   47   61
```

[2,] 59 67 89

3. Norma (euclideana) de un Vector. Dado el vector n –dimensional **a**=(a_1,…a_n) entonces su norma euclideana se define por

$$||\mathbf{a}||=\sqrt{a_1^2+....+a_n^2}=\sqrt{\mathbf{a'a}}$$

En general si p es un número real mayor o igual que uno se define la norma p del vector a por

$$||a||_p=(|a_1|^p+|a_2|^p+....+|a_n|^p)^{1/p}$$

Si p=1 se obtiene la norma Manhattan y si p=∞ se obtiene la norma Chevyschev.

4. Normas de matrices: Dada una matriz A de orden m x n, y una norma vectorial ||.|| se define la norma p de una matriz por

$$||A||_p=\max_{x\neq 0}\frac{||\mathbf{Ax}||_p}{||\mathbf{x}||_p}$$

En particular, para p=1 se obtiene $||A||_1=\max(\sum_{i=1}^{m}|a_{ij}|)$, la mayor de la suma de las columnas, para p=∞ se obtiene $||A||_1=\max(\sum_{j=1}^{n}|a_{ij}|)$, la mayor de la suma de las filas.

Para p=2, se obtiene,

$$||A||_2=\max_{x\neq 0}\sqrt{\frac{\mathbf{x'A'Ax}}{\mathbf{x'x}}}=(\text{mayor eigenvalue de A'A})^{1/2}$$

También es bastante usada la norma de Frobenius.

5. Matriz Identidad. Es una matriz cuadrada cuyos elementos de su diagonal son todos unos y los que no estan en la diagonal son todos ceros. La matriz identidad de orden n se denota por I_n. Por ejemplo,

$$I_4=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

Propiedad: Si A es de orden mxn entonces $AI_n=A$ y $I_mA=A$.

6. Inversa de una matriz. La inversa de una matriz cuadrada A se representa por A^{-1} y es tal que $AA^{-1}=A^{-1}A=I$. Existen varios métodos de calcular inversas de matrices. El comando **solve** de R calcula la inversa de una matriz.

Ejemplo 7. Calcular usando R la inversa de la matriz A del ejemplo 1 y verificarla.

```
> A
     x1 x2 x3
[1,] 1.0  4  9
[2,] 0.5  5  7
[3,] 2.0  6 11
> inva<- solve(A)
>inva
      [,1]    [,2]     [,3]
x1 -0.81250 -0.6250  1.06250
x2 -0.53125  0.4375  0.15625
x3  0.43750 -0.1250 -0.18750
> inva%*%A

              x1           x2           x3
  x1  1.000000e+00 2.220446e-15  4.218847e-15
  x2 -2.775558e-16 1.000000e+00 -1.471046e-15
  x3  1.665335e-16 6.106227e-16  1.000000e+00

> A%*%inva

              [,1]         [,2]         [,3]
  [1,]  1.000000e+00 3.053113e-16 8.049117e-16
  [2,] -2.775558e-16 1.000000e+00 8.326673e-17
  [3,] -5.551115e-17 4.718448e-16 1.000000e+00
>
```

Propiedad: Si A y B son dos matrices cuadradas entonces

$$(AB)^{-1}=B^{-1}A^{-1}$$

7. Traza de una matriz. Si A es una matriz cuadrada entonces su traza es la suma de los elementos que están en su diagonal.

Ejemplo 8. La traza de la matriz A del ejemplo 1 es 1+5+11=17.

Propiedades.

(i) tr(A+B)=tr(A)+tr(B)
(ii) tr(AB)=tr(BA) siempre que AB y BA puedan efectuarse.

8. Rango de una matriz. Indica el número de columnas (o filas) independientes que tiene una matriz. Algunas veces ocurre que una columna (o fila) de una matriz es una combinación lineal de las otras columnas o fila. Si el rango de la matriz es igual al número de columnas entonces se dice que la matriz es de rango completo.

Propiedad Una matriz cuadrada de rango completo tiene inversa.

9. Matriz Simétrica. Una matriz cuadrada A es simétrica si es igual a su transpuesta. Es decir, al intercambiar filas por columnas se obtiene la misma matriz.

Por ejemplo, la matriz $M=\begin{bmatrix} 1 & 7 & 5 \\ 7 & 4 & 9 \\ 5 & 9 & 2 \end{bmatrix}$ es simétrica. La matriz P=M'M también es simétrica

Propiedad. Si una matriz A es simétrica entonce también lo es su inversa A^{-1}. O sea, si A es simétrica $(A^{-1})'=A^{-1}$.

10. Determinante de una matriz cuadrada.

El determinante de una matriz cuadrada A consiste de la suma de ciertos productos de los elementos de A, cada uno de los productos es multiplicado por +1 o –1 de acuerdo a ciertas reglas. El determinante de la matriz A se representa por |A| o det(A).

Ejemplo 9. El determinante de la matriz A de orden 2x2 est'a dado por

$$\det(A) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Si una matriz tiene determinante cero se dice que es sngular.

Propiedades
i) Si una matriz es singular entonces no tiene inversa
ii) $|AB|=|A||B|$
iii)$|A^{-1}|=1/|A|$

11. Matriz Idempotente. Una matriz A es idempotente si es simétrica y si $A^2=A$

12. Matriz Triangular. Si los elementos debajo de la diagonal de la matriz son todos ceros entonces se dice que es del tipo triangular superior y si los elementos por encima de la diagonal son todos ceros entonces es llamada matriz triangular inferior. Sistemas de ecuaciones lineales asociados con matrices triangulares son fáciles de resolver.
El determinante de una matriz triangular es el producto de los elemntos que están en su diagonal

13. Matriz Ortogonal. Una matriz A cuyos vectores columnas son de norma uno y ortogonales (es decir su producto interno da 0) es llamada una matriz orthogonal. Si una matriz cuadrada A es ortogonal entonces A'A=A'A=I, o equivalentemente A^{-1}=A'.

Propiedad. El determinante de una matriz ortogonal es +1 o –1.

14. Forma Cuadrática. Dado un vector columna **z** de dimensión n y una matriz cuadrada A de dimensión n x n. Entonces,

$$\mathbf{z'Az} = \sum_{i=1}^{n} a_{ii} z_i^2 + 2\sum_{i<j} a_{ij} z_i z_j$$

es llamada una forma cuadrática en **z** con matriz A. Notar que la forma cuadrática es un escalar. Se dice que la matriz A es definida positiva si **z'Az>0** para todo **z≠0** y es semi definida positiva si **z'Az≥0** para todo z, pero **z'Az=0** para algún **z≠0**.

15. **Valores propios y Vectores propios.** Sea A una matriz de orden n x n. Los valores propios $\lambda_1, \lambda_2,\ldots, \lambda_n$ de la matriz A se obtiene resolviendo la ecuación

$$|A-\lambda\mathbf{I}|=0$$

Asociado con el i-ésimo valor propio λ_i, hay un vector $\mathbf{v_i}$ que resulta de resolver

$$(A-\lambda_i\mathbf{I})\mathbf{v}_i=\mathbf{0}$$

Propiedades:

a) Si la matriz A es simétrica entonces todos sus valores propios serán reales.
b) Traza(A)= $\sum_{i=1}^{n} \lambda_i$
c) Si V es una matriz cuyas columnas son los vectores propios de A entonces se cumple que **V'AV**=diag($\lambda_1, \lambda_2,\ldots, \lambda_n$).

16. La descomposición de una matriz en valores singulares (SVD). Sea A una matriz real de dimension m por n. Existen matrices ortogonales U de orden m por m y V de orden n por n tales que:

V'A U=diag($\sigma_1,\sigma_2,\ldots,\sigma_p$) con p=min(m,n) y donde

$\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_p\geq 0$ son llamados los valores singulares de A.

Propiedad: Los valores propios de la matriz A'A son los cuadrados de los valores singulares de A.

Referencias

1. Belsley, D., Kuh, y Welsh, R. (1980). Regression Diagnóstics. John Wiley, New York.

2. Draper, N y Smith, H. (1998). Applied Regression Analysis, Third Edition. John Wiley, New York.

3. Haerdle, W. (1990). Applied nonparametric Regression. Cambridge University Press. New York.

4. Hastie, T. y Tibshirani, R. (1990). Generalized additive models. Chapman and Hall, London.

5. Hosner, D y Lemeshow, S. (2000). Applied Logistic Regression. Second Edition. John Wiley, New York.

6.Myers, R. (1990). Classical and modern regression with applications. Duxbury Press, Belmont, California.

7. Neter, J., Wasserman, W., Kutner, M.H, y Nachtsheim, C. (1996). Applied Linear Statistical Models, McGraw-Hill, Boston

8.Rao, C.R. (1973). Linear Statistical Inference and its applications. John Wiley and Sons, New York.

9. Rawlings, J.O., Sastry, G.P. y Dickey D.A. (1998). Applied Regression Analysis: A Research Tool, Springer-Verlag, New York.

10. Rousseeuw, P. y Leroy A. (1987). Robust Regression and outlier detection . John Wiley. New York.

11 Ryan, T. (1996). Modern Regression Methods. John Wiley, New York.

12. Seber, G.A.F and Lee, A. (2003). Linear Regression Analysis. Second Edition. John Wiley, New York.

13. Weisberg, S. (2005). Applied Linear Regression. Third Edition. John Wiley, New York.

14. Venables, W.N. and Ripley, B.D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer-Verlag, New York.