TEORÍA DE REGRESIÓN Y CORRELACIÓN

Constituye una herramienta estadística básica para muchas aplicaciones sobre todo en el campo de la estimación o predicción de una variable por otra u otras, a través de modelos que al cumplir una serie de condiciones estadísticas permitan inferencias válidas con respecto a su comportamiento. La relación entre variables puede ser funcional o estadística. La funcional es una fórmula matemática, por ejemplo, la ecuación de una recta, el cálculo del área basal de un árbol conocido el dap, el precio de venta de un lote de madera, conocido su precio unitario, etc. Figura 2.1.

	m3	costo
1	1	50000
2	2	100000
3	3	150000
4	4	200000
5	5	250000

En R

```
comad<-read.table("clipboard")
attach(comad)
names(comad)
[1] "m3"    "costo"
plot(m3,costo,type = "b",ylab="costo", xlab="metros cúbicos",main="Relacion funcional costo,
metros cúbicos de madera")
```

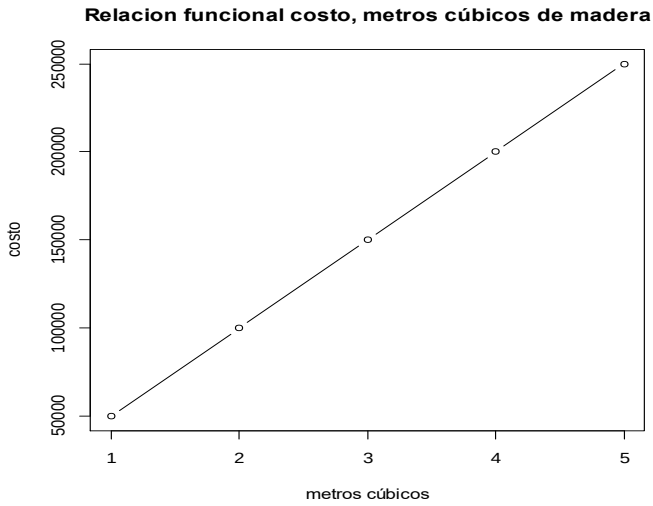


Figura 2.1. Relación funcional de costo de X m^3 de madera a $ 50000 m^3

Todos los puntos de ella caen en la línea de ajuste en forma precisa por lo cual se habla de una relación funcional perfecta $Y = f(x) = 50000\,x$.

A diferencia de la anterior la relación estadística surge de un modelo propuesto a priori por el investigador para unos datos aislados de una población, tomados generalmente al azar. Los datos caen generalmente por fuera de la función (modelo matemático), por lo que se habla de una relación funcional imperfecta.

Por ejemplo, en un bosque se midió el diámetro a la altura del pecho (dap) en cm y la altura total (alt) en m a 10 árboles con el fin de verificar la posible relación estadística

lineal entre estas variables, y poder hacer predicciones si es el caso, de acuerdo con los datos de la Tabla 2.1,

TABLA 2.1. Datos de diámetro (d) y altura (H) para 10 árboles tomados al azar

	dap	alt
a1	23.2	20.1
a2	45.5	29.4
a3	26.3	22.1
a4	39.1	27.2
a5	31.4	25.7
a6	29.6	25.2
a7	25.4	23.8
a8	26.3	18.1
a9	35.2	24.7
a10	29.6	20.2

Con una gráfica de dispersión se apreciaría mejor si tal relación es posible o no. La línea roja sugiere que, es posible.

```
aldi<-read.table("clipboard")
attach(aldi)
names(aldi)
[1] "dap" "alt"


scatter.smooth(dap, alt, main = "Diagrama de dispersion dap alt")
abline(lm(alt~dap))
abline(lm(alt~dap),col="red")
```

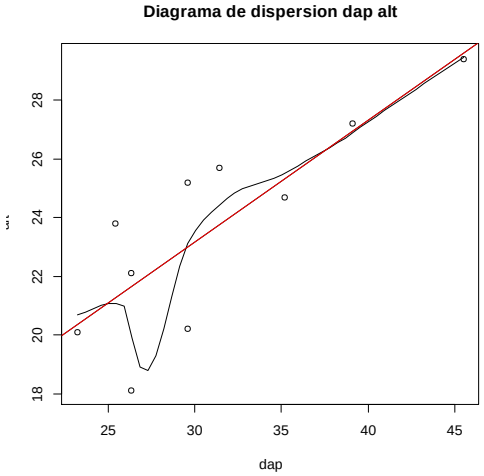


Figura 2.2 Diagrama de dispersión de los datos Tabla 2.1

```
mode1<-lm(alt~dap)
summary(mode1)
Call:
lm(formula = alt ~ dap)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5303 -0.5322  0.0204  1.5801  2.5438

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.7004     3.2531   3.289  0.01103 *
dap           0.4156     0.1021   4.069  0.00359 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 2.127 on 8 degrees of freedom
Multiple R-squared:  0.6742,    Adjusted R-squared:  0.6335
F-statistic: 16.55 on 1 and 8 DF,  p-value: 0.00359
```

Al modelar, los datos se pueden graficar como la función imperfecta mostrada en la Figura 2.3, en la cual se observa sus ubicaciones (datos de campo) con respecto al modelo:

$alt = 10.7004 + 0.4156\, dap$,

obtenido por un método desarrollado posteriormente.

2.1. Regresión.

El análisis de regresión es una de las herramientas más usadas para establecer alguna relación funcional estadística entre una variable dependiente y otra u otras variables explicatorias continuas (Chatterjee y Price, 1977). Como lo vimos la forma de saber si es apropiada es mediante una gráfica de dispersión. Existen varios tipos de ellas

- Regresión lineal, la más usada y simple de todas;
- La regresión polinomial (ante falta de ajustes de la anterior);
- Regresión segmentada (dos o más modelos adyacentes);
- Regresión robusta (viola algunos supuestos y menos sensitiva a observaciones remotas);
- Regresión múltiple (cuando una sola explicatoria no basta);
- Regresión no lineal (ajusta unos datos a un modelo prescrito);
- Regresión no paramétrica (sin una forma funcional obvia).

2.1.1. Regresión lineal. Para su mejor comprensión se planteará el modelo lineal con distribución no especificada de los términos del error como:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i1} + \cdots + \beta_1 X_{i1} + \varepsilon_i; \quad \beta_0 + \sum_{j=1}^{p-1} \beta_1 X_{i1} + \varepsilon_i \quad (2.1)$$

modelo de p-1 variables independientes o modelo con p variables en el cual Y_i es el valor de una variable respuesta para un valor dado de otras variables X_i, conocidas, independientes y constantes llamadas variables independientes, ε_i un término aleatorio para el error que será posteriormente analizado.

2.1.2 Diagrama de dispersión. El primer indicio para intuir una relación lineal entre dos variables se obtiene del gráfico de los pares cartesianos (X, Y) correspondientes a las observaciones relativas de las variables confrontadas (puntos negros):

```
plot(dap,alt,pch=16, main="Diagrama de dispersión altura vs dap")
abline(lm(alt~dap),col="red")
```

En el diagrama de dispersión cada punto representa una observación o tratamiento, generalmente no caen en la línea y la dispersión, alrededor de ella, representa la variación de la altura no asociada con el diámetro, considerada de naturaleza aleatoria.

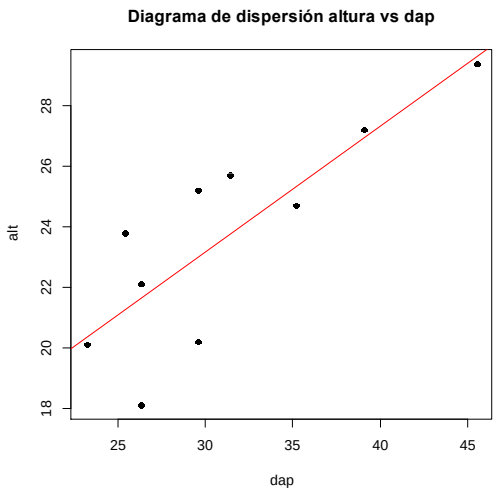


Figura 2.3 Línea ajustada de regresión con datos de la Tabla 2.1.

2.1.3 Ingredientes esenciales de una regresión. Una regresión presenta dos ingredientes esenciales:

1) La tendencia de una variación sistemática de Y con una o más variables independientes (línea roja de la figura 2.2). Ejemplo para los distintos valores de dap, se obtiene en forma sistemática los valores:

```
predichos<-predict(mode1,list=dap)

predichos<-as.data.frame(predichos)
predichos
   predichos
1   20.34196
2   29.60946
3   21.63027
4   26.94973
5   23.74974
6   23.00169
7   21.25624
8   21.63027
9   25.32896
10  23.00169
```

2) La dispersión de las observaciones alrededor de la línea o relación estadística establecida como modelo (puntos negros, figura 2.2). Ambos ingredientes se incorporan al modelo de regresión postulando que:

a) En la población de observaciones asociadas con el proceso muestral hay una distribución de probabilidades de Y para cada nivel de la otra variable X; y;

b) Las medias de estas distribuciones de probabilidades varían en forma sistemática con los cambios de valor de X. Con base en ello es posible la representación de un modelo de regresión lineal, usando algunos puntos del ejemplo anterior, Figura 2.4.

Los usos del modelo permiten generalmente: a) descripción de un fenómeno, b) ejecución de algún tipo de control, c) predicciones, d) verificación de algunas hipótesis y otros que serán explicitados más adelante una vez asumida la estructura teórica.

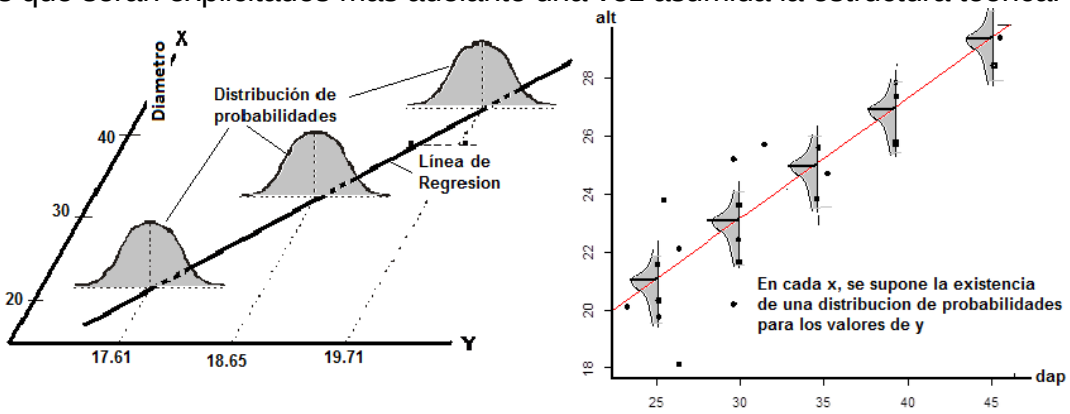


Figura 2.4 Distribuciones de probabilidades de y en cada x. La media cae sobre la línea ajustada

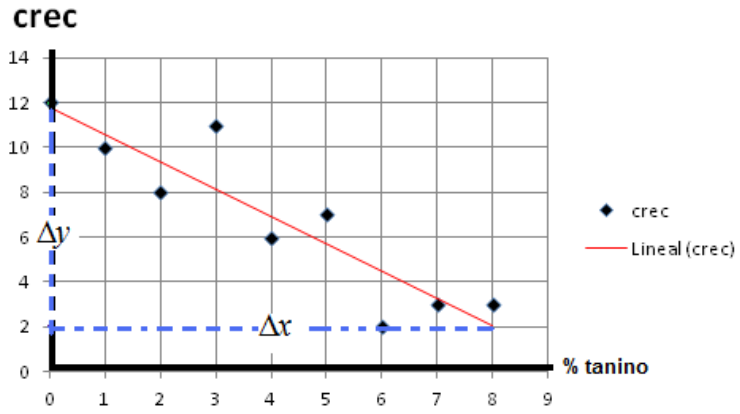
2.2 Regresión lineal simple.

Su esencia es la estimación, a partir de una muestra, de los valores de los parámetros y sus errores estándar del, por ahora, más simple de los modelos: $y = a + bx$, con dos variables continuas (x y y) y dos parámetros (a y b):
y = variable dependiente, o variable respuesta
x = variable independiente o explicatoria
a = intercepto del modelo (equivale a un valor de y cuando x = 0)
b = pendiente del modelo (cambio en y, dividido el cambio en x

Ejemplo sean los datos obtenidos en campo del crecimiento de orugas sometidas a dietas con un contenido diferencial de % de taninos:

X	tan	0	1	2	3	4	5	6	7	8
Y	crec	12	10	8	11	6	7	2	3	3

Que se graficarían así:



Al graficar estos datos, se ve que la línea roja podría ser un modelo adecuado. A mayor porcentaje de tanino en las dietas menor el crecimiento de las orugas.

Se puede tener inicialmente un estimado burdo de los parámetros a ojo: Al incrementarse 8 unidades de tanino (8-0) = 8, se tuvo un cambio de 10 en el crecimiento (2 - 12) = -10. La pendiente b, es entonces el cambio en y dividido por el cambio en x: $b = \frac{\Delta y}{\Delta x} = \frac{-10}{8} = -1.25$. El intercepto es el valor alcanzado por y cuando x = 0, aproximadamente 12. El modelo burdo sería entonces: $y = 12 - 1.25x$

El proceso de regresión no puede ser a ojo, entonces se buscan los mejores estimados de a y b.

La convención moderna usa los llamados, estimados de máxima verosimilitud (*maximum likelihood estimates*) en otras palabras los mejores estimados de unos datos seleccionados y habiendo escogido el modelo lineal, toman como parámetros aquellos que hagan los datos con los valores más probables posibles. La máxima verosimilitud es dada por el método de los mínimos cuadrados cuando se cumplen ciertas asunciones. El concepto o frase "mínimos cuadrados" se refiere entonces a residuales una distancia, d, entre un dato puntual medido en campo y un valor predicho por el modelo para su respectivo x: $d = residual = y - \hat{y}$

Para una mejor comprensión se analizará en primera instancia el *modelo de regresión lineal simple* con distribución no especificada de los términos del error obtenido cuando en la ecuación (2.1), p toma el valor de 2 así:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (2.2)$$

Y_i = valor de la variable llamada dependiente o respuesta en la iésima observación, X_i = constante conocida, valor de una variable j-ésima llamada independiente en el iésimo tratamiento, ε_i término aleatorio del error con $E(\varepsilon_i) = 0$, y varianza $\sigma^2(\varepsilon_i) = \sigma^2$ y además $\varepsilon_i, \varepsilon_j$ no correlacionados de modo que la covarianza $COV(\varepsilon_i \varepsilon_j) = 0$, para todo $i \neq j$.

$ß_0$ y $ß_1$ parámetros estimables por algún procedimiento, y posteriormente caracterizados en el numeral 2.3.2. Este modelo así presentado es: simple (una sola variable independiente), lineal en los parámetros (ninguno aparece de exponente, o multiplicado o dividido por otro), y lineal en la variable independiente (por estar elevada a la potencia 1). Constituye entonces un modelo conocido como de primer orden.

2.3.1 Características importantes del modelo.

1. El valor observado de la Y en la iésima observación es la suma de dos componentes: a) un término constante:

$$\beta_0 + \beta_1 x_i \quad (2.3)$$

b) un término aleatorio ε_i.

2. Puesto que $E(\varepsilon_i) = 0$ entonces

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \quad E(\beta_0 + \beta_1 X_i) + E(\varepsilon_i) = \quad \beta_0 + \beta_1 X_i \qquad (2.4)$$

El modelo de regresión queda entonces así:

$$E(Y) = \beta_0 + \beta_1 X \qquad (2.5)$$

que relaciona la media de las distribuciones de Y para un X dado. Cae sobre la recta de regresión.

3. El valor observado de Y en el iésimo tratamiento excede o queda corto del valor funcional de Y en una cantidad ε_i.

4. Se asume que los términos del error ε_i tienen una varianza constante σ^2 (lo que se conoce como homocedasticidad). O sea, todas las curvas de la figura anterior iguales. Si además la varianza de Y_i es σ^2; sin importar el valor de la X_i ;

$$\sigma^2(Y_i) = \sigma^2(\beta_0 + \beta_1 X_i + \varepsilon_i) = \sigma^2(\varepsilon_i) = \sigma^2 \qquad (2.6)$$

ya que varianza de una constante es cero:

$$\sigma^2(\beta_0 + \beta_1 x_i) = 0 \qquad (2.7)$$

5. Los términos del error se asumen incorrelacionados, es decir el suceso de una observación no afecta los términos del error en otro tratamiento.

Ejemplo: ya se dejó explícito que para los datos de la Tabla 1, el modelo propuesto: $alt = 10.7004 + 0.4156\, dap$ podría describir bien la relación estadística *alt* contra *dap*. Suponga que en un nuevo ensayo, asociado con un *dap* = 31.4 cm se encuentra un valor *alt* = 25.75 m,

```
predict(mode1,list(dap=31.4))
       1
23.74974
> ei=25.7-23.75
> ei<- 25.7-23.75
> ei
[1] 1.95
```

el valor del error será entonces: 1.95. En Figura 2.5. ε_i es simplemente la desviación de Y_i con respecto a su media $E(Y_i)$ y paralelo al eje Y.

2.3.2 Significado de los parámetros de regresión. Los parámetros $ß_0$ y $ß_1$ son llamados coeficientes de regresión. $ß_1$ equivale a la pendiente de la ecuación de la línea recta asociada e indica el cambio promedio en la media de las distribuciones de Y por cada unidad de incremento en X. $ß_0$ es el intercepto y da la media cuando X = 0. Cuando el rango de valores de X no cubre este valor, este parámetro no tiene significado como término separado del modelo de regresión.

2.3.3 Versiones alternativas del modelo. Algunas veces se presenta el modelo en formas equivalentes para lo cual se acude a una variable ficticia $X_0 = 1$

$$Y_i = \beta_0 X_0 + \beta_1 X_1 + \varepsilon_i \qquad (2.9)$$

Otras veces en lugar de las variables se usan sus desviaciones con respecto a sus medias

$$\chi_i = x_i - \bar{x}; \quad \gamma_i = y_i - \bar{y} \quad (2.10)$$

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i = \beta_0^* + \beta_1 \chi_i + \varepsilon_i \quad (2.11)$$

Por ejemplo, si se trabaja con las desviaciones de X

en la cual:

$$\beta_0^* = \beta_0 + \beta_1 \bar{X} \qquad (2.12)$$

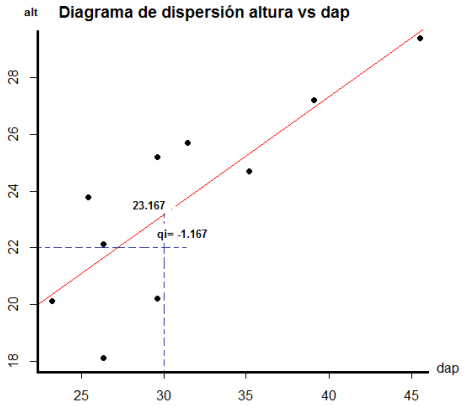
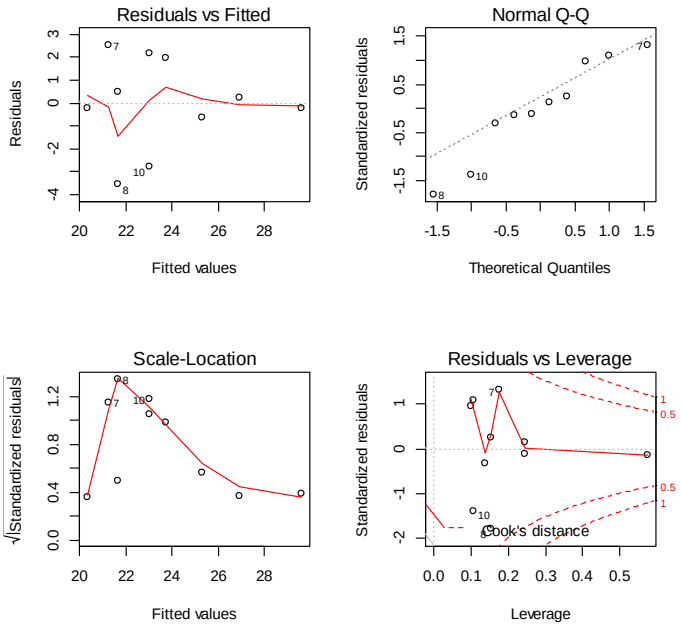


Figura 2.5. Error en ensayo cuya X = 30 cm.

```
par(mfrow=c(2,2))
> plot(model)
```



El primer gráfico muestra los residuales contra los valores ajustados, por ahora, sin distorsiones con respecto al cero, el segundo probable normalidad, el tercero no muestra residuales estandarizados por encima de 1.2 y el cuarto muestra una distancia de Cook, que detecta si alguna observación perturba fuertemente el modelo.

2.3.4 Estimación de la función de regresión. Ordinariamente ß$_0$ y ß$_1$ se desconocen por lo cual es necesario estimarlos por métodos experimentales o no, controlados o no, lo que conduce a los modelos I y II en la regresión.

2.3.4.1 Modelos I y II en regresión. Se habla de Modelo I, o modelo fijo cuando los valores de X son seleccionados de antemano. Ej. al tomar las alturas a unos árboles

cuyos diámetros (dap) sean de 5, 10, 15...cm. que definen las llamadas regresiones de Y en X. El Modelo II aparece cuando ambas variables toman valores aleatoriamente, ej. al intentar establecer relaciones altura diámetro de una parcela. Si además, Y como X pertenecieran a una población normal bivariada (o sea que para cada valor de Y, la X también presenta una distribución normal), sería posible obtener también una relación de X en Y. El modelo II es útil cuando se hable de correlación. Los desarrollos presentados son para el modelo I, aunque con ambas variables aleatorias los métodos de estimación y las pruebas de hipótesis en el II siguen siendo válidas (Johnston, 1963); (Walpole, 1983).

2.3.4.2 Método de los mínimos cuadrados. Los mejores estimadores de los parámetros ß$_0$ y ß$_1$ se obtienen con el método de los mínimos cuadrados en el cual para cada par cartesiano de las observaciones (X_i, Y_i) se consideran las desviaciones q_i de Y_i observado con respecto a su valor calculado por la recta de regresión:

$$q_i = Y_i - (\beta_0 + \beta_1 X_i) \quad (2.13)$$

En particular el método requiere que se considere la suma de n desviaciones de q², que se denotarán como Q:

$$Q = \sum_{i=1}^{n} q_i^2 = \sum_{i=1}^{n} \left[Y_i - \left(\beta_0 + \beta_1 X_i \right) \right]^2 \quad (2.14)$$

en cuyo caso los valores de ß$_0$ y ß$_1$ que hagan Q mínimo darán los mejores estimativos de ellos, lo cual se obtiene tomando las derivadas parciales con respecto a ß$_0$ y ß$_1$ de Q y buscándoles su mínimo:

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)(-1) = -2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right) \quad (2.15)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 * \sum_{i=1}^{n} X_i \left(Y_i - \beta_0 - \beta_1 X_i \right) \quad (2.16)$$

Para algunos valores especiales de ß$_0$ y ß$_1$, llamados b_0 y b_1, ambas ecuaciones se hacen iguales a cero y configuran además un mínimo Q. De (2.15) y (2.16):

$$\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i \right) = 0 \quad \therefore \quad \sum_{i=1}^{n} Y_i = n b_0 + b1 \sum x_i \left(2.17 \right)$$

$$\sum_{i=1}^{n} \left(X_i Y_i - b_0 X_i - b_1 X_i^2 \right) = 0 \quad \therefore \quad \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \quad \left(2.18 \right)$$

(2.17) y (2.18) configuran las llamadas ecuaciones normales con b_0 y b_1 como estimadores puntuales de ß$_0$ y ß$_1$ obtenidas así:

$$b_1 = \frac{\sum X_i Y_i - \frac{\left(\sum X_i \sum Y_i \right)}{n}}{\sum X_i^2 - \frac{\left(\sum X \right)^2}{n}} = \frac{\sum \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum \left(X_i - \overline{X} \right)^2} = \frac{SPCXY}{SCCX} \quad (2.19)$$

$$b_0 = \frac{\left(\Sigma Y_i - b_1 \Sigma X_i \right)}{n} = \overline{Y} - b_1 \overline{X} \quad (2.20)$$

2.3.4.3 Propiedades de los estimadores mínimocuadráticos. El teorema de Gauss-Markov define algunas de las propiedades más importantes de estos estimadores. Neter et al,1983; Wonnacott y Wonnacott, 1981; Walpole y Myers,1984:

1. b_0 y b_1, bajo las condiciones anotadas del modelo trabajado, resultan insesgados y tienen la mínima varianza entre todos los estimadores lineales insesgados o sea que:

$$E(b_0) = \beta_0 \qquad y \qquad E(b_1) = \beta_1 \quad (2.21),$$

por lo cual no dan sobre o subestimaciones sistemáticas.

2. Las distribuciones muestrales de b_0 y b_1 tienen menor variabilidad que cualesquier otros estimadores lineales y por ello el método de mínimos cuadrados es el más preciso de los estimadores lineales insesgados.

2.3.4.4 Residuales. Un residual se define como la diferencia entre un valor observado y su respectivo valor estimado. Se denota como:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_O - b_1 X_i \quad (2.22)$$

cuya magnitud se puede obtener por la distancia vertical entre el valor observado y la línea de regresión (el qi de la Figura 2.3). No se debe confundir el término del error $\varepsilon_i = Y_i - E(Y_i)$ y con el residual e_i ya definido. El primero involucra la desviación vertical de Y_i de una línea desconocida y no ajustada, mientras el segundo con respecto a la línea de regresión ajustada. Los residuales son rectas paralelas al eje Y.

2.3.4.5 Propiedades de la línea ajustada de regresión. La línea de regresión ajustada presenta las siguientes propiedades:

1. La sumatoria de los residuales es cero, o sea:

$$\sum_{i=1}^{n} e_i = 0 = \sum_{i=1}^{n} Y_i - n b_o - b_1 \sum_{i=1}^{n} X_i \quad (2.23)$$

que surge al reemplazar b_0 por la igualdad (2.19) y que coincide además con la primera ecuación normal.

2. La suma de los cuadrados de los residuales es un mínimo:

$$\sum_{i=1}^{n} e_i^2 = minimo \quad (2.24)$$

Este fue uno de los requisitos que debió satisfacerse para derivar los estimadores mínimocuadráticos de los parámetros de la regresión.

3. La suma de los valores observados de Y_i es igual a la suma de los valores ajustados Y_i o sea:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}i \quad (2.25)$$

obtenidas reemplazando el término b_0 de (2.20) involucrado en el segundo de (2.25).

4. Por el numeral anterior $\bar{Y} = \bar{\hat{Y}}$
En efecto:

$$\overline{Y} = \frac{\Sigma Y_i}{n} = \frac{\Sigma \overline{\hat{Y}_i}}{n} = \overline{\hat{Y}_i} \quad (2.26)$$

al dividir por n los términos de (2.25)

5. La suma de los residuales ponderados en cada observación con su respectiva variable independiente (o dependiente estimada) es cero:

$$a) \sum_{i=1}^{n} e_i X_i = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) X_i = \sum_{i=1}^{n} Y_i X_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0 \quad (2.27)$$

por ser la segunda ecuación normal.

$$b) \sum_{i=1}^{n} e_i \hat{Y}_i = \sum_{i=1}^{n} e_i (b_0 + b_1 X_i) = b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} e_i X_i = 0 \quad (2.28)$$

6. La línea de regresión siempre pasará por $(\bar{X}, \bar{Y})$. Usando la versión alternativa del modelo (2.11) y el estimado de ß$_0$* del numeral 2.3.3 se tiene para x promedio:

$$b_0^* = b_0 + b_1 \bar{X} = \bar{Y} - b_1 \bar{X} + b_1 \bar{X} \rightarrow \hat{Y} = \bar{Y} + b_1 (\bar{X} - \bar{X}) = \bar{Y} \quad (2.29)$$

7. Las distribuciones de Y y e son idénticas excepto por sus medias que difieren. En efecto, la distribución de $\mathbf{e_i}$ es justamente la distribución de Y trasladada a su media cero (Wonnacott y Wonnacott, 1981). Para enfatizarlo se puede apreciar **Figura 2.6**.

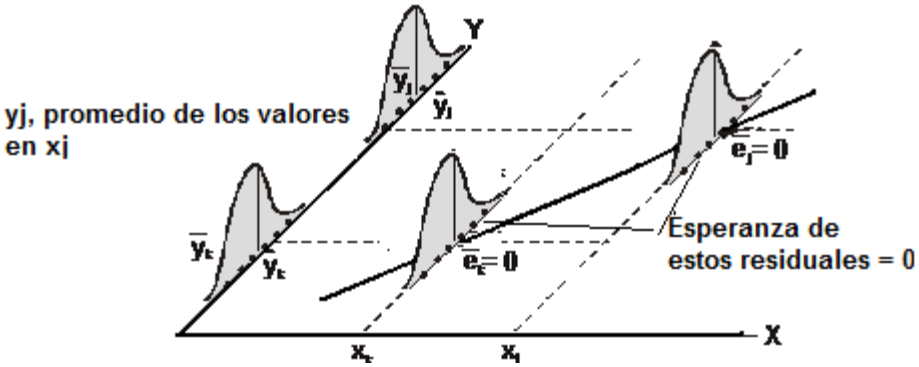


Figura 2.6. Distribuciones de Y_{kj} y e_{kj}, iguales excepto por sus promedios que difieren.

2.3.4.6 Estimación de la varianza de los términos del error. Con base en la evidencia aportada por la séptima propiedad se ve que la varianza de los términos de *y* permite obtener una indicación de la variabilidad de las distribuciones de probabilidades del error. Para una población particular se sabe que la varianza se estima a partir de la suma de cuadrados corregidos de *Y* la cual dividida por sus grados de libertad da un estimador insesgado de la varianza σ² de una población infinita, calculada como:

$$SSY = SCCY = \sum_{i=1}^{n} \left(y_i - \overline{Y} \right)^2 \rightarrow s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{Y})^2}{n-1} \quad (2.30)$$

s_y^2 es a menudo llamada cuadrado medio del error. Obsérvese en la Figura 2.6 que la varianza de Y es la misma que la de los e_i o sea σ^2. Para calcular la suma de cuadrados de las desviaciones, se debe recordar que los Y_i provienen de diferentes distribuciones

de probabilidades con diferentes medias, dependiendo del nivel de X_i, por lo cual la desviación de Y_i debe ser calculada alrededor de su media estimada, o sea de la recta ajustada $\hat{Y}_i$. Dichas desviaciones son los residuales e_i. La apropiada suma de cuadrados será llamada SSE (suma de cuadrados del error o SCE):

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X\right)^2 = \sum_{i=1}^{n} e_i^2 \quad (2.31)$$

con (n-2) grados de libertad por tener dos parámetros ya estimados b_0 y b_1 para la obtención de $\hat{Y}_i$. El cuadrado medio residual MSE (varianza de los errores) se calcula entonces como:

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \quad (2.32)$$

Se puede demostrar que $E(MSE) = \sigma^2$ o sea que resulta insesgada para el modelo de regresión. (Neter et al, 1983) proponen como fórmulas alternativas para el cálculo de la SSE:

$$1)\quad SSE = \Sigma Y_i^2 - b_0 \Sigma Y_i - b_1 \Sigma Y_i X_i \quad (2.33)$$

$$2)\quad SSE = \sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2 - \frac{\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right]^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad (2.34)$$

$$3)\quad SSE = \Sigma Y_i^2 - \frac{\left(\Sigma Y_i\right)^2}{n} - \frac{\left[\Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}\right]^2}{\Sigma X_i^2 - \frac{\left(\Sigma X_i\right)^2}{n}} \quad (2.35)$$

2.4 Modelo de regresión con el error normal.

Sin importar la forma funcional de la distribución de ε_i o de Y_i, se obtuvieron los estimadores puntuales insesgados para $ß_0$ y $ß_1$ con varianza mínima. La necesidad de hacer inferencias acerca de los valores estimados, pruebas de significación y cálculos de intervalos, presupone adoptar una asunción acerca de ε_i, los cuales se considerarán de ahora en adelante normalmente distribuidos: $\varepsilon_i \rightarrow N(0, \sigma^2)$, y por lo tanto similar suerte tendrá la distribución de los Y_i, excepto por su media.

2.4.1 Inferencias en el análisis de regresión. Aceptar la normalidad de las distribuciones de **ε** y Y, permite entonces hacer ciertas inferencias válidas con respecto a los parámetros.

2.4.1.1 Inferencias con respecto a $ß_1$. Importante para saber si existe relación funcional o no, analizar si $ß_1$ es igual o diferente de cero, pues de ser igual se descartaría la existencia de una relación lineal válida entre las variables. Para escrutarlo se plantean las hipótesis:

$$\left.\begin{array}{l} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{array}\right\} \quad (2.36)$$

lo que implica conocer la distribución muestral de los diferentes valores de b_1 que podrían obtenerse con muestreos repetidos, con los niveles de la variable independiente mantenidos constantes de muestra en muestra, ya que se trata de una variable aleatoria función de Y_i. Para el modelo especificado por la ecuación (2.1) la distribución muestral de b_1 es normal, aceptados los postulados de 2.4 o sea que:

$$b_1 \rightarrow \left(\beta_1, \frac{\sigma^2}{\Sigma\left(X_i - \bar{X}\right)^2} \right); \; en \;\; que \;\; E(b_1) = \beta_1 \;\; y \;\; \sigma^2(b_1) = \frac{\sigma^2}{\Sigma\left(X_i - \bar{X}\right)^2} \quad (2.37)$$

Para probarlo basta reconocer; que b_1 sea combinación lineal de las observaciones Y_i, cada una con distribución asumida como normal (aunque no es necesario ni siquiera asumirla). Por ejemplo, se puede llegar al mismo resultado al usar directamente la esperanza de b_1. De acuerdo con el cálculo propuesto para b_1 por la ecuación (2.19), se puede ver que el numerador de ella es expresable así:

$$\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \Sigma\left(X_i - \bar{X}\right)Y_i - \Sigma\left(X_i - \bar{X}\right)\bar{Y} \quad (2.38)$$

$$\Sigma\left(X_i - \bar{X}\right)Y_i - \bar{Y}\,\Sigma\left(X_i - \bar{X}\right) = \Sigma\left(X_i - \bar{X}\right)Y_i - \bar{Y}*0 = \Sigma\left(X_i - \bar{X}\right)Y_i \quad (2.39)$$

con lo que b_1 se puede entonces reescribir como:

$$b_1 = \sum \frac{\left(X_i - \bar{X}\right)Y_i}{\Sigma\left(X_i - \bar{X}\right)^2} \quad (2.40)$$

llamando: $\dfrac{\left(X_i - \bar{X}\right)}{\Sigma\left(X_i - \bar{X}\right)^2} = k_i$, se tiene que:

$$b_1 = \Sigma k_i Y_i \quad (2.41)$$

o sea que es una combinación lineal de Y_i. Usando estos conceptos y las propiedades de k_i, es posible llegar a muchas de las demostraciones necesarias para comprender mejor estos temas, aunque ello se puede lograr de otras formas. Las propiedades son:

$$1) \quad \sum_{i=1}^{n} k_i = 0 \; ; \; \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{0}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = 0 \quad (2.42)$$

$$2) \quad \sum_{i-1}^{n} k_i X_i = 1; \;\; y \;\; 3) \quad \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad (2.43)$$

demostrables como la (2.42). La normalidad de b_1 puede surgir entonces como una consecuencia lógica de ser combinación lineal de Yi, de acuerdo con la ecuación (2.37) ya que en efecto:

$$\left.\begin{aligned} E(b_1) &= E(\Sigma k_i Y_i) = \Sigma k_i E(Y_i) = \Sigma k_i(\beta_0 + \beta_1 X_i) = \beta_0 \Sigma k_i + \beta_1 \Sigma k_i X_i = \beta_1;\ y \\ \sigma^2(b_1) &= \sigma^2(\Sigma k_i Y_i) = \Sigma k_i^2 \sigma^2(Y_i) = \Sigma k_i^2 \sigma^2 = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} \end{aligned}\right\} (2.44)$$

2.4.1.2 Varianza estimada de b_1. La varianza estimada de b_1 se obtiene como:

$$s^2(b_1) = \frac{MSE}{\sum(X_i - \bar{X})^2} \quad (2.45)$$

que resulta ser un estimador insesgado de $\sigma^2(b_1)$.

2.4.1.3 Distribución muestral de $(b_1 - ß_1)/s(b_1)$. Como la distribución muestral de $b_1 \sim N(\beta_1, \sigma^2(b_1))$ es normal, al estandarizar $(b_1 - \beta_1)/s^2(b_1)$ se obtiene una Z*. Ante la imposibilidad de calcular $\sigma^2(b_1)$, se acude al estimado de $\sigma^2(b_1) = s^2(b_1)$ que se tiene de (2.37), con lo cual, en vez de z usamos la distribución t

$$\frac{(b_1 - \beta_1)}{s(b_1)} \rightarrow t_{\left(1-\frac{\alpha}{2};(n-2)gl\right)} \quad (2.46)$$

2.4.1.4 Intervalos de confianza para $ß_1$. Se puede establecer que:

$$P\left\{ t_{\left(\frac{\alpha}{2};n-2\right)} \leq \left[\frac{(b_1 - \beta_1)}{s(b_1)}\right] \leq t_{\left(1-\frac{\alpha}{2};n-2\right)} \right\} = 1 - \alpha \quad (2.47)$$

Por la simetría de t:

$$t_{\left(\frac{\alpha}{2};n-2\right)} = -t_{\left(1-\frac{\alpha}{2};n-2\right)} \quad (2.48)$$

con lo cual se obtienen los intervalos de confianza para $ß_1$

$$\beta_1 = b_1 \pm t_{\left(1-\frac{\alpha}{2};n-2\right)} s(b_1) \quad (2.49)$$

2.4.1.5 Pruebas acerca de $ß_1$. Se utilizan para verificar si efectivamente se da una relación lineal entre X y Y.

1. **Prueba de dos colas:** H_0: $B_1 = 0$ vs H_a: $B_1 \neq 0$. Bajo H_0 siempre se cumple:

$$t^* = \left[\frac{(b_1 - \beta_1)}{s(b_1)}\right] = \frac{b_1}{s(b_1)} \quad (2.50)$$

que se encara con las siguientes reglas de decisión:

$$si \quad |t^*| \leq t_{(1-\frac{\alpha}{2};n-2)} \rightarrow H_0 \quad (2.51) \qquad si \quad |t^*| > t_{(1-\frac{\alpha}{2};n-2)} \rightarrow H_0 \quad (2.52)$$

2. **Prueba de una cola para saber si $ß_1$ es positivo o no: H_0: $B_1 \leq 0$** vs **H_a: $B_1 > 0$**

$$si \quad t^* \leq t_{(1-\alpha;n-2)} \rightarrow H_o; \ si \quad t^* \succ t_{(1-\alpha;n-2)} \rightarrow H_a \quad (2.53)$$

3. **$ß_1$ cumple alguna norma.** Ocasionalmente existe un interés particular de verificar si el valor de $ß_1$ cumple alguna norma conocida de antemano, o por tradición $ß_1 = ß_1^{**}$ ($ß_1^{**}$ es un $ß_1$ conocido). Se plantea una prueba similar a la expuesta en (2.50 a 2.52).

2.4.1.6 Inferencias acerca de ß₀. A menos que exista un $X_i = 0$ dentro de las observaciones, el término $ß_0$ no tiene un significado particular en el modelo. No obstante, dada la existencia de la regresión condicionada, es necesario a veces hacer inferencias acerca de $ß_0$, para lo cual a la manera de lo hecho para b_1 se puede demostrar que b_0 es otra combinación lineal de las Y_i de acuerdo con lo visto en la ecuación (2.20) ya que b_1 es combinación lineal de Y_i, y $b_0 \sim N\left[\beta_0; \frac{\sigma^2(\sum X_i^2)}{n\sum(X_i)^2}\right]$. Para ello entonces:

$$E(b_0) = \beta_0;\ \sigma^2(b_0) = \sigma^2\left(\frac{\Sigma X_i^2}{n*(\Sigma X_i - \bar{X})^2}\right) \quad (2.54)$$

$$E(b_0) = E(\beta_0 + \beta_1\bar{X} - b_1\bar{X}) = E(\beta_0) + \bar{X}\,E(\beta_1) - \bar{X}\,E(b_1) = \beta_0 \quad (2.55)$$

$$\left.\begin{aligned} \sigma^2(b_0) &= \sigma^2\left(\frac{\Sigma Y_i}{n}\right) - \sigma^2(b_1\bar{X}) = \frac{1}{n^2}\sigma^2(\Sigma Y_i) + (\bar{X})^2\sigma^2(b_1) = \\ &\frac{n\sigma^2}{n^2} + \bar{X}^2\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right] \end{aligned}\right\} (2.56)$$

En esta demostración se asumió que $\bar{Y}$ y b_1 son independientes según demostración dada en 2.4.3.1.1. De igual manera que para b_1, $s^2(b_0)$ estima bien a $\sigma^2(b_0)$, por lo cual:

$$s^2(b_0) = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right] \quad (2.57)$$

2.4.1.7 Distribución muestral de $(b_0 - ß_0)/s(b_0)$. Se acude a un tratamiento similar a lo planteado para b_1.

2.4.2 Algunas consideraciones al hacer inferencias sobre $ß_0$ y $ß_1$. Si las distribuciones de probabilidades de Y_i no se apartan seriamente de la normal, las de b_0 y b_1, o sea los diferentes valores de estos parámetros que podrían obtenerse con muestreos repetidos -al mantener los mismos niveles de la variable independiente- tampoco lo harán, o sea, serán aproximadamente normales y se podrá usar la distribución de t para validar su significancia, así como para el establecimiento de sus límites de confianza. Si las distribuciones de Y_i se alejaran fuertemente de la normal, b_0 y b_1 presentan la propiedad de la normalidad asintótica por medio de la cual sus distribuciones se aproximan a la normalidad bajo diversas circunstancias a medida que se incremente el tamaño de la muestra, en cuyo caso el valor de t podrá ser reemplazado por Z para la distribución normal estándar.

El manejo de los límites de confianza y riesgos de errores se interpretan en el sentido de que siempre que se repitan los experimentos con el mismo nivel de confianza y con los mismos valores de X, el tanto por ciento de los niveles de confianza de los intervalos contendrá el verdadero valor de $ß_1$ o $ß_0$. Así mismo, el análisis de las fórmulas de $s^2(b_0)$ y $s^2(b_1)$ indica que para un n y σ^2 dados, las varianzas son afectadas por el espaciamiento de los niveles de X, de tal manera que a más espaciamiento entre las X_i, mayor es la cantidad de la sumatoria $\Sigma(X_i)^2$ y menor resulta la varianza.

2.4.3 Estimación de medias e intervalos de confianza.

2.4.3.1 Estimación del intervalo *E(Y_h)*. Uno de los mayores éxitos de la regresión es poder estimar la media para una o más distribuciones de probabilidades de Y. Sea X_h el nivel de la variable independiente para estimar la respuesta promedia a las distribuciones de Y_i asociadas, que no necesariamente deben coincidir con alguna de las X_i que originaron el modelo, pero sí estar dentro del rango de sus valores. La respuesta promedia cuando $X = X_h$ se denotará como E(Y_h), y además,

$$\hat{Y}_h = b_0 + b_1 X_h \quad (2.58)$$

2.4.3.1.1 Distribución muestral de $\hat{Y}_h$. Se refiere a los diferentes valores de $\hat{Y}_h$ que se obtendrían en muestreos repetidos con los mismos valores de la variable independiente. Se hace constar que su distribución es normal por ser combinación lineal de los Y_i. Para probarlo entonces se acude a demostrar que la media:

$$E(\hat{Y}_h) = E(b_0 + b_1 X_h) = E(b_0) + X_h E(b_1) = \beta_0 + \beta_1 X_h \quad (2.59)$$

La varianza debe partir de demostrar que b_1 y $\bar{Y}$ no están correlacionados, o sea que $\sigma(\bar{Y}, b_1) = 0$. Si se acude a las ecuaciones $\bar{Y} = \frac{1}{n}\sum y_i \;\; y, \;\; b_1 = \sum k_i Y_i$ se ve que las observaciones Yi son independientes y que la covarianza es obtenible como una suma de productos corregidos por medio de la expresión:

$$\sigma(\bar{Y}, b_1) = \Sigma \bar{Y}\, b_1 - \frac{\Sigma \bar{Y}\, \Sigma b_1}{n} = n\bar{Y} b_1 - \frac{n\bar{Y}\, n\, b_1}{n} = 0 \quad (2.60)$$

o por medio del teorema que establece que para variables independientes Y_i la covarianza de dos funciones lineales $\sum a_i Y_i \; y, \sum b_i Y_i$ se obtiene como $\sigma\left(\sum a_i Y_i, \sum b_i Y_i\right) = \sum a_i b_i \sigma^2(Y_i) \rightarrow \sigma(\bar{Y}, b_1) = \sum \frac{1}{n} k_i \sigma^2(y_i) = \frac{1}{n}\sigma^2(y_i)\sum k_i = 0$.

Para mostrar $\sigma^2(\hat{y}_h)$, se puede escribir:

$$\hat{Y}_h = \bar{Y} - b_1 \bar{X} + b_1 X_h = \bar{Y} + b_1 (X_h - \bar{X}) \quad (2.61)$$

entonces:

$$\left.\begin{aligned} \sigma^2(\hat{Y}_h) = \sigma^2\left[\bar{Y} + b_1(X_h - \bar{X})\right] = \sigma^2(\bar{Y}) + \left(X_h - \bar{X}\right)^2 \sigma^2(b_1) = \\ \frac{n\sigma^2}{n^2} + \left[\frac{\sigma^2 (X_h - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{\left(X_h - \bar{X}\right)^2}{\Sigma\left(X_i - \bar{X}\right)^2}\right] \end{aligned}\right\} \quad (2.62)$$

además $\sigma^2\left(\hat{Y}_h\right)$ queda estimada por $s^2\left(\hat{Y}_h\right)$ así:

$$s^2(\hat{Y}_h) = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}\right] \quad (2.63)$$

de donde surgen los intervalos de confianza para E(Y_h) cuya media será:

$$E(\hat{Y}_h) = \beta_0 + \beta_1 X_h \quad (2.64)$$

2.4.3.1.2 Distribución muestral de E(Y_h). Tiene, como debe suponerse, una distribución de $t_{(1-\alpha/2,(n-2)gl)}$. Los intervalos de confianza se hallan en la forma acostumbrada:

$$E(Y_h) = \hat{Y}_h \pm t_{1-\frac{\alpha}{2};(n-2)gl}\, s(\hat{Y}_h) \quad (2.65)$$

Por ejemplo el valor esperado, error y límites de confianza de la altura media para la distribución de las alturas con un diámetro de 30 cm con base en el modelo $H = 15.514 + 0.1048d$ escogido para los datos de la Tabla 1 fueron, para un $\alpha = 0.05$: $\hat{Y}_h = \text{alt} = 10.7004 + 0.4156 * 31.16 = 23.65\,\text{m}$ y

$$s^2(\hat{Y}_h) = 4.52504\left[\frac{1}{10} + \frac{(30-31.16)^2}{403.704}\right] = 0.4675; \quad s(\hat{y}_h) = 0.6838$$

de donde: $Y_h = 23.65 \pm 2.306 * 0.8269 = \begin{cases} 22.07 \\ 25.23 \end{cases}$.

Salidos del Análisis de Regresión del modelo lineal: H = a + b*d

Parámetro	Estimado	Err est	t	Valor-P
Intercepto	10.70	3.25	3.29	0.01
Pendiente	0.415	0.10	4.07	0.00

Análisis de Varianza

Fuente	SS..	Gl	MS..	Razón-F	Valor-P
Modelo	74.9047	1	74.9047	16.55	0.0036
Residuo	36.2003	8	4.52504		
Total (Corr.)	111.105	9			

2.4.3.2 Predicción de una nueva observación Y_{h_n}. Se debe interpretar como el resultado para una nueva observación independiente de aquellas que originaron el modelo. El caso anterior estimaba la media de una distribución de Y, y acá se trata de predecir un suceso aislado extraído precisamente de esa distribución de Y. Como desde luego la mayoría de sucesos aislados se desvían de esta respuesta promedio, se afectan los límites de confianza. (Wonnacott y Wonnacott, 1981) se refieren a este tema como "predicción del intervalo para un Y_0 individual" que ayuda a aclarar la sutil diferencia con el caso anterior. Se deja además constancia de lo útil de esta estimación en muchas actividades forestales en las cuales, a menudo, se deben conocer valores para nuevas observaciones con modelos ya estructurados. La estadística lo asume de dos maneras diferentes, dependiendo si los parámetros de la regresión son conocidos o no.

$$E\left(\hat{Y}_{h_n}\right) = \beta_0 + \beta_1 X_{h_n} \quad (2.66)$$

Si se conocieran ß$_0$, ß$_1$ y σ, se obtendrían así los límites de confianza para un X_h y un nivel α dados:

$$Y_{h_n} = E\left(Y_y\right) \pm Z_{\left(1-\frac{\alpha}{2}\right)}\sigma \quad (2.67)$$

Como generalmente los parámetros β_0, β_1 y σ son desconocidos, (2.68) ya no sería adecuada pues aparecen dos distribuciones extremas de probabilidades de Y a izquierda y derecha alrededor de $\hat{Y}_h$ debiendo acudir al siguiente teorema que establece:

$$\frac{\hat{Y}_h - Y}{s\left(Y_{h_n}\right)} \sim t_{\left(1-\alpha/2;\, n-2\right)} \quad (2.68)$$

en el cual, el estadístico estandarizado usa el estimador puntual $\hat{Y}_h$ en el numerador en vez del verdadero valor medio $E\left(Y_h\right)$ y la nueva observación Y por ser este desconocido, de donde se sigue que la predicción usual con una probabilidad (1 - α), para la nueva observación, según (Neter et al, 1983) será:

$$\hat{Y}_{h_n} = \hat{Y}_h \pm t_{\left(1-\alpha/2;\, n-2\right)} s\left(Y_{h_n}\right) \quad (2.69)$$

lo cual requiere el cálculo de $s\left(Y_{h_n}\right)$ obtenida a partir de la independencia de la nueva observación Y y las observaciones muestrales originales en la que se basa $\hat{Y}_h$ con lo cual se denotaría la varianza del numerador de (2.68) con $\sigma^2\left(Y_{h_n}\right)$ y:

$$\sigma^2\left(Y_{h_n}\right) = \sigma^2\left(\hat{Y}_h - Y\right) = \sigma^2\left(\hat{Y}_h\right) + \sigma^2 \quad (2.70)$$

en el que se observan dos componentes: la varianza de la distribución muestral de $\hat{Y}_h$ y la varianza de las distribuciones individuales de Y. Como lo anotan (Wonnacott y Wonnacott, 1981) se trata de la estimación de una observación individual más que de una media estable para todos los posibles Y. Un estimador insesgado de $\sigma^2\left(Y_{h_n}\right)$ es:

$$s^2\left(Y_{h_{nueva}}\right) = s^2\left(\hat{Y}_h\right) + MSE \quad (2.71)$$

que conduce sin más detalles, por lo similar de las presentaciones anteriores a:

$$s^2\left(Y_{h_{nueva}}\right) = MSE\left[1 + \frac{1}{n} + \frac{\left(X_h - \bar{X}\right)^2}{\sum\left(X_h - \bar{X}\right)^2}\right] \quad (2.72)$$

Por ejemplo, supóngase interesado en predecir la altura de un árbol cuyo d = 30 cm con los datos de la Tabla 1 y analizar el resultado obtenido.

$$\hat{H}_{hn} = 15.514 + 0.1048 * 30 = 18.66\, m$$

$$s^2\left(\hat{H}_{hn}\right) = 4.525\left[1 + \frac{1}{10} + \frac{(30 - 31.16)^2}{433.704}\right] = 4.993; \rightarrow s\left(\hat{H}_{hn}\right) = 2.2344$$

Los límites de confianza serán:

$$Y_{hn} = 23.65 \pm 2.306 * 2.2364 = \begin{cases} 18.50 \\ 28.80 \end{cases}$$

más amplios que los anteriores por tratarse de una observación individual, con una t(8) gl = 2.306, ya que resulta más difícil predecir una observación individual que la media de las distribuciones de Y_i, (Wonnacott y Wonnacott,1981).

2.4.3.3 Predicción de m nuevas observaciones para una X_h dada. A veces se precisa el promedio de un lote de m nuevas observaciones para un nivel dado de la

variable independiente, con el fin de predecir el valor esperado. Sin entrar en detalles, para este caso:

$$s^2\left(\bar{Y}_{hnueva_m}\right) = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{\left(X_h - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right] \quad (2.73)$$

Por ejemplo, qué rango de valores se espera para la altura de 4 observaciones con d = 30 cm ?. Entonces, $s^2(Y_{hn}) = 1.5988$, $s(Y_{hn}) = 1.2644$, cuyos límites de confianza son: 20.73 y 26.56, que lógicamente dan unos límites más estrechos que para una observación individual, ya que se trata de grupos de ellas pero más amplio que para las distribuciones de Y_i.

2.5 Interpolación y extrapolación.

En la sección anterior se considera como éxito del proceso el poder estimar valores dentro del rango de la variable independiente asumida como continua. Se ve, además, en las fórmulas (2.63), (2.72) y (2.73) que los intervalos de predicción serán más amplios a medida que X_h, o X_{hn} se alejen más de $\bar{X}$, esto porque el estimado de la media Y_h y Y_{hn} es menos preciso a medida que los X_h, o X_{hn} se localicen más lejos de $\bar{X}$. Según (Wonnacott y Wonnacott, 1981), es notable la diferencia entre intervalos de confianza e intervalos de predicción, pues estos se refieren a la toma de muestras repetidas basadas en el mismo conjunto de la variable X (como aleatoria), en tanto que los intervalos de confianza representan una inferencia sobre un parámetro, entendidos como un intervalo, que cubre el valor del parámetro. Con todo, la anterior teoría enfatiza la posibilidad de inferencias válidas dentro del rango de las X muestrales, al interpolar. El caso contrario, cuando X_h está por fuera de este rango, o sea la extrapolación, puede prestarse a malas interpretaciones de resultados por salir del área de influencia $X_h - \bar{X}$ en la cual el modelo no tiene prácticamente incumbencia, se pierde la asunción de la linealidad y, hasta cambiar aún drásticamente por las X cada vez más alejadas de su valor central.

2.6 Mínimos cuadrados con X aleatorias.

Resulta como caso muy común uno en que no se controlan los valores de la variable independiente, sino que se toman al azar. Sorprendente con algunas reinterpretaciones de conceptos, lo estudiado seguirá siendo válido, bastando solamente que: "el error e sea estadísticamente independiente de X". Asumidas pues todas las asunciones incluida esta última se puede probar que: 1. b_1 es un estimador insesgado de $ß_1$. 2. Que los intervalos de confianza para $ß_1$ siguen siendo válidos. 3. Que si se asume un error normal, b_1 es aún un estimador máximo verosímil de $ß_1$.

2.7 Errores en la variable independiente.

Los métodos mínimos cuadráticos asumen las variables independientes sin error, sin embargo, cuando tienen varianzas considerables, el estimado de la pendiente presenta sesgos notables hacia cero. Una regla de apuño es que se puede tener confianza en los

mínimos cuadrados si la varianza de X es menor que un décimo de la desviación promedia de dispersión de las X a partir de su media. (Daniel y Wood, 1980). Si sucede lo contrario se recomienda un método propuesto por Bartlett, citado por el anterior, que escapa al alcance de este libro.

2.8 Análisis de varianza de la regresión.

Este concepto se estudia para proyectarlo a la regresión múltiple pues con respecto a la regresión simple no aporta más información que la dada por las pruebas ya conocidas. Se basa en la partición de la suma de cuadrados totales y de los grados de libertad asociados con la variable dependiente Y, en las distintas desviaciones con respecto a cada Y_i empezando con la más conocida, la suma de cuadrados de $\left(Y - \bar{Y}\right)$, que estima la varianza por definición y es conocida como la suma de cuadrados totales.

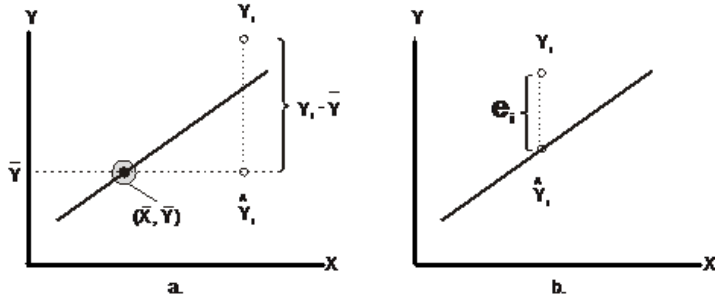
2.8.1 Suma de cuadrados totales SSTO. La suma de cuadrados totales:

$$SSTO = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \quad (2.74)$$

permite ver que si todas las observaciones fueran iguales entonces SSTO = 0, y que, a mayor variación con respecto a la media, mayor SSTO. Figura 2.7. A la SSTO se le llama "variación total".

2.8.2 Suma de cuadrados del error. Cuando se utiliza la regresión, la variación se estima alrededor de la línea ajustada de regresión. Las desviaciones son entonces $y_i - \hat{y}_i = e_i$.Figura 2.7 b). La suma de cuadrados del error será entonces

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 \quad \left(2.75\right)$$



Figuras 2.7. a) Diferencias entre y_i y su media. b) Diferencias entre y_i y su línea ajustada.

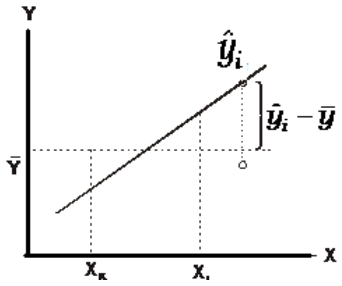
Si todos los puntos cayeran en la línea, SSE = 0. A mayor SSE, mayor variación alrededor de la línea de regresión. Cuando se pierde claridad frente a las diversas pruebas o estimados, el modelo que minimice la SSE puede resultar ventajoso. A menudo a la SSE se le llama "variación no explicada", concepto que quedará claramente entendido al abordar el estudio del coeficiente de correlación.

2.8.3 Suma de cuadrados de la regresión. Adicionalmente a las ya anotadas se puede generar una suma de cuadrados con base en las diferencias de $\hat{y}_i - \bar{Y}$ que se denominará por contexto suma de cuadrados de la regresión, la cual se considera como una medida de la variabilidad de las Y por la línea, con respecto al promedio. **Figura 2.8.**

$$SSR = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{Y}\right)^2 \quad (2.76)$$

que conduce a

$$SSR = \sum_{i=1}^{n}\left(b_0 + b_1 X_i - b_0 - b_1\bar{X}\right)^2 = b_1^2\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \quad (2.77)$$



Figuras 2.8, Diferencias entre $\hat{y}_i$ y $\bar{Y}$ y valores simétricos x_k y x_l.

Cuando SSR = 0 la línea de regresión sería horizontal y no habría relación estadística. Mientras mayor es SSR con respecto a SSTO, mayor es el efecto de la relación de regresión al computar la variación total de las observaciones de Y. Se nota además que para los valores X_k simétrico con X_l las desviaciones son iguales y de signo contrario, situación digna de tenerse en cuenta a la hora del cálculo de los grados de libertad. A menudo a la SSR se le llama "variación explicada" en Y. En la Figura 2.9 es posible observar elementos gráficos intuitivos que conducen al ANAVA pues con base en las observaciones se ve $y_i - \bar{Y} = y_i - \hat{y}_i + \hat{y}_i - \bar{Y}$. Véase entonces lo sucedido con las respectivas sumas de cuadrados:

$$\left.\begin{aligned}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2 &= \sum_{i=1}^{n}\left[\left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \bar{Y}\right)\right]^2 = \\ \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 &+ \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n}\left[\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right)\right]\end{aligned}\right\} \quad (2.78)$$

El tercer sumando $2\sum_{i=1}^{n}\left[\left(y_i - \hat{y}_i\right)\left(\hat{y}_i - \bar{Y}\right)\right]$ se puede descomponer así:

$$\left.\begin{aligned}\sum_{i=1}^{n}\left[\left(Y_i - \hat{Y}_i\right)\hat{Y}_i - \left(Y_i - \hat{Y}_i\right)\bar{Y}\right] &= \sum_{i=1}^{n}\hat{Y}_i\left(Y_i - \hat{Y}_i\right) - \sum_{i=1}^{n}\bar{Y}\left(Y_i - \hat{Y}_i\right) = \\ \sum_{i=1}^{n}\hat{Y}_i e_i - \bar{Y}\sum_{i=1}^{n}e_i = \sum_{i=1}^{n}\left(b_0 + b_1 X_i\right)e_i &= b_0\sum_{i=1}^{n}e_i + b_1\sum_{i=1}^{n}e_i X_i = 0\end{aligned}\right\} \quad (2.79)$$

por ello es posible entonces que

$$SSTO = SSE + SSR \quad (2.80)$$

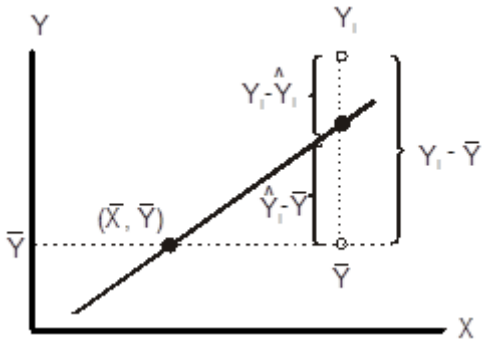


Figura 2.9. Igualdades de base para el análisis de varianza de la regresión.

Se presentan las siguientes fórmulas computacionales para ellas:

$$\left.\begin{aligned} SSTO &= \Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma Y_i^2 - \frac{\left(\Sigma Y_i\right)^2}{n} = \Sigma Y_i^2 - n\bar{Y}^2 = SCCY \\ SSR &= b_1\left[\Sigma X_i Y_i - \frac{\left(\Sigma X_i \Sigma Y_i\right)}{n}\right] = b_1 SPCXY = \\ & b_1\left[\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right] = b_1^2 \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \end{aligned}\right\} \quad (2.81)$$

2.8.4 Partición de los grados de libertad. Con cada suma de cuadrados se asocia cierto número para sus grados de libertad así: SSTO tiene (n-1) gl, se pierde un grado por tener un valor estimado $\hat{Y}$. La SSE pierde 2 grados por tener dos estimados b_0 y b_1, entonces queda con (n-2) gl. La SSR se compone de una suma de desviaciones no independientes, o sea que se pueden eliminar las simétricas con respecto a la media, quedando al final dos sumandos, de los cuales uno es estimado, lo que da un grado de libertad. También se obtiene sabiendo que son dos parámetros estimados de los cuales solo se usa 1. De acuerdo con lo anterior, los grados de libertad resultan aditivos, lo que configura según el teorema de Cochran la posibilidad del análisis de varianza, por la partición de la suma de cuadrados totales en dos sumas parciales y la aditividad mencionada.

2.8.5 Análisis de varianza.

2.8.5.1 Cuadrados medios esperados. La esperanza de MSE = E(MSE) = σ^2 , según las hipótesis planteadas puede demostrarse por el teorema de Cochran y la por la propiedad de $\frac{(n-1)s^2}{\sigma^2}$ de distribuirse como una χ^2 cuya $E\left(\chi^2\right)$ iguala a sus grados de libertad. Como $(n-1)s^2$ es siempre una suma de cuadrados:

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)gl \rightarrow E\left(\frac{SSE}{\sigma^2}\right) = n-2 \;\therefore E\left(\frac{SSE}{n-2}\right) = \sigma^2 \;\therefore\; E(\,MSE\,) = \sigma^2 \quad (2.82)$$

La esperanza de MSR se puede obtener de la expresión dada para SSR

$$SSR = b_1^2 \Sigma\left(X_i - \bar{X}^2\right) \therefore E\left(MSR\right) = E\left[\ b_1^2 \Sigma\left(X_i - \bar{X}\right)^2\ \right] \quad (2.83)$$

Se sabe que:

$$\sigma^2\left(b_1\right) = E\left(b_1^2\right) - \left[E(b_1)E(b_1)\right] \therefore E\left(b_1^2\right) = \sigma^2(b_1) + \left[E(b_1)\right]^2 = \sigma^2\left(b_1\right) + \beta_1^2 \quad (2.84)$$

Entonces:

$$\left.\begin{aligned} &E\left(MSR\right) = \left[\sigma^2\left(b_1\right) + \beta_1^2\right]\Sigma\left(X_i - \bar{X}\right)^2 = \\ &\frac{\sigma^2\Sigma\left(X_i - \bar{X}\right)^2}{\Sigma\left(X_i - \bar{X}\right)^2} + \beta_1^2\Sigma\left(X_i - \bar{X}\right)^2 = \sigma^2 + \beta_1^2\Sigma\left(X_i - \bar{X}\right)^2 \end{aligned}\right\} \quad (2.85)$$

2.8.5.2 Prueba de F para $\beta_1 = 0 \quad vs \quad \beta_1 \neq 0$**.** La prueba de F para probarlo se plantea como F* = MSR/MSE. Altos valores de F soportan H_a, entonces, si:

$$F^* \leq F(1-\alpha;1,n-2) \rightarrow H_0 \ ;\ si\ \ F^* > F(1-\alpha;1,n-2) \rightarrow H_a \quad (2.86)$$

TABLA 2.2 análisis de varianza para la regresión.

FUENTE DE VARIACION	SUMAS DE CUADRADOS SS..	GRAD DE LIBERT.	CUADRADOS MEDIOS	ESPERANZA DE MS E(MS)
REGRESION	$SSR = b_1^2\Sigma\left(x_i - \bar{X}\right)^2$	1	MSR=SSR/1	$\sigma^2 + \beta_1^2\Sigma\left(x_i - \bar{X}\right)^2$
ERROR	$SSE = \Sigma\left(y_i - \hat{y}_i\right)^2$	n-2	MSE=SSE/(n-2)	σ^2
TOTAL	$SSTO = \Sigma\left(y_i - \bar{Y}\right)^2$	n-1		
CORRECCIÓN PARA LA MEDIA	$F_{ac\ de\ corr} = n\bar{Y}^2$			
TOTAL, SIN CORREGIR	$SSTO = \Sigma y_i^2$	n		

2.8.5.3 Equivalencia entre las pruebas de F y de t. Por lo ya visto:

$$SSR = b_1^2\sum_{I=1}^{N}\left(x_i - \bar{X}\right)^2 \ ;\ F^* = \frac{\dfrac{SSR}{1}}{\dfrac{SSE}{n-2}} = \frac{b_1^2\sum_{I=1}^{N}\left(x_i - \bar{X}\right)^2}{MSE} \quad (2.87)$$

pero $\sigma^2\left(b_1\right)$ se estima a partir de $\sigma^2\left(b_1\right) = \dfrac{MSE}{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}$

de donde:

$$MSE = s^2(b_1)\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 \quad (2.88)$$

y por ello:

$$F^* = \frac{b_1^2 \Sigma(X_i - \overline{X})^2}{s^2(b_1)\Sigma(X_i - \overline{X})^2} = \left[\frac{b_1}{s(b_1)}\right]^2 = t^2 \quad (2.89)$$

Por lo cual, en la regresión lineal simple, el análisis de varianza no aporta mas información que la obtenida con las pruebas anteriormente vistas, pero si sirve de fundamento a muchos análisis de la regresión lineal múltiple.

2.8.6 Medidas descriptivas de asociación entre X y Y.

2.8.6.1 Coeficiente de determinación R^2. Se vio que SSTO mide la variación en las observaciones Y_i con respecto a su media, independientemente de las X_i, lo que se representa como variabilidad no asociada, o no modelada, en la Figura 2.11, o sea, la incertidumbre de predecir Y sin considerar X. Similarmente, SSE mide la variación de las Y_i cuando el modelo de regresión emplea las X. El coeficiente de determinación R^2 es entonces una medida del efecto logrado por las X para reducir la variación de Y_i. La Figura 11 muestra el efecto en forma intuitiva.

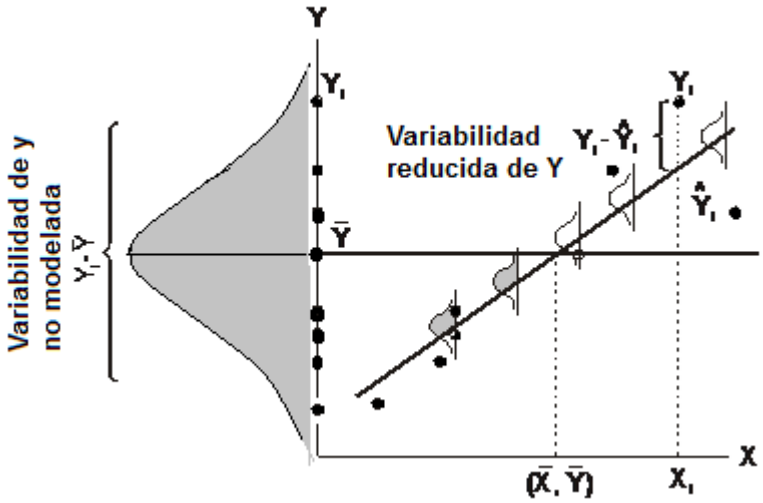


Figura 11. Efecto de la línea de regresión para reducir la variabilidad natural de y_i

El coeficiente de determinación se calcula como:

$$R^2 = \frac{(SSTO - SSE)}{SSTO} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad (2.90)$$

Como $0 \le SSE \le SSTO \quad \rightarrow \quad 0 \le R^2 \le 1$. A la $\sqrt{R^2} = \pm R$ se le conoce como coeficiente de correlación. Otra fórmula de R es:

$$R = \frac{SPCXY}{\sqrt{SCCX}\sqrt{SCCY}} \quad (2.91)$$

El coeficiente de correlación R varía de acuerdo con el signo de la pendiente ya que $-1 \le R \le +1$. Se ampliará un poco más el concepto al estudiar incipientemente la correlación. Basta decir que existen tablas y transformaciones para calificar R

obtenido dados (n-p) grados de libertad (Snedecor y Cochran, 1970; Steel y Torrie, 1985).

		variables independientes			
gl error	p	1	2	3	4
1	0.05	0.997	0.999	0.999	0.999
1	0.01	1.000	1.000	1.000	1.000
2	0.05	0.950	0.975	0.983	0.987
2	0.01	0.990	0.995	0.997	0.998
3	0.05	0.878	0.930	0.950	0.961
3	0.01	0.959	0.976	0.983	0.987
4	0.05	0.811	0.881	0.912	0.930
4	0.01	0.917	0.949	0.962	0.970
5	0.05	0.754	0.836	0.874	0.898
5	0.01	0.974	0.917	0.937	0.949
17	0.05	0.456	0.545	0.601	0.641
17	0.01	0.575	0.647	0.691	0.724
18	0.05	0.444	0.532	0.587	0.628
18	0.01	0.561	633.000	0.678	0.710
19	0.05	0.433	0.520	0.575	0.615
19	0.01	0.537	0.620	0.665	0.698
20	0.05	0.413	0.509	0.563	0.604
20	0.01	0.526	0.608	0.652	0.685
21	0.05	0.404	0.498	0.522	0.592
21	0.01	0.515	0.596	0.641	0.674

Libro de Snedecor,Etatistical Methods,1946

		variables independientes			
gl error	p	1	2	3	4
24	0.05	0.388	0.470	0.523	0.562
24	0.01	0.496	0.565	0.609	0.642
25	0.05	0.381	0.462	0.514	0.553
25	0.01	0.487	0.550	0.600	0.633
26	0.05	0.374	0.454	0.506	0.545
26	0.01	0.478	0.546	0.590	0.624
27	0.05	0.367	0.446	0.498	0.536
27	0.01	0.470	0.538	0.582	0.615
28	0.05	0.361	0.439	0.490	0.529
28	0.01	0.463	0.530	0.573	0.606
125	0.05	0.174	0.216	0.246	0.269
125	0.01	0.228	0.266	0.294	0.316
150	0.05	0.159	0.198	0.225	0.247
150	0.01	0.208	0.244	0.270	0.290
200	0.05	0.138	0.172	0.196	0.215
200	0.01	0.181	0.212	0.234	0.253
300	0.05	0.113	0.141	0.160	0.176
300	0.01	0.148	0.174	0.192	0.208
400	0.05	0.098	0.122	0.139	0.153
400	0.01	0.128	0.151	0.167	0.180
500	0.01	0.088	0.109	0.124	0.137
500	0.05	0.115	0.135	0.150	0.162
1000	0.01	0.062	0.077	0.088	0.097
1000	0.01	0.081	0.096	0.106	0.115

El término R^2 se acepta algunas veces como la proporción de la variación total de Y, que puede explicarse al asociarla con una nueva variable X. Así, tomada literalmente esta explicación puede conducir a interpretaciones erróneas ya que un modelo de regresión no implica que necesariamente Y depende de X en forma causal. Los modelos de regresión tampoco contienen parámetros que sean estimados por el R^2, que es simplemente una medida descriptiva del grado de asociación lineal entre X y Y de la muestra analizada. Sólo al estudiar la correlación se podrá generar propuestas para tomar a R^2 como en un estimador válido de tal asociación. Ha existido una tendencia a juzgar una regresión solo por este estimador; situación muy usual en el sector forestal, por lo cual conviene alertar sobre ello, pues un ajuste puede ser malo y el valor de R elevado. (Caillez, 1980) reporta algunos ejemplos analizando

regresiones para volumen que presentan tal situación, como se confirmará luego del análisis de residuales.

2.8.6.2 Otras concepciones del R^2. Existen estudios sobre el R^2, incluso propuestas de ligarlo de alguna forma a distribuciones de probabilidades que permitieran un juicio más objetivo de su significancia. Además, puede prestarse a usos indebidos, por lo cual se aportan algunos criterios que de alguna manera inquieten al usuario. Para la fórmula exacta se debe distinguir entre regresiones con y sin término constante. (Rousseeuw y Leroy, 1987). Para el caso de ecuaciones lineales mínimo cuadráticas la mayoría de las expresiones que se presentarán resultan equivalentes, pero para otros tipos de regresión como los no lineales (en los parámetros), los resultados difieren y se falla en el análisis del problema (Kvålseth, 1985), quien presenta otras expresiones para el R^2 mencionadas a través de la literatura, a parte de las ya mencionadas:

$$\left.\begin{array}{l} R_1^2 = 1 - \dfrac{\Sigma\left(Y - \hat{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}; \quad R_2^2 = \dfrac{\Sigma\left(\hat{Y} - \bar{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}; \quad R_3^2 = \dfrac{\Sigma\left(\hat{Y} - \bar{\hat{Y}}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}; \\ R_4^2 = 1 - \dfrac{\Sigma\left(e - \bar{e}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}; \; R_5^2 = R^2\, entre \left[\dfrac{regresandos}{regresores}\right]; \; R_6 = R^2\, entre\, Y\, y\, \hat{Y}; \\ R_7^2 = 1 - \dfrac{\Sigma\left(Y - \hat{Y}\right)^2}{\Sigma Y^2}; \quad R_8^2 = \dfrac{\Sigma \hat{Y}^2}{\Sigma Y^2} \end{array}\right\} \qquad (2.92)$$

El R_1^2 es como el resumen de los R_1^2 a R_6^2 , computable para modelos linealizados con transformaciones en algunas de sus variables, también llega a ser el coeficiente de correlación producto - momento de Pearson en regresión lineal bivariada. La escogencia apropiada del R^2 se debe basar en varias consideraciones:- Tipo de modelo formulado, - Técnica de ajuste utilizada, - Propósito para el cual es usado, - Propiedades deseables de él. Por ejemplo, se reporta que R_1^2 a R_6^2 resultan iguales en modelos lineales simples, R_7^2 y R_8^2 parecen recomendables sólo en casos de modelos lineales sin intercepto. (Montgomery, 1982) reporta las diferencias así: Para modelos con intercepto lo define como la proporción de la variabilidad medida por la suma de cuadrados acerca de , mientras que en modelos sin intercepto mide la proporción de la variabilidad con los **Y** acerca del origen, explicadas por la regresión y aclara: Si se usa R^2 para modelos sin intercepto, al basarse en s_{yy} podrían esperarse valores de R^2 negativos puesto que la variación acerca de la línea de regresión (SSE) podría exceder S_{yy}. Para estos casos además podrían esperarse valores de R_2^2 y R_3^2 mayores que 1. R_1^2 y R_4^2 podrían ser posiblemente negativos en regresiones inapropiadas o modelos en que resulta notoria la presencia de observaciones remotas, aunque generalmente resultarán menores que 1 y no negativos. Para casos lineales sin intercepto y para modelos no lineales es recomendable rechazar el uso de R_4^2. A pesar de que R_5^2 proporciona una indicación

indirecta del ajuste de modelos originales no lineales, ello no proporciona una descripción de lo estrecho de la relación. El R_6^2 para regresiones no lineales puede producir errores potenciales de estimación pues resulta claramente posible para Y y $\hat{Y}$ estar altamente correlacionados aunque sus correspondientes valores se desvíen sustancialmente. Por último aunque $0 \leq R_7^2 \leq 1$, tanto el como R_8^2 puede resultar mayor que 1 para modelos no lineales.

Podríamos concluir hasta acá que, la t de student, la F de Fisher, el Estadísitico de Durbin Watson y el R^2 son estadísticos importantes de un modelo, los 3 primeros deben estar respaldados por p-values <0.05, el R^2 concebido como una medida de la disminución de la variabilidad de Y por asociarla con X, y con tablas para su evaluación

2.9 Aptitud del modelo - Análisis de residuales.

Aparte de los elementos estadísticos descritos para evaluar el modelo es necesario estudiar la aptitud o capacidad de cumplir los postulados que sustentan la regresión lineal, por medio de un análisis de sus residuales ya definidos como $e_i = y_i - \hat{y}_i$ diferentes a los ε_i y además asumidos como $\mathrm{N}\left(0, \sigma^2\right)$. Aunque se hizo expresa la diferencia $\varepsilon_i \neq e_i$ si el modelo resulta apropiado para los datos, entonces los e_i reflejarán las propiedades de los ε_i, idea que soporta el análisis de residuales.

2.9.1 Propiedades de los residuales. Las más importantes son:

1. La media de los residuales es cero:

$$\bar{e} = \frac{\sum e_i}{n} \quad (2.93)$$

a causa de esta situación no es posible saber si $E\left(\varepsilon_i\right) = 0$.

2. La varianza de los residuales es:

$$\frac{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2}{n-2} = \frac{SSE}{n-2} = MSE \quad (2.94)$$

Si el modelo es apropiado, la MSE es un estimador insesgado de la varianza de los términos del error.

3. Para muchos análisis es conveniente la normalización de residuales así:

$$e_{i_{est}} = e_{is} = \frac{e_i}{\sqrt{MSE}} = \frac{\left(e_i - \bar{e}\right)}{s} \quad (2.95)$$

4. Los residuales no son variables aleatorias independientes porque involucran los valores ajustados $\hat{y}_i$ basados en los estimadores b_0 y b_1 por lo cual se asocian con (n-2) grados de libertad. Sin embargo, cuando el número de residuales es grande,

comparado con el número de parámetros del modelo, el efecto de no independencia se vuelve relativamente intrascendente.

2.9.2 Desajustes del modelo analizables con residuales

Al examinar los residuales es posible conocer algo del comportamiento teórico de los modelos por medio de algunos diagnósticos gráficos más o menos simples o intuitivos algunos, y otros matemáticos con base en propuestas, algunas de las cuales escapan al alcance de este libro (Draper & Smith 1966, Daniel & Wood 1980, Neter et al. 1983). Los desarreglos más visibles son:

1. Que el modelo propuesto no se acomode a la regresión lineal.

2. Que la varianza de los términos del error no resulte constante, o sea verificar la presencia de heterocedasticidad.

3. Que los términos del error no sean independientes: autocorrelación de errores.

4. Que algunas observaciones para el modelo queden fuera de lugar, remotas, o su término clásico: "outliers".

5. Que los términos del error no se distribuyan normalmente.

6. Que de pronto falten variables al modelo, que siendo importantes, no se hubieran tenido en cuenta.

La mayoría de los análisis expuestos tendrán carácter gráfico, aunque se darán otros simples que complementen su diagnóstico.

2.9.2.1 No linealidad de la función. Análisis simple en el cual al graficar los puntos se detectan posiciones sistemáticas al inicio, intermedio y final de la supuesta recta de regresión por encima o por debajo de ella. Ejemplo: Los siguientes valores de diámetro y altura fueron obtenidos en campo (Tabla 2.) y graficados en la Figura **2.**11 salida del modelo de los datos de la Tabla 2.3:

Tabla 2.3. Valores de diámetro (d), altura (h) y residuales (e_i).

d	5	10	15	20	25	30	35	40	7	10	44
h	7	16	18	22	23	23	22	21	10	13	21
e_i	-5.41	1.99	2.39	4.79	4.19	2.59	-0.01	-2.61	-3.05	-1.01	-3.89

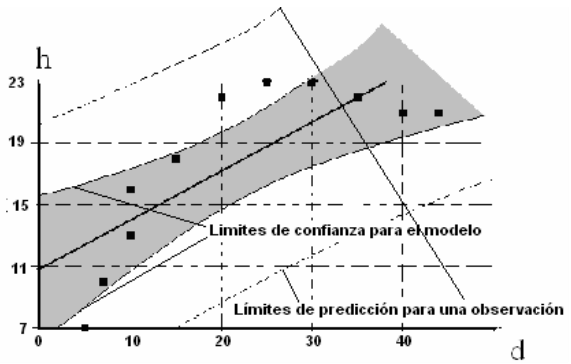


Figura 2.13 Línea de regresión y datos de la Tabla 2.

Parámetro	Estimado	Err est	t	Valor-P
Intercepto	10.70	3.25	3.29	0.01
Pendiente	0.415	0.10	4.07	0.00

Análisis de Varianza

Fuente	SS..	Gl	MS..	Razón-F	Valor-P
Modelo	74.9047	1	74.9047	16.55	0.0036
Residuo	36.2003	8	4.52504		
Total (Corr.)	111.105	9			

Coeficiente de correlación = 078 $R^2 = 0.622$; Error estándar de est. = 3.62729.
La regresión ajustada $\hat{h} = 10.70 + 0.415d$ dio un $R = 0.79$, significativo al 95% y al 99%, por lo cual induce a pensar en un buen ajuste para el modelo. Para verificarlo se debe acudir entonces al gráfico de residuales contra los d. Figura 2.1 .

El gráfico muestra los residuales que se apartan de cero en forma sistemática -, +, -, lo que hace presumible algún desajuste. Además de lo anterior el análisis de residuales permite otras pruebas, independientemente de la escala que no resulta importante para visualizar muchos desajustes de los modelos (Draper & Smith 1966).

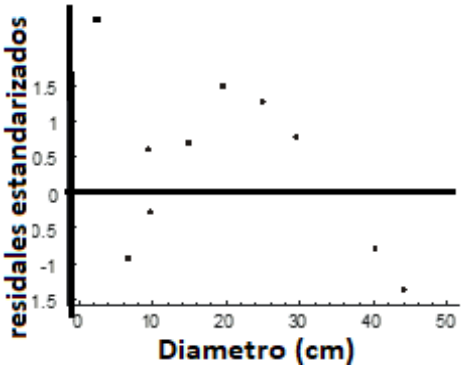


Figura 2.14. Residuales para el modelo generado con los datos de la Tabla 2.3

Para lo anterior se acude a una serie de gráficos prototipo, que posibilitan diagnósticos intuitivos reforzables posteriormente. Cuando el modelo es apropiado los residuales tienden a localizarse en una banda alrededor del eje X_i contra los e_i Fig 2.15.

Cuando se dan situaciones de alejamiento de la regresión lineal se pueden dar bandas circulares, de ancho más o menos constante, para la franja de residuales, con tendencia a una variación sistemática entre positiva y negativa de los e_i Figura (2.16)

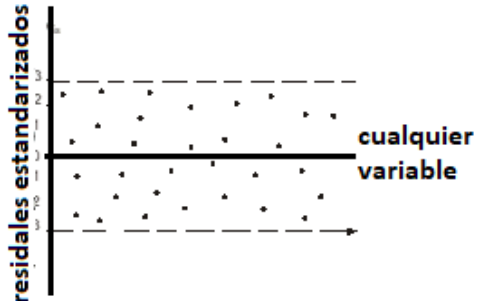


Figura 2.15. Prototipo del comportamiento adecuado de los residuales

Esta situación puede mostrar entre otras cosas la necesidad de una función de regresión curvilineal.

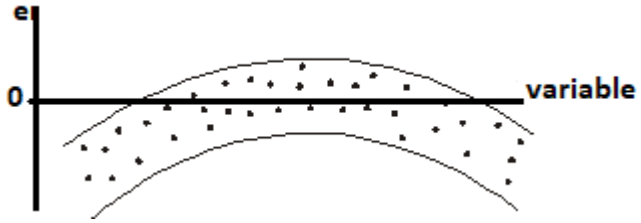


Figura 2.16. Residuales con alejamientos sistemáticos positivos y negativos

2.9.2.2 No homogeneidad de varianza

Otro gráfico podría mostrar una situación en la cual se observa una tendencia constante en el crecimiento o disminución de los residuales con el tamaño de las variables independientes o dependientes, mostrando entonces la llamada heterocedasticidad. Figura 2.17

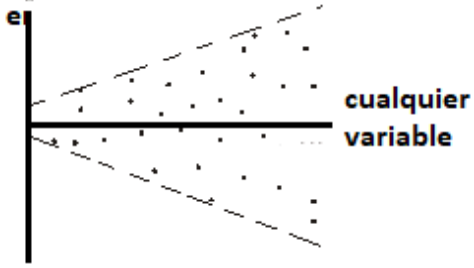


Figura 2.17. Residuales con presencia de heterocedasticidad

Existen otros diagnósticos, algunos simples, para detectar heterocedasticidad:

1. Ordenar los datos de acuerdo con la variable a investigar, y separar el arreglo en dos o más muestras de igual o desigual tamaño.

2. Usar una de las siguientes pruebas: a. Prueba de *F* para comparación de varianzas, entre dos grupos:

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ contra } H_a : \sigma_1^2 \neq \sigma_2^2 \quad (2.96)$$

el grupo 1 para los datos de 1 a k, el grupo 2 de $k+1$ a n.

$$\text{si } F^* = \frac{s_1^2}{s_2^2} > F_{(\alpha/2;\, n_1-1,\, n_1-1)} \rightarrow \text{rechase } H_0 \quad (2.97)$$

Existen otras pruebas como las de Bartlett y la de Cochran que calculan unos estadísticos T_1, T_2 basado en *k* agrupamientos de γ observaciones, para lo cual existen tablas de k y γ. Es de notar que las 3 pruebas reseñadas presentan una alta sensibilidad a la no normalidad de los residuales. Pero la más contundente es la prueba de Golfeld–Quandt. Muy similar, a la reseñada antes, es la reportada por (Judge, et al, 1987) bajo ese nombre basada en el estadístico $\left(\hat{\sigma}_1^2 / \hat{\sigma}_2^2\right)$ de aquella, reportada bajo la asunción que bajo H_1 la muestra podría ser particionada en dos subconjuntos de observaciones con la varianza de los errores diferente en cada uno,

mientras permanece constante dentro de cada uno de ellos. Los estimados de las varianzas serían computadas por medio de dos regresiones separadas en cada subconjunto. Para esta prueba la hipótesis alternativa es $H_1 : \sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_n^2$ o sea que las observaciones pueden ser ordenadas de acuerdo con varianzas crecientes. Aunque la prueba se plantea acá para dos subgrupos la base es la misma. Los pasos propuestos para implementar la prueba son los siguientes:

1. Asumir que H_1 es cierta y ordenar las varianzas en orden creciente.
2. Omitir m observaciones centrales.
3. Correr dos regresiones separadas, una con $(n-m)/2$ observaciones y la otra con las últimas $(n-m)/2$ observaciones.
4. Se calcula un pseudoestadístico $F : F_{GQ} = MSE_2 / MSE_1$ con MSE_i; $i = 1,\ 2$ en las dos regresiones se compara con un $F_{(1-\alpha/2;((n-m-2p)/2),((n-m-2p)/2))}$, con la cual se toma la decisión pertinente.

El valor de m conlleva que grandes valores de él tienden a incrementar la potencia de la prueba a través de un incremento en el valor de F pero la decrece en términos de los grados de libertad. Los autores dan valores de $m = 4$ para $n = 30$ y $m = 10$ para $n = 60$, incluso ni siquiera se requiere el mismo número de observaciones para las dos regresiones. Por ultimo parece que la prueba GQ es exacta y no descansa en propiedades asintóticas. En R:

```
library(lmtest)
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

gqtest(mode1, point = 0.5, fraction = 0, alternative = c("greater", "two.sided", "less"),
   order.by = NULL, data = list())

        Goldfeld-Quandt test

data:  mode1
GQ = 9.8691, df1 = 3, df2 = 3, p-value = 0.04605
alternative hypothesis: variance increases from segment 1 to 2
```

2.9.2.3 No independencia en los términos del error

2.9.2.3.1 Método gráfico. Para encontrar esta falta de ajuste se acude a veces a una ubicación temporal u ordenación de los datos de menor a mayor, ya que el tiempo o el orden a veces influyen en los errores, haciéndolos dependientes unos de otros en forma sistemática. Entonces se acude al gráfico de residuales contra tiempo y se espera un comportamiento como el de la Figura 2.**,** si los términos del error no son independientes. Una situación como esta es prototípica de efectos de variables relacionadas con el tiempo sin que éste se hubiera considerado en el modelo (posible autocorrelación de los errores).

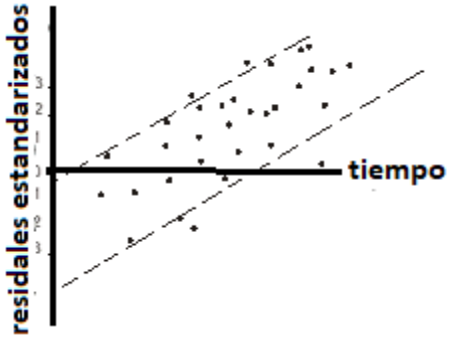


Figura 2.18. Presencia de autocorrelación de los errores.

Además este tipo de gráficas de variación sistemática de los e_i contra las variables puede mostrar otros desajustes, generalmente omisión de variables. Cuando se halla presente la autocorrelación de errores se afecta así el análisis de regresión: 1. Los estimadores minimocuadráticos, aunque insesgados ya no tienen varianza mínima. 2. Los valores estimados para la varianza de los coeficientes de regresión $s^2(\beta_k)$ pueden subestimar seriamente las varianzas de los estimadores mínimos cuadráticos de β_k, y 3. Los intervalos de confianza y las pruebas de hipótesis pueden perder validez (Canavos 1987).

2.9.2.3.2 Estadístico de Durbin-Watson. Esta prueba, asume modelos autorregresivos del error de primer orden de forma:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \mathcal{E}_t \\ \mathcal{E}_t &= \rho_{\mathcal{E}_{t-1}} + \mu_t \end{aligned} \qquad (2.98)$$

ρ es un parámetro tal que $|\rho| < 1$, u_t términos independientes $N(0,\ \sigma^2)$. El modelo es idéntico al presentado en la ecuación (2.1) excepto para los términos del error que consisten en una fracción del error previo si $\rho > 0$ y un término de disturbio, a veces llamado ruido, u_t. El parámetro ρ es llamado parámetro de autocorrelación y se considera en cierta forma como una medida de asociación entre los e_i. Este estadístico plantea las hipótesis siguientes: $H_0 : \rho = 0$ no hay autocorrelación de errores; $H_a : \rho \neq 0$ hay autocorrelación positiva. No se trata de un procedimiento exacto pero existen tablas de este estadístico con unos límites inferior (d_l) y superior (d_u) para un nivel α, unos grados de libertad y un número *k* de variables dadas, de modo que se tome la decisión más acertada así; llamando *D* al valor obtenido para el estadístico en el modelo, calculado por medio de la expresión:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2} \qquad (2.98.1)$$

asumiendo que $\sum_{t=1}^{n} e_t^2 \simeq \sum_{t=2}^{n} e_t^2 \simeq \sum_{t=1}^{n} e_{t-1}^2$. Desarrollándolo y simplificándolo se puede ver que

$$d \cong \frac{2\sum_{t=1}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \cong 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}\right) = 2(1-\hat{\rho}) \quad (2.98.2)$$

En que

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \quad (2.99)$$

Como $-1 \le \hat{\rho} \le 1$ reemplazando en (118) se obtiene que $0 \le \boldsymbol{d} \le 4$, además si la correlación es positiva, d varía entre 0 y 2; y si es negativa d varía entre 2 y 4. Durbin y Watson demostraron que aunque el estadístico d depende de los datos, está entre dos valores *dl* y *du*, tabulados en función del número de parámetros del modelo, nivel de significancia de la prueba, tipo de prueba de hipótesis y número de datos.

De la tabla construida por Durbin-Watson se leen d_u y d_l tales que $0 < d_l < d_u < 2$, con los cuales se define la región crítica para los diferentes tipos de prueba de hipótesis:
Para dos colas:

$H_0: \rho = 0$ / $H_1: \rho \neq 0$

- $\boldsymbol{d}_u^* < \boldsymbol{d}_{cal} < 4 - \boldsymbol{d}_u^*$ — *No* se rechaza $H_0 \Rightarrow \rho = 0$
- $\boldsymbol{d}_{cal} < \boldsymbol{d}_L^*$ *o* $\boldsymbol{d}_{cal} > 4 - \boldsymbol{d}_L^*$ — *Se* rechaza $H_0 \Rightarrow \rho \neq 0$
- $\boldsymbol{d}_L^* < \boldsymbol{d}_{cal} < \boldsymbol{d}_u^*$ *o* $4 - \boldsymbol{d}_u^* < \boldsymbol{d}_{cal} < 4 - \boldsymbol{d}_L^*$ — No puede tomarse una decision (ND)

De cola a derecha

$H_0: \rho = 0$ / $H_1: \rho > 0$

- $\boldsymbol{d}_{cal} > \boldsymbol{d}_u^*$ — *No* Se rechaza $H_0 \Rightarrow \rho = 0$
- $\boldsymbol{d}_{cal} < \boldsymbol{d}_L^*$ — *Se* rechaza $H_0 \Rightarrow \rho > 0$
- $\boldsymbol{d}_L^* \le \boldsymbol{d}_{cal} \le \boldsymbol{d}_u^*$ — No puede tomarse una decision (ND)

De cola izquierda

$H_0: \rho = 0$ / $H_1: \rho < 0$

- $\boldsymbol{d}_{cal} < 4 - \boldsymbol{d}_u^*$ — *No* se rechaza $H_0 \Rightarrow \rho = 0$
- $\boldsymbol{d}_{cal} > 4 - \boldsymbol{d}_L^*$ — Se rechaza $H_0 \Rightarrow \rho < 0$
- $4 - \boldsymbol{d}_u^* \le \boldsymbol{d}_{cal} \le 4 - \boldsymbol{d}_L^*$ — No puede tomarse una decision (ND)

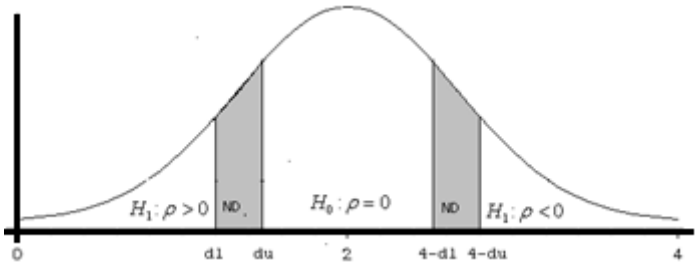


Figura 2.19 Regiones de rechazo y aceptación en la prueba de D-W

Si el estadístico de prueba cae en la región de indecisión existe la opción de calcular el punto exacto de la región de significancia utilizando los métodos de (D-W, 1971),

Fig. 2.19. Alternativamente puede considerarse la región de indecisión como parte de la región de rechazo.

Con muestras $d \simeq 2(1-\gamma_1)$, γ_1, coeficiente de autocorrelación de orden 1 de los residuales de tal manera que si $\gamma_1 = 0 \rightarrow d = 2$, si $\gamma_1 > 0$ d se sitúa entre 0 y 2 y sí $\gamma_1 < 0 \rightarrow d < 4$; con las siguientes reglas de decisión: si $d < d_l$ se rechaza H_0, si $d > d_u$ no se debe rechazar H_0, si $d_l < d < d_u$ prueba no concluyente. Presenta el inconveniente que aunque exista fuerte dependencia entre residuales γ_1 puede resultar bajo, a veces. En R para estudiar la Correlación serial de residuales

```
resid(modelo1)
1           2           3           4           5           6
-5.405077789  1.994273170  2.393624129  4.792975088  4.192326048  2.591677007
          7           8           9          10          11
-0.008972034 -2.609621075 -3.045337406 -1.005726830 -3.890140307
```

Esta prueba en R se encuentra en la librería *car* (Instal packages)

```
 library(car)
durbinWatsonTest(modelo1)
 lag Autocorrelation D-W Statistic p-value
   1       0.4336288     0.7582298   0.016
 Alternative hypothesis: rho != 0
```

Parece que hay correlación serial por el p-value<0.05

2.9.2.4 Observaciones remotas o *outliers*

Una observación remota es un dato extremo que parece apartarse de una norma común, quedando aislado y bastante lejos de los otros datos en el diagrama de dispersión. Existen algunas normas, como las intuitivas de Draper & Smith (1966) quienes proponen sospechar de observaciones con valores mayores de 3-4 desviaciones normalizadas o estandarizadas desde el cero. Otras más eficientes, determinan el impacto de una observación en los coeficientes de regresión estimados, como la llamada distancia de Cook, que se muestra incipientemente más adelante.

2.9.4.1 Forma gráfica para detectar observaciones remotas. Los valores normalizados del error $e_k > 4$ pueden marcarse como posibles observaciones remotas, debiendo evaluarse esta suposición. De los residuales estandarizadas para el ejemplo analizado el mayor es -1.726 para el primer valor, por lo cual no resulta sospechoso como posible remoto. El gráfico típico en que se da tal situación se parece al de la Fig 2.20.

Las observaciones remotas pueden crear gran dificultad, al empujar desproporcionadamente la línea ajustada de regresión, pero sólo serían descartables ante evidencias como mala toma de datos o procesamiento de los datos. Pueden constituirse de otro lado en fuente importante de decisiones como evidencia de interacciones con variables omitidas en el modelo.

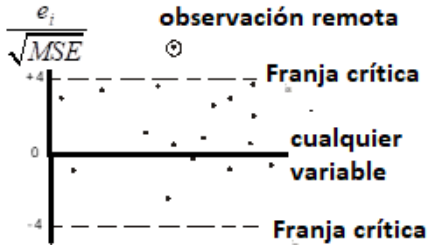


Figura 2.20. Detección de observaciones remotas

2.9.5 Vector palanca (*Leverage*)

Los puntos incrementan su influencia, estando en su entorno, en la medida que se alejen del valor medio de x (a izquierda o derecha). La medida más familiar para cuantificarlo es

$$\bar{h} = \frac{2p}{n}; \quad \text{si } p = 2 \rightarrow \bar{h} = \frac{4}{n} \quad (2.100)$$

Llamado vector palanca o *leverage*, en que p es el número de variables a usar en el modelo. Aquellos valores de h que superen este valor son indicadores de una alta influencia en el modelo. Como práctica cuidadosa pueden descartarse. Para cuantificarlo se propone otra medida del leverage, la proporcionalidad de x y $\bar{x}$ contra SSX: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX}$.

En R puede verse un conjunto de medidas influenciales, no descritas acá, mediante una función propia de él. Por ejemplo

```
influence.measures(modelo1)
Influence measures of
•               lm(formula = h ~ d) :
•           dfb.1_     dfb.d    dffit cov.r   cook.d    hat inf
•       1  -1.09805  0.872875 -1.10607 0.760 4.64e-01 0.2410   *
•       2   0.24670 -0.172977  0.25779 1.397 3.59e-02 0.1654
```

Según esta tabla, es muy probable que la observación 1, única con asterisco sea altamente influencial.

2.9.2.6 No normalidad de los términos del error

Pequeños alejamientos de la normal no crean problemas serios a causa de la normalidad asintótica debida a un buen tamaño de la muestra, pero los grandes sí pueden afectar enormemente el modelo. En primer lugar, la variación aleatoria puede resultar particularmente perjudicada cuando se estudia la naturaleza de la distribución de probabilidades a menos que el tamaño de la muestra sea muy grande, y aun peor, otros tipos de alejamiento de la normalidad pueden afectar la distribución de los residuales debido a un modelo inapropiado o por efectos de la heterocedasticidad. Existen varias formas de detectar el problema, por ejemplo con un histograma de residuales para ver su configuración, aunque para valores pequeños de la muestra (n) el método puede desorientar un poco al observador. Otro procedimiento es la gráfica de los residuales unos encima de otros, *Overall*, propuesto por Draper (1966)

para compararla con tablas preparadas para ello. Otra alternativa es el uso del papel de probabilidades normales en el cual se deberá tener una línea recta (Daniel & Wood 1980). Y, de pronto, el más conocido, implica el uso de la desviación normal unitaria:

$$s^2 = \frac{\sum (e_i - \bar{e})^2}{n-p} = \frac{\sum e_i^2}{n-p} \quad (2.101)$$

que estima a σ^2. (En el caso de la regresión simple $p = 2$, en el caso de la múltiple se verá que p es igual al número de variables independientes). Con esta concepción, el 68% de los residuales normalizados se localizarán entre +1 y -1 y aproximadamente el 95% entre +2 y -2. Como una prueba más objetiva se presenta el concepto de los residuales esperados bajo normalidad que se pueden definir de acuerdo con un postulado estadístico que establece que: para una variable aleatoria normal con media cero y s estimado como $\sqrt{MSE}$, el valor esperado de la *i*-ésima observación ordenada en una muestra aleatoria de la más pequeña a la más grande se da aproximadamente por:

$$E\left(\frac{e_i}{\sqrt{MSE}}\right) = Z\left(\frac{i-0.375}{n+0.25}\right) \therefore E(e_i) = \sqrt{MSE}\, Z\left(\frac{i-0.375}{n+0.25}\right) \quad (2.102)$$

en la cual $Z^*\left(\frac{i-0.375}{n+0.25}\right)$ es el percentil de Z para la observación *i* de *n*. Se presenta como ejemplo para los datos antes analizados los residuales esperados bajo normalidad en la.

Tabla 2.4 Residuales esperados bajo normalidad

i	e_i ascendente	*K*	*Z*(*K*)	$E(e_i)$
1	-5.405	0.055	-1.598	-5.797
2	-3.890	0.144	-1.063	-3.854
3	-3.045	0.233	-0.729	-2.644
4	-2.609	0.322	-0.462	-1.676
5	-1.005	0.411	-0.225	-0.816
6	-0.009	0.5	0.000	0.000
7	1.994	0.588	0.222	0.807
8	2.393	0.677	0.459	1.666
9	2.591	0.766	0.726	2.632
10	4.192	0.855	1.058	3.838
11	4.792	0.944	1.589	5.765

i = orden ascendente, e_i = residual según el orden i, $k = (i-0.375)/(n+0.25)$, $E(e_i) = MSE^{0.5} * Z(k)$.

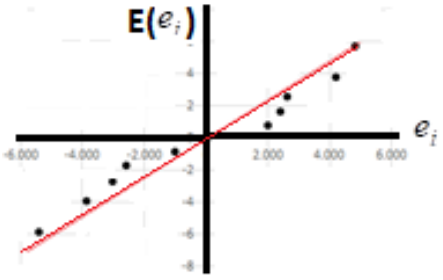


Figura 221. Gráfica de residuales ascendentes versus residuales esperados bajo normalidad

Con los valores dados se hace una gráfica de valores esperados $E(e_i)$ contra los residuales ordenados. Fig 2.21. Si el gráfico parece seguir una línea recta se acepta normalidad de los términos del error. Para complementar lo anterior, Neter *et al*. (1983) proponen que valores de $r \geq 0.9$ lo corroboran. Para el caso se tuvo un $r \geq 0.985$, lo cual presupone entonces la normalidad de los términos del error, aún para este modelo que ya se aparta de la linealidad.

Como un criterio más elaborado se conoce la tabla de Blom $(s.\ d)$, de valores críticos para distintos niveles de significación y tamaños de muestra para calificar aproximadamente este coeficiente de correlación, Tabla 2.

Tabla 25. Valores críticos de Blom para el coeficiente de correlación en distintos niveles de significación α y distintos tamaños de muestra.

n	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}10$	n	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}10$
10	0.879	0.918	0.934	40	0.959	0.972	0.977
15	0.910	0.939	0.951	50	0.966	0.977	0.981
20	0.926	0.951	0.960	60	0.971	0.980	0.984
25	0.939	0.959	0.966	70	0.976	0.984	0.987
30	0.947	0.964	0.971	100	0.982	0.987	0.989

2.9.2.7 Omisión de variables

Es posible graficar residuales contra variables omitidas. (En el ejemplo se graficó e_i contra edad, (Fig 2.22).

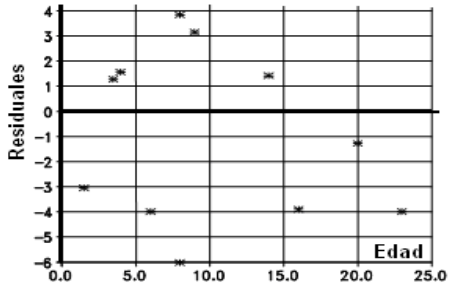


Figura 2.22. Gráfica de residuales contra edad, variable no usada en el modelo

Se observa cierto cambio en la tendencia de los residuales, ahora más aleatoriamente agrupados, lo que indica el efecto de esta variable omitida, lo que se esperaba, si se tiene en cuenta que la "variable tiempo" está asociada con casi todos los fenómenos en que se dan con el crecimiento, o aún con datos en que su variación es concomitante con él. Cuando se asigna un orden temporal a las observaciones y se presenta una banda continua, para los residuales, como la de la Fig 2.15, se supone que no hay desajustes en el modelo, o que a largo plazo el tiempo no influenció los datos. Si se da un comportamiento diferente del presentado como en la Fig 2.17 se concluye lo contrario, implicando esto el manejo de la regresión ponderada, o que debió incluirse algún termino lineal en el tiempo. Si se da un comportamiento como el de la Fig 2.16, podría haberse incluido en el modelo otro término cuadrático. Además, es posible tener de pronto combinaciones de los efectos mostrados como la llamada variación estacional en la cual los residuales parecen asumir comportamientos organizados en estructuras como las mostradas en la Fig 2.23.

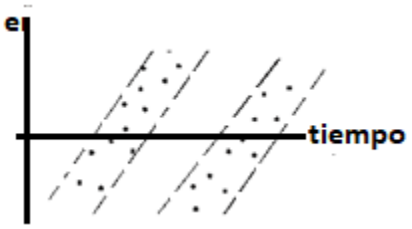


Figura 2.23. Presencia de residuales con variación estacional

2.9.3 Pruebas de falta de ajuste

Con frecuencia conviene verificar el ajuste de los datos al modelo lineal. Si ello no fuera así, el *MSE* ya no sería un estimador insesgado de σ^2 y las inferencias pertinentes resultarían inapropiadas (Gómez 1989). Se propone entonces una prueba de bondad de ajuste con base en las asunciones fijadas al modelo: independencia en las observaciones Y_i, distribución normal de las Y_i para cada nivel de X_i, y homocedasticidad en las distribuciones anteriores (Neter *et al*. 1983). Para esta prueba se recomienda que los investigadores obtengan por norma varias observaciones para cada nivel de la variable X_i (Walpole & Myers 1984).

2.9.3.1 Hipótesis consideradas para falta de ajuste

Las hipótesis analizadas son:

$$\left.\begin{aligned} H_o &: \hat{Y} = \beta_0 + \beta_1 X \\ H_a &: \hat{Y} \neq \beta_0 + \beta_1 X \end{aligned}\right\} \quad (2.103)$$

Si H_0 se cumple entonces $E\left(s^2\right) = \sigma^2$. Si H_a se cumple:

$$E(s^2) = \sigma^2 + \frac{\sum\left(\hat{Y} - \beta_0 - \beta_1 X\right)^2}{n - 2} \quad (2.104)$$

Si el modelo no es correcto, los residuales tienen dos componentes: σ^2 constituido por los errores aleatorios ya estudiados y, el constituido por los errores sistemáticos formados por las diferencias entre los valores aportados por el modelo supuesto y el verdadero modelo: $\hat{Y}_i - \beta_0 - \beta_1 X_i$.

Los errores aleatorios constituyen el error puro, y los sistemáticos la falta de ajuste entre los modelos, y juntos la suma ya definida *SSE*.

2.9.3.1.1 El error puro. — Al trabajar con el concepto de poblaciones de Y_i, para cada valor de X, los errores puros deben extraerse de ahí, por lo cual la suma de cuadrados del error puro; para un número dado de clases c de la variable independiente y n_j valores de Y_{ij} para cada X_i; se obtendrá como:

$$SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 \quad (2.105)$$

Esta suma tendrá $(n-c)$ grados de libertad, ya que hay $(n_1 - 1)$ grados de libertad asociados con la variable X_1, $(n_2 - 1)$ con la variable X_2, y así para el resto, de tal

manera que $\sum(n_j - 1) = \sum n_j - c = n - c,\ (j = 1,\ldots,\ c)$, con lo cual error puro medio cuadrático *MSPE* será:

$$MSPE = \frac{SSPE}{n - c} \quad (2.106)$$

El término error puro se justifica porque *MSPE* es un estimador insesgado que coincide con el concepto de σ^2, o sea que mide la variabilidad de las distribuciones de Y sin ligarse a ninguna asunción acerca de la naturaleza de la función de regresión, resultando una medida pura de la varianza del error.

2.9.3.1.2 Componente del error por falta de ajuste del modelo. — El segundo componente de la *SSE* sale de la fórmula (2.104). Se llamará suma de cuadrados por falta de ajuste *SSLF* y se obtiene como:

$$SSLF = SSE - SSPE = \sum_{j=1}^{c} n_j \left(\bar{Y}_j - \hat{Y}_j\right)^2 \qquad (2.107)$$

suma ponderada de cuadrados; con ponderadores n_j de las desviaciones $\bar{Y}_j - \hat{Y}_j$, que representan la diferencia entre la media y el valor ajustado por el modelo. Mientras más cercanos estos valores mejor la evidencia del ajuste. Los grados de libertad asociados con *SSLF* son $(c-2)$, ya que se tienen c clases y se pierden 2 grados de libertad por los parámetros estimados β_0 y β_1. Entonces el cuadrado medio de la falta de ajusta *MSLF* se obtendrá como: $MSLF = SSLF/(c-2)$.

2.9.3.2 ANAVA para verificar falta de ajuste

Se puede configurar según la Tabla 2.. La prueba de *F* para la falta de ajuste sigue la rutina tradicional y se obtiene con: $F = \frac{MSLF}{MSPE}$. Ya se había anotado que *MSPE* tiene σ^2 como esperanza sin importar la naturaleza de la función de regresión. Se puede ver entonces que:

$$E(MSLF) = \sigma^2 + \frac{\sum_{j=1}^{c} n_j \left[E(Y_j) - (\beta_0 + \beta_1 X_j)\right]^2}{c-2} \quad (2.108)$$

en la que $E(Y_j)$ será la verdadera media de la distribución de Y en X_j y $(\beta_0 + \beta_1 X_j)$ es el valor esperado bajo la suposición del modelo lineal. Entonces si esto último ocurre (linealidad del modelo) , el segundo término del lado derecho deberá valer 0, con lo que $E(MSLF) = \sigma^2$. Pero si la función no es lineal:

$$E(Y_j) \neq \beta_0 + \beta_1 X_j \text{ y la } E(MSFL) > \sigma^2 \quad (2.109)$$

La prueba de *F* se evalúa en la forma acostumbrada:

$$\left.\begin{array}{l} H_0 : E(Y) = \beta_0 + \beta_1 X \text{ vs } H_a : E(Y) \neq \beta_0 + \beta_1 X \\ \text{si } F^* > F_{[(1-\alpha);(c-2);(n-c)]} \rightarrow H_a \end{array}\right\} \quad (2.110)$$

Por ejemplo, en un bosque natural degradado en Piedras Blancas se obtuvieron los siguientes datos de diámetros y volúmenes de 36 árboles, con los cuales se generó

el modelo mostrado al final de la siguiente salida del computador y los datos se organizaron como lo muestra la Tabla 2. (Puche 1989).

Tabla 2.6. Diámetros y volúmenes de 36 árboles de un bosque natural degradado en Piedras Blancas.

	Volumenes	nj	*SSPE*
9	0.0288; 0.0283; 0.0455; 0.0384	4	2.82E+00
10	0.0489; 0.0334; 0.0265; 0.0175; 0.0298	5	5.30E+00
11	0.0408; 0.0750; 0.0674; 0.0281	4	1.46E+00
12	0.0851; 0.0483; 0.0568; 0.0493	4	8.92E+00
12.5	0.0914; 0.0468; 0.0717	3	9.99E+00
14	0.0631; 0.1259; 0.0826; 0.0627; 0.1050; 0.0776	6	3.10E+00
15	0.0705; 0.1061; 0.0514; 0.1417; 0.0862	5	4.81E+00
15.5	0.1611	1	0
19	0.1984; 0.1858	2	8.19E-01
21	0.1250; 0.1793	2	1.47E+00

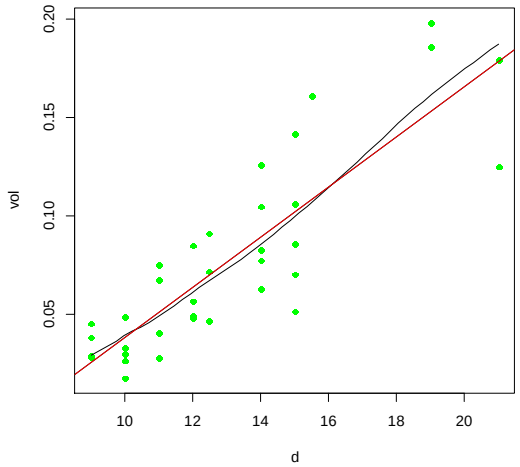
Que insinúa una tabla de falta de ajuste como la mostrada. Tabla 2..

Tabla 2.7 Anava por falta de ajuste.

Fuente de variación	gl.	Sumas de cuadrados *SS*..	Cuadrados medios *MS*..
Regresión	1	*SSR*	*MSR*
Error	$n-2$	*SSE*	*MSE*
Falta de ajuste	$c-2$	*SSLF*	*MSLF*
Error puro	$n-c$	*SSPE*	*MSPE*
Total	$n-1$	*SSTO*	

Análisis de regresión lineal $Y = a + bX$. En R

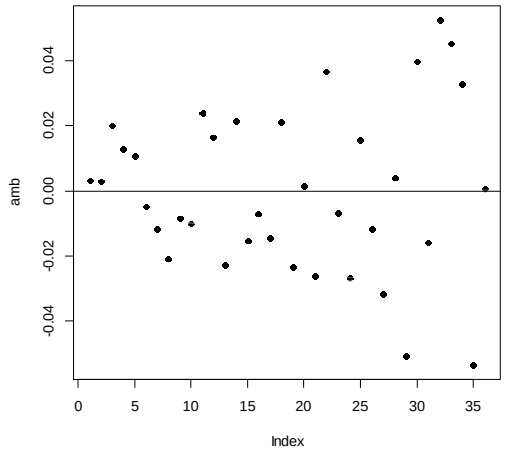
```
vodi<-read.table("clipboard")
attach(vodi)
names(vodi)
[1] "d"   "vol"
scatter.smooth(d,vol,col="green",pch=16)
abline(lm(vol~d),col="red")
```




```
rlsiaj<-lm(vol~d)
summary(modelo1)
Call:
lm(formula = vol ~ d)
Residuals:
      Min        1Q    Median        3Q       Max
-0.053581 -0.015552 -0.002061  0.017304  0.052589
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
        (Intercept) -0.089339   0.018010  -4.961 1.93e-05 ***
d            0.012758   0.001341   9.517 4.08e-11 ***
```

```
Residual standard error: 0.0255 on 34 degrees of freedom
Multiple R-squared:  0.7271,    Adjusted R-squared:  0.7191
F-statistic: 90.58 on 1 and 34 DF,  p-value: 4.078e-11
```

eis<-resid(rlsiaj)
plot(eis,pch=16)
abline(h=0)



Este gráfico podría sugerir heterocedasticidad, pero son mejores otros análisis. Como existen replicaciones para algunos niveles de la variable independiente se puede estimar la *SSPE*, vista. No es más que el resultante de un análisis de varianza, haciendo *d* como factor
fac.d<-factor(d)

```
fac.d
 [1] 9    9    9    9    10   10   10   10   10   11   11   11   11   12   12   12   12   12.5
12.5 12.5 14   14   14   14   14   14
[27] 15   15   15   15   15   15.5 19   19   21   21
Levels: 9 10 11 12 12.5 14 15 15.5 19 21
modelo1<-aov(v~fac.d)
summary(modelo1)
            Df  Sum Sq  Mean Sq F value   Pr(>F)
fac.d        9 0.06746 0.007496   14.38 5.26e-08 ***
Residuals   26 0.01355 0.000521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entonces el *SSPE* = 0.01355 con 26 grados de libertad. Recordamos el aov del modelo ajustado

```
summary(aov(rlsiaj))
            Df  Sum Sq Mean Sq F value   Pr(>F)
d            1 0.05890 0.05890   90.58 4.08e-11 ***
Residuals   34 0.02211 0.00065
```

La diferencia entre 0.02211 y 0.01355 da la medida de falta de ajuste, 0.02211-0.01355 = 0.00856. En R podemos también comparar ambos modelos para ver si existen diferencias en sus variables explicatorias.

anova(rlsiaj,modelo1)

```
Analysis of Variance Table
Model 1: v ~ d
Model 2: v ~ fac.d
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     34 0.022111
2     26 0.013551  8 0.0085595 2.0528 0.0793 .
```

Entonces no hay problemas de falta de ajuste por el p-value >0.05. En R también una simple tabla de anava mostrando las SS.. por falta de ajuste (Lack of fit), así

anova(lm(v~d+fac.d))

```
Analysis of Variance Table

Response: v
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
d          1 0.058904 0.058904 113.0160 5.842e-11 ***
fac.d      8 0.008559 0.001070   2.0528    0.0793 .
Residuals 26 0.013551 0.000521
```

que conducen a la misma conclusión, no parece haber problemas por falta de ajuste. En R con la librería alr3

```
library(alr3)
summary(modelo1)
Call: lm(formula = vol ~ d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.053581 -0.015552 -0.002061  0.017304  0.052589

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.089339   0.018010  -4.961 1.93e-05 ***
d            0.012758   0.001341   9.517 4.08e-11 ***
Residual standard error: 0.0255 on 34 degrees of freedom
Multiple R-squared:  0.7271,    Adjusted R-squared:  0.7191
F-statistic: 90.58 on 1 and 34 DF,  p-value: 4.078e-11
```

TABLA ANAVA POR FALTA DE AJUSTE.

pureErrorAnova(modelo1)#Funcion R para falta de ajuste

```
Analysis of Variance Table
Response: vol
             Df   Sum Sq  Mean Sq  F value    Pr(>F)
d             1 0.058904 0.058904 113.0160 5.842e-11 ***
Residuals    34 0.022111 0.000650
 Lack of fit  8 0.008559 0.001070   2.0528    0.0793 .
 Pure Error  26 0.013551 0.000521
```

El modelo no presenta falta de ajuste. Se ensayará el modelo lineal: vol vs dap^2

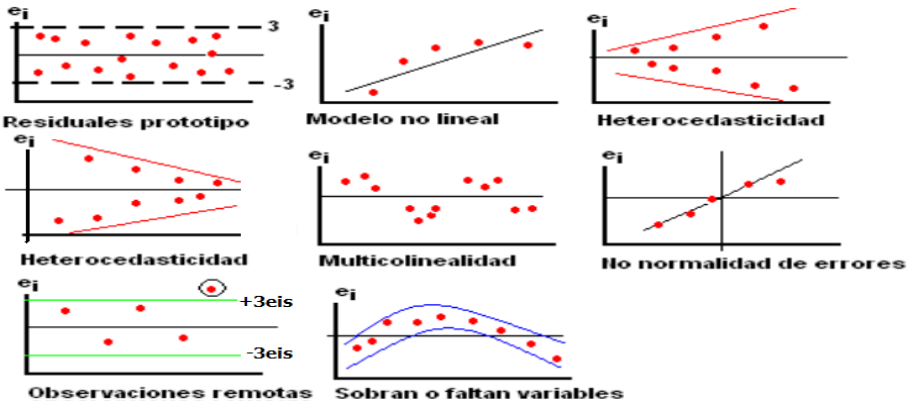
```
 lm(formula = vol ~ I(d^2))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0007674  0.0095570  -0.080    0.936
I(d^2)       0.0004321  0.0000471   9.173 1.01e-10 ***
Residual standard error: 0.02619 on 34 degrees of freedom
Multiple R-squared:  0.7122,    Adjusted R-squared:  0.7038
F-statistic: 84.15 on 1 and 34 DF,  p-value: 1.014e-10

pureErrorAnova(modelo3)# Este modelo presentará falta de ajuste

Analysis of Variance Table

Response: vol
             Df   Sum Sq  Mean Sq  F value    Pr(>F)
I(d^2)        1 0.057700 0.057700 110.7069 7.276e-11 ***
Residuals    34 0.023314 0.000686
 Lack of fit  8 0.009763 0.001220   2.3415   0.04824 *
 Pure Error  26 0.013551 0.000521
```

En resumen, se muestran los gráficos analizados:



2.10 Medidas remediales

Cuando se observan los desajustes del modelo se debe asumir algún criterio que permita volver el modelo lo más apto posible.

2.10.1 Modelo lineal no apropiado. Se abandona el modelo y se busca otro más apropiado. Se hacen transformaciones a las variables o a los datos para que el modelo resultante se vuelva lineal. Con esta acción se puede oscurecer el modelo, aparecer interrelaciones no deseadas, o problemas de interpretación de resultados. Una primera transformación puede consistir en adicionar un término cuadrático, que pasa el modelo a una regresión múltiple. Como ejemplo se propone pasar del lineal al modelo cuadrático:

$$Y = \beta_0 + \beta_1 X \text{ a } E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 \quad (2.111)$$

Otra transformación se da asumiendo la función exponencial:

$$E(Y) = \beta_0 * X^{\beta_1} \rightarrow Y' = \log \beta_0 + \beta_1 * \log X \rightarrow Y' = \log Y \quad (2.112)$$

Los modelos multiplicativos, como el último propuesto, conducen a transformaciones logarítmicas de amplio uso, por ejemplo, en el sector forestal, por la ley de alometría reinterpretada. Se tienen actualmente algunas propuestas para resolver parte de los problemas por ellos planteados, como estimar un promedio geométrico más bien que uno aritmético, en que el primero resulta siempre subestimado con respecto al segundo.

2.10.1.1Transformaciones logarítmicas. Los modelos multiplicativos, como el propuesto:

$$E(Y) = \beta_0 X^{\beta_1} \rightarrow Y' = \log \beta_0 + \beta_1 \log X \rightarrow Y' = \log Y \quad (2.112.1)$$

conducen a transformaciones logarítmicas de amplio uso, por ejemplo, en el sector forestal por la ley de alometría reinterpretada. Se tienen actualmente algunas propuestas para resolver parte de los problemas por ellos planteados, como estimar un promedio geométrico más bien que uno aritmético, en que el primero resulta siempre subestimado con respecto al segundo.

Este sesgo es ínsito a estos procesos por cuanto los mayores valores son comprimidos en la escala logarítmica y tienden a tener un menor “leverage” que los más pequeños, (Beauchamp y Olson, 1973). Esto ha llevado a algunos autores como Magdwick, 1970, citado por (Satoo y Magdwick, 1982) a proponer algunas

correcciones al respecto, basados en (Heien, 1968). Entonces dado un conjunto de n observaciones en Y y X, tales que:

$$Y = \alpha_0 X^{\beta_1} \quad (2.112.2)$$

se puede reescribir (112) como:

$$Y_i = \alpha_1 X_i^{\beta_1} V_i \quad (2.112.3)$$

Con V_i *un error aleatorio multiplicativo para una distribución lognormal*, justificado en el teorema del Limite Central Multiplicativo que establece, que el producto de n variables aleatorias independientes tienden a una *lognormal* cuando $n \to \infty$ por lo cual, V puede verse como el producto de varios errores independientes y aun con distribuciones diferentes:

$$V = \prod_{i=1}^{n} V_i, \quad \to \ln V = \sum_{i=1}^{n} \ln V_i = \sum_{i=1}^{n} v_i \quad (2.112.4)$$

en que $v_i = \ln V_i \to \mathrm{N}(0, \theta^2)$ y por las teorías de la lognormal se sabe que si v es normal, entonces V es lognormal. Por ello al transformar (2.112.3) como:

$$\ln Y_i = y_i = \beta_0 + \beta_1 \ln X_i + v_i = \beta_0 + \beta_1 x_i + v_i \quad (2.112.5)$$

se debe proceder con esta como un análisis convencional de regresión lineal, teniendo en cuenta las siguientes relaciones entre la media (m) y la varianza (σ^2) de la función madre y la media de su lognormal respectiva, así:

$$\eta = \exp\left(m + \frac{1}{2}\sigma^2 \right) \quad (2.112.6)$$

Como la media de v es 0, la media del error V será:

$$\eta = \exp\left(\frac{1}{2}\theta^2 \right) \quad (2.112.7)$$

Resulta claro entonces que, de acuerdo con lo anterior que (2.112.3):

$$E(Y_i) = \alpha_1 X_i^{\beta_1} \exp\left(\frac{1}{2}\theta^2 \right)_i \neq \alpha_0 X_i^{\beta_1} \quad (2.112.8)$$

por lo cual se debe reparametrizar el modelo. Para ello se reescribe (2.112.3) como:

$$Y_i = \alpha_1 X_i^{\beta_1} U_i \quad 2.112.9$$

y si se requiere que

$$E(Y_i) = \alpha_0 X_i^{\beta_1} \quad (2.112.10)$$

Se debe asumir que $E(U_i) = 1$, con lo cual la relación transformada sería:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (2.112.11)$$

en que los $u_i \to \mathrm{N}\left(\frac{1}{2}\theta^2, \theta^2 \right)$ de (2.112.11) y de que $E(U_i) = 1$, una forma alternativa de escribir (2.112.1) es

$$y_i = \beta_0' + \beta_1 x_i + v_i \quad (2.112.12)$$

en la cual $\beta_0' = \beta_0 + \frac{1}{2}\theta^2$, con $v \to \mathrm{N}(0, \theta^2)$.

Al transformar (2.112.1) en (2.112.12), o (2.112.2) en (2.112.3) y tomando esperanzas aparecerán los estimadores mínimocuadráticos de (2.112.2), así:

$$\hat{\beta}_0' = \frac{\sum_{i=1}^{n} y \sum_{i=1}^{n} x^2 - \sum_{i=1}^{n} x \sum_{i=1}^{n} xy}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2}; \qquad \hat{\beta}_1' = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2} \quad (2.112.13)$$

y un estimado de σ^2 será:

$$s^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0' - \hat{\beta}_1 x_i\right)^2}{n-2} \quad (2.112.14)$$

Con los principios ya conocidos de la regresión entonces se sabe que

$$E\left(\hat{\beta}_0'\right) = \beta_0', \quad E\left(\hat{\beta}_1'\right) = \beta_1; \quad y\ E\left(s^2\right) = \theta^2 \quad (2.122.5)$$

con lo cual, se hace evidente que:

$$E\left(\hat{\beta}_0' + \frac{s^2}{2}\right) = \beta_0' + \frac{\theta^2}{2} = \beta_0 \quad (2.112.16)$$

de tal manera que

$$\hat{\beta}_0 = \beta_o' + \frac{s^2}{2} \quad (2.112.17)$$

es un estimador insesgado de $\hat{\beta}_0$, con lo cual se corrige la distorsión por las transformaciones, al adicionar $\frac{MSE}{2}$ al transformado β_0.

2.10.2 Heterocedasticidad

Cuando es detectada, se modifica el modelo usando el proceso de los mínimos cuadrados ponderados, en el que cada observación se liga a un ponderador o peso particular W_i, cuyo valor puede determinarse empíricamente o derivarse de razonamientos teóricos. También se da el caso en que unas observaciones puedan ponderarse más que otras. El método entonces consiste en minimizar:

$$QW = \sum W_i \left(Y_i - \beta_0 - \beta_1 X_i\right)^2 \quad (2.113)$$

cuyo resultado es:

$$\sum W_i Y_i = b_0 \sum W_i Y_i + b_1 \sum W_i X_i \quad (2.114)$$

$$\sum W_i Y_i X_i = b_0 \sum W_i X_i + b_1 \sum W_i X_i^2 \quad (2.115)$$

De (2.114) y (2.115), que se convierten en las ecuaciones normales se obtienen:

$$b_1 = \frac{\sum W_i X_i - \frac{\sum W_i Y_i \sum W_i Y_i}{\sum W_i}}{\sum W_i X_i^2 - \frac{\left(\sum W_i X_i\right)^2}{\sum W_i}} \quad (2.116)$$

$$b_0 = \frac{\sum W_i Y_i - b_1 \sum \left(W_i X_i\right)}{\sum W_i} \quad (2.117)$$

Son muy utilizadas ponderaciones como $W_i = 1/s_i^2$, en la cual s_i es la desviación estándar residual de Y_i. Otras empíricas pueden obtenerse así: ajustar el modelo sin ponderar y calcular $e_i^2 = \left(Y_i - \hat{Y}_i\right)^2$, luego realizar uno de los siguientes pasos

1. Tabular valores promedios de $w_k = \frac{1}{e_i^2}$, por clases de Y_i.

2. Otra forma muy eficiente es ajustar una regresión que relacione e_i^2 con Y_i así: $\hat{e}_i^2 = b_o + b_1 Y$, con la cual predice la ponderación $w_i = \frac{1}{e_i^2}$ para cada Y_i.

Se recuerda que la *función lm* tiene la siguiente sintaxis:

```
lm(formula, data, subset, weights, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
contrasts = NULL, ...)
formula: para el modelo expresado: y~x1+x2+...+xn.
data:  especificar el dataframe con las variables del modelo.
subset: especificar un subconjunto de observaciones para validar posteriormente el modelo.
weights: especificar los pesos (mínimos cuadrados ponderados).
method: otros métodos diferentes a mínimos cuadrados.
model: si = TRUE decimos a R que guarde en el objeto, la matriz del modelo, la frame,...
contrast: podemos especificar objetos con los que realizar contrastes sobre los parámetros.
```

Otras transformaciones efectivas para la estabilización de la varianza son las propuestas por Neter *et al*. (1983), dadas a continuación:

Hacer $Y' = b_0 + b_1 X$ en que $Y' = \sqrt{Y}$ sí σ² es proporcional a $E(Y_i)$.

Hacer $Y' = b_0 + b_1 X$ en que $Y' = \log_{10} Y$ sí σ es proporcional a $E(Y_i)$.

Hacer $Y' = b_0 + b_1 X$ en que $Y' = \frac{1}{Y}$ sí $\sqrt{\sigma}$ es proporcional a $E(Y_i)$.

En idéntica forma se pueden generar transformaciones para X similares para las propuestas para Y. Un gráfico tomado de los autores mencionados puede servir de guía de acuerdo con el diagrama de dispersión Fig 2.24. : En A se transformaría Y en $Y' = \sqrt{Y}$ ó $\log_{10} Y$ ó $1/Y$. En B se transformaría X en $X' = \sqrt{X}$ ó $\log_{10} X$ ó $1/Y$. En C se transformarían de a una o ambas. Cuando la variable Y es una proporción o un porcentaje se acostumbra la transformación:

$$Y' = 2 * arcsen\sqrt{Y} \qquad (2.118)$$

Figura 2.24. Diagramas de dispersión linealizables

2.10.3 No independencia en los términos del error

Si se da una situación de no independencia como las sugeridas por la Figura 2. y Figura 2. es posible recurrir a ajustar una nueva regresión múltiple con un término lineal en el tiempo. Así para la Figura 2., en vez de:

$$Y = b_0 + b_1 X \rightarrow Y = b_0 + b_1 X + b_2 t \quad (2.119)$$

Cuando se trata de variación estacional se acude al uso de una variable ficticia, agrupando los residuales de acuerdo con su tendencia y asumiendo pendientes comunes para las tendencias observadas. Para ello se puede adicionar al modelo un término de la forma $\delta\left[(t-1)\text{módulo } n_i\right]$ en la cual δ denota un coeficiente por estimar, y, $(t-1)\text{módulo } n_i$ es la variable obtenida hallando el residuo del valor $(t-1)$ por n_i, en la cual t es el tiempo, o el número de orden del residual y n_i los grupos de residuales (Draper & Smith 1966). Por ejemplo: suponga que unas observaciones se agrupan así ordenadamente en una variable t.

Variable t	(1, 2, 3)	(4, 5, 6, 7, 8)	(9, 10, 11)	(12, 13)
((t-1)mod ni)	(0, 1, 2)	(3, 4, 0, 1, 2)	(2, 0, 1)	(1, 0)

En R se hace así: **3%%5= 3; 4%%5= 4, l**a nueva regresión sería:

$$Y = b_0 + b_1 X_i + \delta\left[(t-1)\text{mod}_{n_i}\right] \quad (2.120)$$

Habría que buscar los valores correspondientes a la nueva variable. Neter *et al*. (1983), proponen adicionar una o más variables al modelo, o el uso de variables transformadas, especialmente cuando se dan modelos autorregresivos (series de tiempo), que escapan por ahora al propósito de este libro.

2.10.4 Omisión de variables

Los análisis gráficos estructurados alertan sobre otras posibles definiciones para el modelo, que incluyan otras variables. Por ejemplo, la Fig 2.17 implica que podría incluirse un término cuadrático como variable, de pronto también el tiempo.

2.10.5 Modelos más complicados

Existen otras medidas remediales por uso de modelos más complicados que involucren transformaciones complejas de las variables, uso de variables ficticias, etc. cuando las iniciales propuestas no surtan el efecto deseado, pero deberá estudiarse el efecto de tales transformaciones. Algunas de estas transformaciones ya sugeridas en 2.10.2, se pueden asumir en muchos contextos diferentes al de estabilizar la varianza, con el propósito de lograr un modelo de regresión de forma simple en variables transformadas en vez de uno complicado con las variables originales (Draper & Smith 1981). Cuando no existe una razón *a priori* para seleccionar alguna de las propuestas. Box & Cox (1964) proponen un método empírico basado en el exponente más apropiado para la siguiente función:

$$\left.\begin{aligned} y_i' &= \frac{\left(y_i^r - 1\right)}{\gamma} \text{ para } \gamma \neq 0;\ \text{y} \\ y_i' &= \ln\left(y_i\right) \text{ para } \gamma = 0 \end{aligned}\right\} \quad (2.121)$$

en que y_i' es la transformada a la observación y_i. El valor γ es el que maximiza la siguiente función de verosimilitud:

$$L = -\left(\frac{gl}{2}\right)\ln\left(s_{y'}^2\right) + (\gamma - 1)\left(\frac{gl}{n}\right)\sum\ln\left(y_i\right) \quad (2.122)$$

que produce la mejor transformación a la normalidad. Además, $s_{y_i}^2$ es la varianza de los valores transformados de y_i'. Cuando se analizan varios grupos se reemplaza la

s_y^2, por la varianza dentro de grupos o varianza residual. Se anota que si $\gamma = 1$, la función es una simple transformación lineal, si $\gamma = 1/2$ la función es una transformación raíz cuadrada, si $\gamma = 0$, es una transformación logarítmica y si $\gamma = -1$ se tiene la transformación recíproca. También son famosas otras propuestas como las Box y Tidwell citados por Draper & Smith (1981), que pueden involucrar incluso varias X_j simultáneamente como por ejemplo $Z_t = X_1^{1/2} \ln X_2$. También se acostumbra la llamada escalera de transformaciones para la simetría, de Tukey citada por Finney (1978) y Draper & Smith (1981); en el análisis de regresiones de dosis, para proponer familias de transformaciones de acuerdo con la ecuación:

$$X = \frac{\left(Z^{\lambda} - 1\right)}{\lambda} \quad (2.123)$$

la cual es significativa para todos los valores de λ positivos y negativos incluyendo los límites de $\lambda \to 0$ que configuran: $X = \ln Z$. Los valores más comunes de λ son 0 y 1 y posiblemente muchos datos quedarían adecuadamente representados por otros tales ±2, ±2/3, ±1/2, ±1/3, etc., de forma que:

$$X = Z^{\lambda} \text{ cuando } \lambda \neq 0 \quad X = \log Z \text{ cuando } \lambda = 0 \quad (2.124)$$

Es posible además buscar un óptimo que se hallaría como:

$$\frac{Z^{\lambda} - 1}{Z^* \overline{Z}^{(\lambda - 1)}} = \beta_0 + \beta_1 X \quad (2.125)$$

Se busca *SSR* y se grafica contra λ con lo cual se podría encontrar una buena aproximación a la transformación. Una vez realizadas las nuevas regresiones se desharán las transformaciones, pero se debe siempre revisar de nuevo el comportamiento de los residuales pues las transformaciones en la variable respuesta afectan su distribución incluso llegando al caso de dejar de ser $N\left(0,\ \sigma^2\right)$.

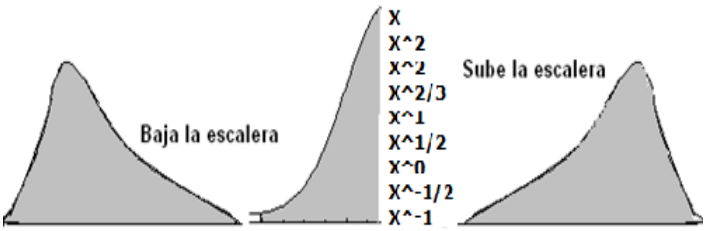


Figura 2.25. Escalera de Tuckey

Cuando la distribución de los datos incluye varios grupos la llamada Ley de Potencia de Taylor (Taylor 1961) es otra propuesta para estabilizar varianzas, relacionándo las medias y varianzas de *k* grupos por medio de la ecuación:

$$s_{y_k}^2 = a\left(\overline{y}_k\right)^b \quad (2.126)$$

cuyos parámetros *a* y *b* pudieran ser hallados hasta por regresión no lineal cuando la linealización no de buenos resultados. Una vez encontrado el valor de *b*, las transformaciones para estabilización de la varianza aplicables a los datos serían:

$$y_i' = y_i^{\left(1 - \frac{b}{2}\right)} \text{ para } b \neq 2; \ o \quad y_i' = \ln\left(y_i\right) \text{ para } b = 2 \quad (2.127)$$

2.11 Regresión múltiple

Es otra herramienta fundamental cuando el investigador tiene un control casi seguro de muchas variables fáciles de medir que podrían ayudar a encontrar otra u otras difíciles de hacerlo. La regresión múltiple se debe concebir como una relación estadística entre una variable dependiente con dos o más independientes, cuando se intuye que una sola de ellas no reduce lo suficiente la variabilidad de otra o se quiere ver el efecto logrado por varias variables en otra llamada respuesta, o simplemente por la necesidad de adicionar variables con base en algunas de las medidas remediales del desajuste de residuales o cuando por conocimientos previos o intuitivos un fenómeno puede ser explicado por otras que de pronto se juzgan como independientes, por ejemplo al estudiar el crecimiento de un árbol con base en altura h, d, índice de sitio y edad. Gran parte de la medición forestal se basa en predicciones por lo que algunos modelos con varias variables pueden ser más útiles que los vistos hasta ahora. Se partirá de la explicación de un modelo de primer orden con 2 variables independientes porque permite una conceptualización en tres dimensiones y una extensión ya sin explicación gráfica visible a otros hiperplanos.

2.11.1 Regresión múltiple con dos variables independientes

Para dos variables independientes X_1 y X_2 es posible generar el modelo llamado de primer orden, lineal en los parámetros y en las variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \mathcal{E}_i \quad (2.128)$$

en que: y_i es la respuesta para *i*-ésima observación, x_{ij} valores en la *i*-ésima observación de las variables $j\,(j = 1,\ 2,\ldots,\ p-1)$. Si se asume que $E(\mathcal{E}_i) = 0$, como en el modelo simple, se obtiene la ecuación de un plano respuesta, como el de la Fi 2.26:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (2.129)$$

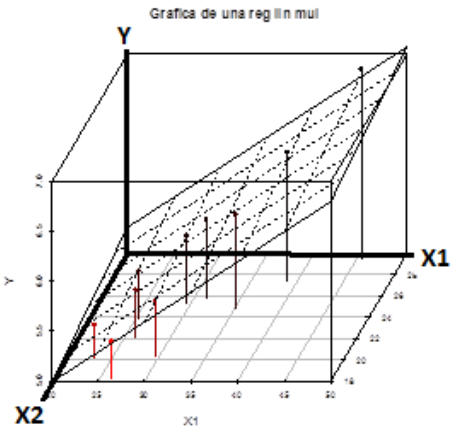


Figura 2.26. Plano respuesta para dos variables independientes

2.11.1.1 Significado de los coeficientes de regresión

β_0 es el intercepto del plano $X_1'X_2'$ con el eje de las Y, y si las observaciones incluyen $X_1 = 0$ y $X_2 = 0$, entonces β_0 adquirirá un valor como respuesta media a ese nivel pues de otra forma no tiene un significado particular. β_1 indicará el cambio en la respuesta

media por cada incremento de una unidad X_1 cuando X_2 se mantenga constante, y caso similar para β_2, llamados coeficientes parciales de regresión.

2.11.1.2 Efecto de las variables

Cuando el efecto de X_1 en la respuesta media no depende del nivel de X_2 y viceversa, las dos variables se dicen aditivas o no interactuantes, siendo esta una de las condiciones de un modelo diseñado con tal propósito, como el mostrado, pues en algunos casos se darán variables interactuantes como en el caso de una $X' = X_1 X_2$. Esto se puede establecer del cálculo diferencial simplemente tomando derivadas parciales a $E(Y)$ con respecto a X_1 y X_2.

$$\frac{\partial E(Y)}{\partial (X_1)} = \beta_1; \quad \frac{\partial E(Y)}{\partial (X_2)} = \beta_2 \quad (2.130)$$

2.11.2 Modelo general de regresión lineal múltiple

Un modelo similar con $p-1$ variables independientes se puede escribir como la ecuación del siguiente hiperplano.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,\,p-1} + \mathcal{E}_i \quad (2.131)$$

A veces se escribe asumiendo una variable $X_0 \cong 1$ como:

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \mathcal{E}_i \quad (2.132)$$

Un modelo con $p-1$ variables independientes, a veces la misma sujeta a alguna transformación, no concebible gráficamente pero que conserve los criterios anotados para el modelo de una y dos variables, tiene la forma expresada por las ecuaciones (2.131) y (2.132). Por ejemplo, β_1 indica el cambio en la respuesta media por cada incremento en la variable X_1 cuando las demás β_{k-1} permanecen constantes. Además en el se conserva que los $\mathcal{E}_i$ son independientes y con tendencia a la distribución normal $N(0,\, \sigma^2)$ y si todas las variables X_{ij} son independientes, no presenta variables interactúantes. Sin embargo, otros se forman simplemente con diversas potencias de una sola variable como el polinomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \quad (2.133)$$

o los modelos de variable combinada, en que alguna variable interactúa con otra, como en el caso de tablas de volumen (capítulo 7),

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \mathcal{E}_i \quad (2.134)$$

cambiando los conceptos relativos a la explicación de los coeficientes de regresión β_1 y β_2 a pesar de lo cual se considerarán como el modelo general. Puede verse que el cambio en la respuesta media por cada incremento en X_1 cuando X_2 es constante es: $\beta_1 + \beta_2 X_{i2}$. Lo mismo para X_2, con lo cual el efecto en una variable depende del nivel asumido por la otra. Para verlo basta con derivar parcialmente:

$$\frac{\partial E(Y)}{\partial X_1} = \beta_1 + \beta_3 X_2 y \frac{\partial E(Y)}{\partial X_2} = \beta_2 + \beta_3 X_1 \quad (2.135)$$

2.11.3 Modelo de regresión lineal en términos matriciales

Es muchísimo más simple el estudio de la regresión con base en matrices.

2.11.3.1 Matrices básicas para la regresión lineal simple

Las matrices básicas para el análisis de regresión lineal simple son entre otras el vector $\mathbf{Y}_{n*1}$, su traspuesta $\mathbf{Y}'_{\mathbf{1*n}}$, así como su producto $\mathbf{Y}'_{\mathbf{1*n}}\mathbf{Y}_{\mathbf{1*n}}$.

$$\mathbf{Y}_{\mathbf{n*1}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix};\ \mathbf{Y}_{\mathbf{1*n}} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix};\ \mathbf{Y'Y} = \begin{bmatrix} \sum y^2 \end{bmatrix} \quad (2.136)$$

La matriz $\mathbf{X}_{\mathbf{n*1}}$ (de valores de las variables independientes) por definición tiene su primer vector de unos (1) y el resto compuesto de los otros valores de X.

$$\mathbf{X}_{\mathbf{n*2}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \mathbf{X}'_{\mathbf{2*n}} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \quad (2.137)$$

En R: modelo1<-lm(alt~dap) con los siguientes datos

	dap	alt
a1	23.2	20.1
a2	45.5	29.4
a3	26.3	22.1
a4	39.1	27.2
a5	31.4	25.7
a6	29.6	25.2
a7	25.4	23.8
a8	26.3	18.1
a9	35.2	24.7
a10	29.6	20.2

X se construye en R como:

```
matriz.modelo1<-model.matrix(modelo1) #creamos la matriz de diseño para un modelo
matriz.modelo1
   (Intercept)  dap
1            1 23.2
2            1 45.5
3            1 26.3
4            1 39.1
...          . ..
9            1 35.2
10           1 29.6
attr(,"assign")
[1] 0 1
```

Igualmente son importantes su traspuesta, así como $\mathbf{X'X}$ y $\mathbf{X'Y}$.

$$\mathbf{X'X}_{\mathbf{2*2}} = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix};\ \mathbf{X'Y}_{\mathbf{2*1}} = \begin{bmatrix} \sum x \\ \sum xy \end{bmatrix} \quad (2.138)$$

La matriz de coeficientes o de coeficientes estimados, $\boldsymbol{\beta}$ ó $\mathbf{b}$, son las compuestas de vectores cuyos componentes son β_0 y β_1 ó b_0 y b_1:

$$\boldsymbol{\beta}_{\mathbf{2*1}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}; \ \mathbf{b}_{\mathbf{2*1}} = \begin{bmatrix} \mathrm{b}_0 \\ \mathrm{b}_1 \end{bmatrix} \quad (2.139)$$

2.11.3.2 Otras matrices importantes

Además de las anteriores se usan para describir y calificar los modelos:

$$\mathbf{E}(\mathbf{Y})_{\mathbf{n*1}} = \begin{bmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_n) \end{bmatrix}; \ \boldsymbol{\mathcal{E}} = \begin{bmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_n \end{bmatrix}; \ \mathbf{E}(\boldsymbol{\mathcal{E}})_{\mathbf{n*1}} = \begin{bmatrix} E(\mathcal{E}_1) \\ E(\mathcal{E}_2) \\ \vdots \\ E(\mathcal{E}_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0} \quad (2.140)$$

y la matriz de varianzas y covarianzas, simétrica expresada por:

$$\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{Y})_{\boldsymbol{n*n}} = \begin{bmatrix} \sigma^2(Y_1) & \sigma^2(Y_1, Y_2) & \dots & \sigma(Y_1, Y_n) \\ \sigma(Y_2, Y_1) & \sigma^2(Y_2) & \dots & \sigma(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(Y_n, Y_1) & \sigma(Y_n, Y_2) & \dots & \sigma^2(Y_n) \end{bmatrix} \quad (2.141)$$

de un vector aleatorio que se compone de las varianzas en la diagonal principal y las covarianzas de las variables en las posiciones (i, j), en la que:

$$\sigma(Y_i, Y_j) = \sigma(Y_j, Y_i); \ \forall\, i \neq j \quad (2.142)$$

para lo cual es útil darse cuenta que es posible escribir $\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{Y})$ como el vector:

$$\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{Y}) = \mathbf{E}\left\{\left[\mathbf{Y} - \mathbf{E}(\mathbf{Y})\right]\left[\mathbf{Y} - \mathbf{E}(\mathbf{Y})\right]'\right\} \quad (2.143),$$

con el cual se obtuvo la matriz anterior.

2.11.3.3 Regresión lineal simple en términos matriciales

El modelo analizado, ecuación (2.1) se puede escribir con base en cada uno de sus vectores $\mathbf{Y}$, $\boldsymbol{\beta}$, $\mathbf{X}$, $\boldsymbol{\mathcal{E}}$, así:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \quad (2.144)$$

Con las asunciones y condiciones vistas de que $E(\mathcal{E}_i) = 0$ y $\sigma^2(\varepsilon_i) = \sigma^2$, se expresan en términos matriciales como:

$$\boldsymbol{\sigma}^{\mathbf{2}}(\boldsymbol{\mathcal{E}}) = \boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \quad (2.145)$$

2.11.3.3.1 Estimación de parámetros. Recordando el numeral 2.3.4.2, las ecuaciones normales (2.17 y (2.18) en términos de matrices se obtienen como:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y} \quad (2.146)$$

de las cuales es posible estimar los coeficientes $(b_0 \text{ y } b_1)$:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \ \therefore \ \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2.147)$$

Por ejemplo: con los anteriores En R:

```
y <- dial[,'alt']#Altura como vector
y
 [1] 20.1 29.4 22.1 etc

b <- solve(t(X) %*% X) %*% t(X) %*% alt#solve(inversa)
> b
          [,1]
[1,] 10.700434
[2,]  0.415583
```

Para obtener las ecuaciones normales se minimiza la cantidad:

$$Q=\sum\left(Y_i-\beta_0-\beta_1X_i\right)^2 \quad (2.148)$$

la cual queda expresada en términos matriciales así:

$$\mathbf{Q}=\left(\mathbf{Y}-\mathbf{X\beta}\right)'\left(\mathbf{Y}-\mathbf{X\beta}\right)=\mathbf{Y'Y}-\mathbf{Y'X\beta}-\boldsymbol{\beta}'\mathbf{X'Y}+\boldsymbol{\beta}'\mathbf{X'X\beta} \quad (2.149)$$

la cual se reduce a

$$\mathbf{Q}=\mathbf{Y'Y}-2\boldsymbol{\beta}'\mathbf{X'Y}+\boldsymbol{\beta}'\mathbf{X'X\beta} \quad (2.150)$$

Para encontrar los valores de $\boldsymbol{\beta}$ que minimizan $\mathbf{Q}$, se diferencia con respecto a $\boldsymbol{\beta}_0$ y $\boldsymbol{\beta}_1$ que, al igualarlas a $\mathbf{0}$ dan los valores de $\mathbf{b}$:

$$\frac{\partial\mathbf{Q}}{\partial\boldsymbol{\beta}}=\begin{bmatrix}\dfrac{\partial Q}{\partial\beta_0}\\ \dfrac{\partial Q}{\partial\beta_1}\end{bmatrix}=-2\mathbf{X'Y}+2\mathbf{X'X\beta}=0\ \therefore\ \mathbf{X'Xb}=\mathbf{X'Y} \quad (2.151)$$

2.11.3.3.2 Análisis de varianza en términos matriciales. Sean los vectores de

$$\hat{\mathbf{Y}}=\begin{bmatrix}\hat{Y}_1\\ \hat{Y}_2\\ \vdots\\ \hat{Y}_n\end{bmatrix}\ \text{y}\ \mathbf{e}=\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{bmatrix}\rightarrow\hat{\mathbf{Y}}=\mathbf{Xb}\rightarrow\mathbf{e}=\mathbf{Y}-\hat{\mathbf{Y}}=\mathbf{Y}-\mathbf{Xb} \quad (2.152)$$

Con base en ellos se pueden encontrar las distintas sumas de cuadrados corregidos, así, la de totales:

$$\mathbf{SSTO}=\mathbf{Y'Y}-\left(\frac{1}{n}\right)\mathbf{Y'11'Y} \quad (2.153)$$

la de los residuales:

$$\boldsymbol{SSE}=\mathbf{e'e}=\begin{cases}\left(\mathbf{Y}-\mathbf{Xb}\right)'\left(\mathbf{Y}-\mathbf{Xb}\right)\\ \mathbf{Y'Y}-\mathbf{Y'Xb}-\mathbf{b'X'Y}+\mathbf{b'X'Xb}\end{cases} \quad (2.154)$$

se observa que:

$$\left(\mathbf{Y'Xb}\right)'=\mathbf{b'X'Y}=\mathbf{b'X'Xb}\rightarrow\boldsymbol{SSE}=\mathbf{Y'Y}-\mathbf{b'X'Y} \quad (2.155)$$

y que la suma de cuadrados de la regresión se obtiene por definición o por diferencias como:

$$\boldsymbol{SSR}=\mathbf{b'X'Y}-\left(\frac{1}{n}\right)\mathbf{Y'11'Y} \quad (2.156)$$

2.11.3.3.4 Matriz de varianzas y covarianzas de los coeficientes. La matriz de varianzas y covarianzas de $\mathbf{b}$ es:

$$\boldsymbol{\sigma^2}(\mathbf{b}) = \begin{bmatrix} \sigma^2(b_0) & \sigma(b_0 b_1) \\ \sigma(b_1 b_0) & \sigma^2(b_1) \end{bmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (2.157)$$

La cual es estimada por $\mathbf{s}^2(\mathbf{b})$ así:

$$s^2(\mathbf{b}) = MSE(\mathbf{XX})^{-1} \quad (2.158)$$

En R encontramos primero $\boldsymbol{\sigma^2}$

```
sigma2<- t(y - X%*%b) %*% (y - X%*%b) /8
sigma2
         [,1]
[1,] 4.52504
sigma<-sigma2^.5
sigma
          [,1]
[1,] 2.127214
```

Y (2.158) así:

```
sigma2n<-as.numeric(sigma2)
 sig2b<- sigma2n*solve(t(X) %*% X)
sig2b
            [,1]        [,2]
[1,] 10.5828422 -0.32510713
[2,] -0.3251071  0.01043348
```

2.11.4 Regresión lineal múltiple en términos matriciales.Tiene un tratamiento similar a la regresión lineal simple. pero cambiando las dimensiones de las matrices:

$$\mathbf{Y}_{\boldsymbol{n*1}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{X}_{\boldsymbol{n*p}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,\,p-1} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,\,p-1} \end{bmatrix}; \quad \boldsymbol{\beta}_{\boldsymbol{p*1}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}; \quad \boldsymbol{\varepsilon}_{\boldsymbol{n*1}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (2.159)$$

x_{ik} identifica la variable *k* de la observación *i*. De acuerdo con lo anterior:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon};\ E(\boldsymbol{\varepsilon}) = 0;\ \sigma^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} = \sigma^2;\ E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta};\ \sigma^2(\mathbf{Y}) = \sigma^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} \quad (2.160)$$

2.11.4.1 ANAVA para la regresión múltiple. Los estimadores mínimos cuadráticos y el análisis de varianza se rigen por los mismos conceptos dados en el numeral 2.7.5, variando únicamente los grados de libertad que quedan como se muestra en el anava de la Tab 2.8. Resaltamos que lo presentado para la regresión lineal es lo usado en este ítem, pues con las pruebas usadas para la simple bastaba para la mayoría de sus diagnósticos.

2.11.4.2 Tabla 2.8. ANAVA *para la regresión múltiple*

. Fuente de variación	*gl*.	Sumas de cuadrados	Cuadrados medios
Regresión	$p-1$	$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}$	$SSR/(p-1)$
Error	$n-p$	$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$SSE/(n-p)$
Total	$n-1$	$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}$	

2.11.4.3 Prueba de *F* para la relación de regresión. Para verificar si se da la regresión entre la variable dependiente y las independientes $X_1, X_2, \ldots, X_{p-1}$ es preciso escoger entre las alternativas:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_a : \text{al menos un } \beta_k \ (k = 1,\ 2, \ldots,\ p-1) \neq 0 \end{cases}$$

Para ello entonces se plantea:

$$\begin{cases} F_{calc} = \dfrac{MSR}{MSE}; \ \ si \ F_{calc} > F_{(1-\alpha,\ p-1,\ n-p)} \rightarrow H_a \end{cases} \quad (2.161)$$

Es de anotar que la existencia de una relación de estas no asegura que las predicciones basadas en ella resulten útiles.

2.11.4.4 Coeficiente de determinación múltiple.

Se define como ya se había visto en (2.90), y mide la reducción proporcionada de la variación total de la variable dependiente, asociada con el uso del conjunto de variables independientes $X_1, \ldots, X_{p-1}$. El $0 \leq R^2 \leq 1$; 0 cuando todos los $\beta_k = 0 \ (k = 1,\ 2, \ldots,\ p-1)$ y 1 si todas las observaciones caen en el hiperplano ajustado. Al adicionar muchas variables independientes al modelo se incrementa automáticamente su valor, por lo cual se sugiere el $R^2_{ajusta_{d_i}} = R^2_a$ que reconoce el número de variables independientes del modelo así:

$$R^2_a = 1 - \frac{(n-1)}{(n-p)} \frac{SSE}{SSTO} \quad (2.162)$$

2.11.4.5 Inferencias acerca de los parámetros de la regresión

Los estimadores mínimos cuadráticos en b son insesgados o sea $E(b) = \beta$. La matriz de varianzas y covarianzas similar a la propuesta:

$$\boldsymbol{\sigma^2}(\mathbf{b}) = \begin{bmatrix} \sigma^2(b_0) & \sigma(b_0,\ b_1) & \ldots & \sigma(b_0,\ b_{p-1}) \\ \sigma(b_1,\ b_0) & \sigma^2(b_1) & \ldots & \sigma(b_1,\ b_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(b_{p-1},\ b_0) & \sigma(b_{p-1},\ b_1) & \ldots & \sigma^2(b_{p-1}) \end{bmatrix} \begin{cases} \boldsymbol{\sigma}^2(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ \text{estimada por} \\ \mathbf{s}^2(\mathbf{b}) = MSE(\mathbf{X}'\mathbf{X})^{-1} \end{cases} \quad (2.163)$$

2.11.4.6 Estimación de los intervalos para β_k Como es de suponer, la distribución de:

$$\frac{(b_k - \beta_k)}{s(b_k)} \rightarrow t_{(1-\alpha/2,\ n-p)} \quad (2.164);$$

y sus límites de confianza para el nivel $1-\alpha$ serán:

$$\beta_k = b_k \pm t_{(1-\alpha/2,\ n-p)} * s(b_k) \quad (2.165)$$

2.11.4.7 Pruebas para β_k **.**Se asumen en la forma usual así:

$$\begin{cases} H_0 : \beta_k = 0 \\ H_a : \beta_k \neq 0 \end{cases} \quad (2.166)$$

con el cálculo de t^* y sus respectivas reglas de decisión:

$$t^* = \frac{b_k - 0}{s(b_k)} = \frac{b_k}{s(b_k)} \therefore \begin{cases} si\,|t^*| \le t_{(1-\alpha/2,\ n-p)} \rightarrow H_0 \\ si\,|t^*| > t_{(1-\alpha/2,\ n-p)} \rightarrow H_a \end{cases} \quad (2.167)$$

2.11.4.7 Inferencias acerca de la respuesta media. Para valores dados de las variables independientes agrupadas como el vector:

$$\mathbf{X_h} = \begin{bmatrix} 1 \\ x_{h1} \\ x_{h1} \\ \vdots \\ x_{h(p-1)} \end{bmatrix} \quad (2.168)$$

Para valores dados de las variables independientes agrupadas como el vector $\mathbf{X_h}$ la respuesta media denotada por $\mathbf{E(Y_h)}$ será:

$$\mathbf{E(Y_h) = X'_h \beta} \quad (2.169),$$

y su respuesta media estimada como:

$$\mathbf{\hat{Y}_h = X'_h b} \quad (2.170)$$

Este estimador resulta insesgado:

$$E\left(\mathbf{\hat{Y}_h}\right) = \mathbf{X'_h \beta} = \mathbf{E(Y_h)};\ \text{y su varianza es}\ \boldsymbol{\sigma}^2\left(\mathbf{\hat{Y}_h}\right) = \boldsymbol{\sigma}^2 \mathbf{X'_h}\left(\mathbf{X'X}\right)^{-1}\mathbf{X_h} = \mathbf{X'_h}\boldsymbol{\sigma}^2(\mathbf{b})\mathbf{X_h} \quad (2.171)$$

con su varianza estimada dada por:

$$s^2\left(\hat{Y}_h\right) = \begin{cases} MSE\left(\mathbf{X'_h}\left(\mathbf{X'X}\right)^{-1}\mathbf{X_h}\right); \\ \text{ó} \qquad (2.172) \\ \mathbf{X'_h s}^2(\mathbf{b})\mathbf{X_h} \end{cases}$$

2.11.4.8 Predicción de una nueva observación. Los límites de predicción para una nueva observación correspondiente a un $\mathbf{X}_h$ se darán junto con su varianza por:

$$\begin{aligned} Y_{h(nueva)} &= \hat{Y}_h \pm t_{(1-\alpha/2;\ n-p)} * s\left(\bar{Y}_{h(nueva)}\right) \\ \mathbf{s}^2\left(\bar{Y}_{h(nueva)}\right) &= MSE + \mathbf{X'_h s}^2(\mathbf{b})\mathbf{X_h} = MSE\left(1 + \mathbf{X'_h}\left(\mathbf{X'X}\right)^{-1}\mathbf{X_h}\right) \end{aligned} \quad (2.173)$$

2.11.4.9 Predicción de *m* nuevas observacionesa. Para predecir un conjunto de m nuevas observaciones para un $\mathbf{X}_h$ dado, su media $\bar{\mathbf{Y}}_{h(nueva)}$, así como su varianza se darán por:

$$\left.\begin{aligned} Y_{h(nueva)} &= \hat{Y}_h \pm t_{(1-\alpha/2;\ n-p)} * s\left(Y_{h(nueva)}\right) \\ s^2\left(Y_{h(nueva)}\right) &= \frac{MSE}{m} + \mathbf{X_h s}^2(\mathbf{b})\mathbf{X_h} = MSE\left(\frac{1}{m} + \mathbf{X_h}\left(\mathbf{X'X}\right)^{-1}\mathbf{X_h}\right) \end{aligned}\right\} \quad (2.174)$$

2.11.4.10 Observaciones remotas — Distancia de Cook D_i

Se utiliza para medir el impacto de la *i*-ésima observación en el cálculo de los coeficientes de regresión y por ende establecer su carácter influencial o no. Su expresión está dada por:

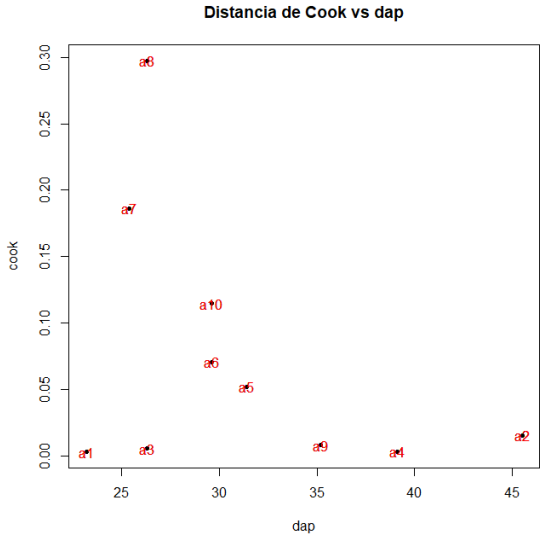
$$\frac{(\mathbf{b}-\boldsymbol{\beta})' \mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})}{p * MSE} = F_{(1-\alpha;\, p,\, n-p)} \rightarrow D_i = \frac{\left(\hat{Y}-\hat{Y}_{(i)}\right)' \left(\hat{Y}-\hat{Y}_{(i)}\right)}{p * MSE} \quad (2.175)$$

la segunda de ellas resume bien el concepto, en la cual: D_i = distancia de Cook, $\hat{\mathbf{Y}}$ = vector de valores ajustados de Y con todos los datos, $\hat{\mathbf{Y}}_{(\mathbf{i})}$ = vector de los valores ajustados de Y con el modelo sin la *i*-ésima observación, p = número de variables en el modelo. Para evaluar D_i también se acude a cambiar en la primera expresión el valor de β por $\mathbf{b}_{(i)}$ coeficiente obtenido sin la *i*-ésima observación que aunque no se distribuye exactamente como una F se asume como tal. En resumen, se debe hacer referencia a la correspondiente distribución de F definida por los grados de libertad $(p,\ n-p)$; así $F\ (p,\ n-p)$. Se debe determinar el valor percentil de tal distribución para tomar la decisión pertinente. Si el valor percentil es menor que el 20% (algunos autores aceptan el 10%), la *i-ésima* observación tiene poca influencia, pero si es mayor del 50% o más, es contundente su efecto. En R:

```
modelo1<-lm(alt~dap)
cook<-cooks.distance(modelo1)
cook
          1           2           3           4           5           6
0.002800914 0.015347167 0.005267428 0.002981919 0.051969934 0.070498919
          7           8           9          10
0.186084619 0.297516459 0.008089571 0.114510455
```

2.11.14.11. Medidas influenciales. Cook i-ésimo calcula la distancia entre los parámetros estimados para ver si se incluye o no la iesima observación, para cada uno de los datos, considerándose significativa si es mayor que 1, calculada con R, con la función mostrada (cooks.distance). Según lo anterior no hay ninguna en los datos analizados. Podemos incluso graficara el mode1<-lm(alt~dap)

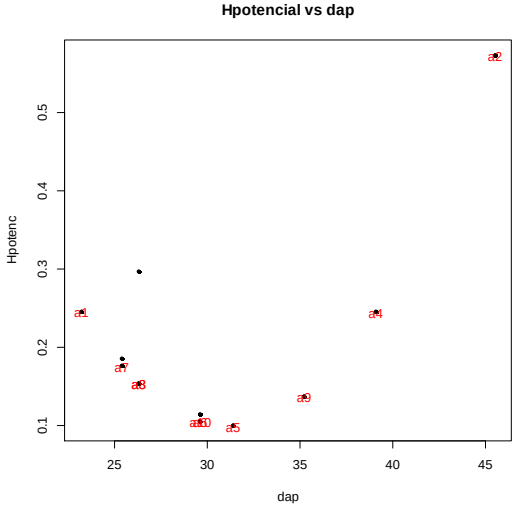
```
cook<-cooks.distance(mode1)
plot(dap,cook,pch=20,main=
"Distancia de Cook vs dap")
points(dap, cook,pch=20)
text(cook~dap,data=dialt,
label=ide,col="red")
```



2.11.14.12. Matriz hat. La matriz *hat* propicia la medida de unos vectores palanca (*leverages*), útiles para cuantificar la influencia de algunas observaciones en la regresión. Esta se conoce también como matriz de proyección al proyectar las observaciones (y) en el vector de predicciones, quizás por ello su nombre. Se define como: $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, pues permite encontrar los $\hat{y} = \boldsymbol{H}y = \boldsymbol{X}b$. Los elementos diagonales de H, son los llamados vectores palanca o leverages que siempre darán $0 \le h_{ii} \le 1$, y $\sum_{i=1}^{n} h_{ii} = p$

2.11.14.13. Otras observaciones potencialmente influyentes. Se analiza para cada caso, el peso de la i-ésima observación al estimar la predicción. Los pesos o potenciales se obtienen matricialmente como la diagonal principal de la matriz HAT. En R. Los elementos de la diagonal principal de hat son unos potenciales de diagnóstico para la regresión

```
madimodelo1<-model.matrix(modelo1) #creamos la matriz de diseño
madimodelo1
   (Intercept)  dap    Matriz Hat en Excel
1            1 23.2
2            1 45.5
3            1 26.3
4            1 39.1
5            1 31.4
6            1 29.6
7            1 25.4
8            1 26.3
9            1 35.2
10           1 29.6
attr(,"assign")
[1] 0 1
Hpotenc<-hat(madimodelo1)
Hpotenc
 [1] 0.2460941 0.5741381 0.1544602 0.2453609 0.1001328 0.1056112 0.1764983 0.1544602 0.1376330
0.1056112
plot(dap,Hpotenc,pch=20,main="Hpotencial vs dap")
points(dap, Hpotenc,pch=20)
text(Hpotenc~dap,data=dialt,label=ide,col="red")
```




```
influence.measures(modelo1)
Influence measures of
         lm(formula = alt ~ dap) :
    dfb.1_ dfb.dap   dffit cov.r  cook.d   hat inf
1  -0.0621  0.0540 -0.0701 1.725 0.00280 0.246
2   0.1318 -0.1491 -0.1641 3.050 0.01535 0.574   *
3   0.0720 -0.0572  0.0964 1.523 0.00527 0.154
```

```
4  -0.0449  0.0557  0.0723 1.723 0.00298 0.245
5   0.0549  0.0117  0.3209 1.132 0.05197 0.100
6   0.1625 -0.0878  0.3808 1.057 0.07050 0.106
7   0.5158 -0.4246  0.6449 0.972 0.18608 0.176
8  -0.7004  0.5565 -0.9371 0.543 0.29752 0.154
9   0.0402 -0.0626 -0.1197 1.476 0.00809 0.138
10 -0.2195  0.1186 -0.5143 0.838 0.11451 0.106
Hpr=4/10
> Hpr
[1] 0.4
which(apply(inflm.SR$is.inf, 1, any)); 2;2
```

2.12 . Mínimos cuadrados ponderados en forma matricial

Diferentes ponderadores pueden asumirse para las observaciones como se hizo anteriormente, basta con colocar los pesos $\mathbf{W}_i$ en una matriz diagonal $\mathbf{W}$ de pesos, con lo cual se calcula $\mathbf{b}=\left(\mathbf{X'WX}\right)^{-1}*\mathbf{X'WY}$. Cuando los términos del error σ_i^2 son desiguales se recomienda tomar pesos inversamente proporcionales a ellos ejemplo $w_i=\frac{1}{\sigma_i^2}$, en que σ_i^2 es la varianza del término del error para la *i-ésima* observación.

$$\mathbf{W}_{\boldsymbol{n*m}}=\begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix} \quad (2.176)$$

De ahí:

$$\mathbf{s}^2\left(\mathbf{b}\right)_{\boldsymbol{p*p}}=MSE\left(\mathbf{X'WX}\right)^{-1} \quad (2.177)$$

Los w_i pueden obtenerse por diversos criterios, por ejemplo, si los términos del error resultan proporcionales por niveles de la variable independiente o a alguno de ellos; x_{ik}^2, entonces $\sigma_i^2=kx_i^2=y$; $w_i=\frac{1}{kx_{ik}^2}$, o algunas veces $w_i=\frac{1}{x_{ik}^2}$. En el Capítulo 7 se amplía el concepto con un ejemplo complementario.

2.13 Pruebas de contribución de variables

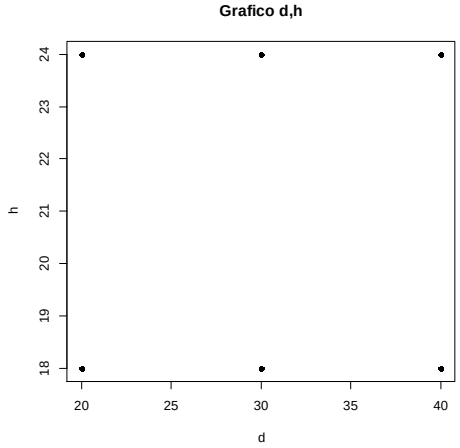
Es a veces interesante conocer el efecto de algunas variables o relación entre algunas independientes y la dependiente, o entre las independientes. Se plantean preguntas como estas:
¿Cuál es el efecto relativo de las diferentes variables independientes?
¿Cuál es la magnitud del efecto de una variable dada en la variable dependiente?
¿Puede eliminarse alguna variable del modelo si su aporte no se considera importante?
¿Podrán considerarse variables aún no incluidas en un modelo apto para ello?
Cuando las variables independientes incluida en el modelo no están correlacionadas entre si, o con cualquier otra variable omitida del modelo, estas preguntas parecen tener respuestas simples, pero cuando esto no ocurre aparece el efecto de multicolinealidad que debe ser removido del modelo.

2.13.1 Efecto de variables independientes no correlacionadas. Para mostrarlo se sacaron datos de un inventario forestal, Tabla 2.9 para establecer una correlación entre las variables diámetro (d) en cm, altura (h) en m y volumen comercial de aserrío (V) en pies tablares. *Exprofeso* se rebuscó una muestra como la presentada en la Tabla 2.

Tabla 2.9. Datos para volumen (V) con base en diámetro (d) y altura (h).

d	*h*	*V*	*d*	*h*	*V*
20	18	101	30	24	151
20	18	105	30	24	149
20	24	118	40	18	216
20	24	127	40	18	202
30	18	142	40	24	225
30	18	138	40	24	238

de la cual se calcula la relación entre d y h: $R^2_{dh} = 0$. La grafica muestra dos variables independientes diámetro altura, como se muestra en la figura



o con los valores covarianza

```
var(d,h)
[1] 0
O correlacion
cor(d,h)
[1] 0
```

A las variables que se hagan presentes en cualquier expresión se las notará entre paréntesis con sus respectivos subíndices, por ejemplo, la notación $SS^*(X_i,\ X_j)$ quiere decir que es una suma de cuadrados * con las variables anotadas para el modelo. Se ajustaron los tres modelos presentados a continuación:

$$1)\ V = -62.583 + 5.375d + 2.889h;$$

$$2)\ V = -1.1917 + 5.375d;$$

$$3)\ V = 98.666 + 2.889h;$$

y se presenta sus análisis de varianza Tab 2.10.

Tabla 2.10. Análisis de varianza para tres modelos con variables independientes no correlacionadas.

Fuente	Sumas de cuadrados		*gl.*	Cuadrados medios	Razón de *F*	*P-value*
Modelo 1	*SSR* (*d, h*)	24013,80	2	12006,90	69,32	0,0000
Residual	*SSE* (*d, h*)	1558,83	9	173,21		
Modelo 2	*SSR* (*d*)	23112,50	1	23112,50	93,95	0,0000
Residual	*SSE* (*d*)	2460,17	10	246,02		
Modelo 3	*SSR* (*h*)	901,33	1	901,33	0,37	0,5590
Residual	*SSE* (*h*)	24671,30	10	2467,13		

Total (Corr)	25572,70	11

Nótese que el b_1 es el mismo para los modelos 1 y 2, el b_2 en 1) es igual a b_1 en 3), sin importar las otras variables incluidas o no en el modelo, cuando las variables independientes no están correlacionadas. También se nota que la $SSE(d,\ h)=1.558,83$, pero cuando solo se incluye el d , $SSE(d)=2.460,17$, por lo cual se puede adjudicar la diferencia $SSE(d)-SSE(d,\ h)=\ \ 2.460,17-1.558,83=901,34=SSR(h)$ al efecto provocado por h . Esta diferencia se denotará por $SSR(h/d)=SSE(d)-SSE(d,\ h)$. Simultáneamente

$SSE(h)-SSE(d,\ h)=24.671,3-1.558,83=23.112,50=SSR(d)$, de donde:

$$1.\ SSR(h)=SSE(d)-SSE(d,\ h);$$
$$2.\ SSR(h|d)=SSE(d)-SSE(d,\ h);\quad (2.178)$$
$$3.\ SSR(d|h)=SSE(h)-SSE(d,\ h)$$

Se concluye que si las dos variables d y h no están correlacionadas $R^2_{dh}=0$ y la contribución de una de estas variables, para reducir la *SSE**, cuando la otra variable independiente está en el modelo es exactamente la misma que cuando esta variable independiente está sola en él, o sea que:

$$SSR(d,\ h)=SSR(d)+SSR(h)\quad (2.179)$$

2.13.1 Efecto de variables independientes correlacionadas

Es un uso práctico en inventarios forestales hacer estimativos de altura con base en los diámetros a la altura del pecho, dada la dificultad física de medirla en el campo, por lo cual las variables independientes d y h pueden presentar correlación. A la manera de lo hecho en el caso anterior se tomó una serie aleatoria de datos sin ningún tipo de arreglo en parcelas a las cuales se les estimo su altura por medio de una regresión diámetro altura ($R^2_{dh}=0,915$) según lo mostrado por la Tabla 2.11. Con base en estos datos se ajustaron tres modelos similares a los ajustados en el primer caso:

Tabla 2.11. Datos de diámetro (d), altura (h) y edad (ed) correlacionados para estimar volumen (V).

d	*h*	*ed*	*V*	*d*	*h*	*ed*	*V*
20	15,9	10	98	35	23,9	22	232
23	18,5	14	112	32	22,0	23	143
25	22,0	13	126	21	16,4	13	99
27	21,1	16	138	37	24,9	24	228
29	21,1	19	142	38	24,8	23	246
40	26,7	24	239	39	25,8	25	242

$$1.\ V=-67.0063+8.569d-1.0915h;$$
$$2.\ V=-75.382+8.0599d;$$
$$3.\ V=-175.75+15.79h$$

A los tres modelos mencionados se les calculó el Anava que se expone en la Tabla 2.12.

Tabla 2.12. Análisis de varianza para tres modelos lineales con variables independientes correlacionadas.

Fuente	Sumas de cuadrados		*gl.*	Cuadrados medios	Razón de *F*	*P-value*
Modelo 1.	*SSR* (*d. h*)	38007.20	2	19003.60	57.36	0.0000

Residual	*SSE* (*d. h*)	2981.73	9	331.30		
Modelo 2.	*SSR* (*d*)	37994.00	1	37994.00	126.86	0.0000
Residual	*SSE* (*d*)	2994.88	10	299.49		
Modelo 3.	*SSR* (*h*)	34591.00	1	34591.00	57.07	0.0000
Residual	*SSE* (*h*)	6397.92	10	639.79		
Total (Corr)		40988.90	11			

Se nota en primer lugar que los coeficientes que afectan al d y h no son los mismos, porque ya el efecto de d depende de la presencia o ausencia de h y viceversa, a causa de su correlación alta, por lo cual sus coeficientes ya no reflejan los efectos inherentes de las variables individuales sobre la variable dependiente, sino únicamente un efecto parcial o marginal que depende de la presencia o ausencia de otras variables.

2.13.2 Efectos de la multicolinealidad en las sumas de cuadrados

Cuando se da alguna correlación entre las variables independientes, no existe una suma de cuadrados única adscrita a ellas que refleje su efecto en la reducción de la variación total en la variable dependiente, por lo cual esta variación se debe contextualizar con respecto a las otras variables. Usando la notación tradicional para las variables independientes:

$SSR(X_k)$ mide la reducción de la variación en Y cuando solo X_k está en el modelo.

$SSR(X_k|X_j)$ mide la reducción adicional en la variación de Y cuando se introduce X_k dado que X_j ya está en el modelo.

En el ejemplo reseñado: $SSR(d|h) = 6.397,92 - 2.981,73 = 3.416,19$, $SSR\ (h|d) = 2994.88 - 2981.73 = 13.50$. La razón por la cual $SSR\ (d|h) < SSR(d)$, se explica porque mucha de la variación ya fue reportada por la presencia de h. Por ello, el efecto marginal de d para reducir la variación en V dado que h está ya en el modelo, es menor que si se introdujera sola. Por similar razón $SSR(h|d) < SSR(h)$. Los términos $SSR(X_k|X_j)$ se llaman sumas extras de cuadrados. en resumen:

$$SSR(d \mid h) = SSE(h) - SSE(d;\ h);$$
$$SSR(h \mid d) = SSE(d) - SSE(d;\ h) \quad (2.180)$$
$$SSR(d \mid h) \neq SSR(h \mid d)$$

2.13.3 Paradoja estadística

Surge una aparente paradoja estadística al evaluar modelos con altísima correlación entre variables independientes, pues de un lado la prueba de F puede rechazar la $H_0 : \beta_k = 0$ contra la H_a: al menos un $\beta_k \neq 0$, cuando por otro lado las pruebas separadas de t (t pequeños) muestran lo contrario, debido a que en cada coeficiente la prueba de t lo que está midiendo es la contribución marginal de las variables independientes dado que otras están en el modelo. Es un caso parecido al del último ejemplo en el cual $t(b_2) = -0,246$ cuando si se deja sola en el modelo alcanza un valor de 7.

2.13.4 Descomposición de SSR en sumas extras de cuadrados

Ya que una suma extra de cuadrados refleja la reducción en la suma de cuadrados del error al adicionar una variable independiente dado que otras están en el modelo,

entonces ello implica un incremento igual en las sumas de cuadrados de la regresión ya que $SSTO = SSE + SSR$. Ejemplo:

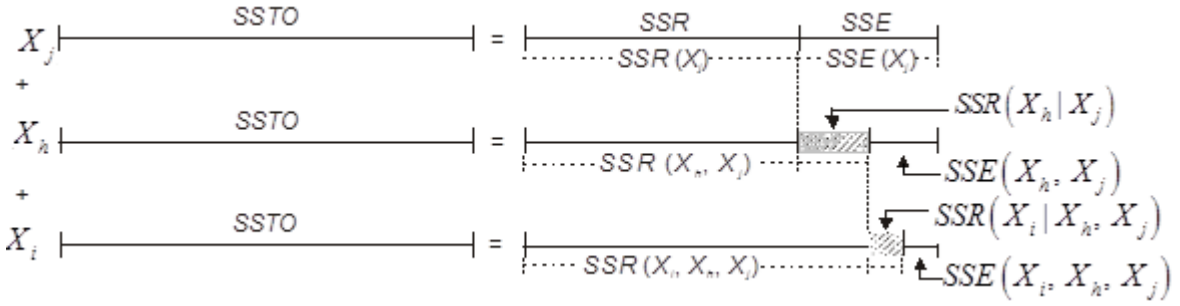


Figura 2.27. Descomposición de una SSR en sumas extras de cuadrados.

Por eso se puede escribir:

$$SSR(X_h \mid X_j) = SSR\ (X_h,\ X_j) - SSR(X_j) \quad (2.181)$$

pero también:

$$SSR\ (X_h \mid X_j) = SSE(X_j) - SSE(X_h,\ X_j) \quad (2.182)$$

Una extensión a tres o más variables conduce a que:

$$SSR(X_i \mid X_h,\ X_j) = SSE(X_h,\ X_j) - SSE(X_i,\ X_h,\ X_j) \quad (2.183)$$

y en forma más amplia, se puede extender a más variables, entonces:

$$SSR(X_i \mid X_h,\ X_j) = SSR(X_i,\ X_h,\ X_j) - SSR(X_h,\ X_j) = SSE(X_h,\ X_j) - SSE(X_i,\ X_h,\ X_j) \quad (2.184)$$

Cada suma de cuadrados que involucre la adición de una variable independiente al modelo de regresión tiene asociado con ella un grado de libertad, y dado que $SSTO = SSR(X_i) + SSE(X_i)$, es posible al reemplazar por ejemplo $SSE(X_i)$ de una expresión como $SSR\ (X_2 \mid X_1) = SSE(X_1) - SSE(X_1,\ X_2)$, entonces:

$$SSE(X_1) = SSR(X_2 \mid X_1) + SSE(X_1,\ X_2) \therefore SSTO = SSR(X_1) + SSR(X_2 \mid X_1) + SSE(X_1,\ X_2) \quad (2.185)$$

pero

$$SSE(X_1,\ X_2) = SSR(X_3 \mid X_1,\ X_1) + SSE(X_1,\ X_2,\ X_3) \quad (2.186)$$

$$SSTO = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1,\ X_2) + SSE(X_1,\ X_2,\ X_3) \ldots \quad (2.187)$$

y así podrían seguirse extendiendo otras descomposiciones de acuerdo con el orden de entrada de las variables al modelo, con el fin de llegar a verificar las hipótesis de las variables individuales por medio de coeficientes de determinación parcial o las pruebas de *F* que se plantean enseguida.

2.13.5 Coeficiente de correlación muestral. Miden la contribución marginal de una variable independiente para reducir la variabilidad de la variable dependiente Y cuando ya otras variables están en el modelo. Por ejemplo en el modelo analizado en el numeral 2.13.2: $SSE(X_2)$ mide la variación de Y cuando X_2 es incluida en el modelo; $SSE(X_1,\ X_2)$ mide la variación de Y cuando las incluidas son X_1, X_2. Pero

la reducción marginal relativa o adicional relativa de la variación en Y asociada con X_1, cuando X_2 ya está en el modelo se obtiene como:

$$\frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)}, \quad (2.188)$$

en idéntica forma:

$$\frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} \quad (2.189)$$

mide el efecto marginal de X_2, cuando X_1, ya estaba en el modelo. Pero ya se sabe que: $SSR(X_2 \mid X_1) = SSE(X_1) - SSE(X_1, X_2)$; entonces:

$$r^2_{Y2.1} = \frac{SSR(X_2 \mid X_1)}{SSE(X_1)} \quad (2.190)$$

mide la proporción de la reducción en la variación de Y cuando se involucra X_2, dado que X_1, ya estaba en el modelo; y

$$r^2_{Y1.2} = \frac{SSR(X_1 \mid X_2)}{SSE(X_2)} \quad (2.191)$$

mide la proporción de la reducción en la variación de Y cuando se involucra X_1, dado que X_2, ya estaba en el modelo, por lo cual coinciden con coeficientes de determinación parcial y adoptan este nombre. Para nuestro ejemplo 2.13.2 se tiene:

$$r^2_{Y1.2} = (6.397,92 - 2.981,73) / 6.397,92 = 0,53; r = 0,73;$$

$$r^2_{Y2.1} = (2.994,88 - 2.981,73) / 2.994,88 = 0,0044$$

lo que muestra que al aporte adicional en la reducción de la variación del volumen al incluir el diámetro es alto comparado con la inclusión de la altura cuando ya está incluido el d. Este es un resultado lógico hasta en términos prácticos ya que es mejor predictora la variable d para el V. Estos coeficientes se utilizan discrecionalmente por parte de los usuarios para tomar decisiones acerca de la cantidad proporcional de ajuste que se lograría con una de las inclusiones, respetando los órdenes de colocación de las variables originalmente. En el sector forestal variables que aporten más del 5% se pueden considerar importantes en casos como los de estimación de volúmenes, por ejemplo. Como caso general se podría ver, aunque existen múltiples combinaciones que:

$$r^2_{Y_i, j_k} = \frac{SSR(X_i \mid X_j, X_k)}{SSE(X_j, X_k)} \quad (2.192)$$

2.13.6 Pruebas para verificar si una $\beta_k = 0$ contra $\beta_k \neq 0$

Se acude a la prueba parcial de *F*, con las hipótesis ya conocidas, cuyas alternativas son:

$$H_0 : \beta_k = 0 \text{ vs } H_a : \beta_k \neq 0 \quad (2.193)$$

que se evalúa con la prueba de:

$$F^* = \frac{\dfrac{SSR(X_k \mid X_1, \ldots, X_{k-1};\ X_{k+1}, \ldots,\ X_{p-1})}{1}}{\dfrac{SSE(X_1, \ldots,\ X_{p-1})}{n - p}} = \frac{MSR(X_k \mid X_1, \ldots,\ X_{k-1};\ X_{k+1}, \ldots,\ X_{p-1})}{MSE} \quad (2.194)$$

que se da bajo H_0. Grandes valores de *F* concluyen H_a. Como ejemplo adicional para este tema, se ajustó un nuevo modelo para estimar V con base en la Tabla 2.13, incluyendo como variable independiente la edad, que muestra la influencia de cada variable adicionada cuando la otra u otras ya están en el modelo, lo que posibilita su simplificación. Además de los resultados mostrados otros valores se pueden consultar en las dos tablas anteriores. Ante valores significativos del $r^2_{X_{ij}}$, de acuerdo con las tablas de significación de *r*; para ciertos grados de libertad y niveles de α por ejemplo; es posible eliminar variables que resulten altamente asociadas entre las independientes (Steel & Torrie 1985).

Tabla 2.13. Modelo de regresión con 3 variables correlacionadas.

Parámetros	Estimados	Desviación estándar	*t*	*P-value*
Constante	-49.1851	37.1213	-1.32498	0.2218
d	17.4849	3.8086	4.59089	0.0018
h	-6.5788	4.6065	-1.42816	0.1911
ed	-8.9973	3.2555	-2.76375	0.0245

Anava

Fuente	Suma de cuadrados	*gl*.	*MSS*	Razón de *F*	*P-value*
Modelo	39463.6	1	13154.50	68.99	0.0000
Residual	1525.35	10	190.67		
Total (Corr)	40988.9	11			

$R^2 = 0,962786 \rightarrow R^2_{ajustado} = 0,948831$; $MSE = 138033$; $DW = 3,00145$

Anava para las variables en orden de ajuste

Fuente	Suma de cuadrados	*gl*.	*MSS*	Razón de *F*	*P-value*
d	37994.00	1	37994.00	199.27	0.0000
d|*h*	13.15	1	13.15	0.07	0.7995
ed|*d*. *h*	1456.38	1	1456.38	7.64	0.0245
Modelo	39463.60	3			

2.14 Otros elementos de multicolinealidad.

Este término a veces se deja sólo para efectos extremos de correlación. La multicolinealidad ocurre por ejemplo si en la matriz de variables independientes las variables X_i y X_j miden casi el mismo efecto, o sea que las columnas son casi linealmente dependientes o también cuando otra variable X_k es combinación lineal de las anteriores. En caso de multicolinealidad el determinante de $(\mathbf{X}'\mathbf{X})^{-1}$ será cercano a cero (Wonnacott & Wonnacott 1981). En general, el efecto más importante de la multicolinealidad es la contribución de las variables a la *SSE*. Cuando las variables independientes están incorrelacionadas la contribución marginal de una variable en presencia de la otra mostró que era igual a la que tendría si estuviera sola en el modelo, lo que condujo a la paradoja estadística en que las pruebas de t separadas lo que miden es el efecto marginal de una variable dado que otras están en el modelo, cuando la prueba de F^* dice que al menos una $\beta_k \neq 0$. Se recomienda consultar las obras de Chatterjee & Price (1977), Neter *et al*. (1983) y Wonnacott & Wonnacott (1981) al respecto.

2.14.1 Métodos informales para detectar multicolinealidad. Se reportan indicaciones de multicolinealidad severa con algunos diagnósticos simples (Neter *et al*. 1983) como los siguientes:

1. Cambios drásticos en los coeficientes de regresión estimados cuando se agregan o quitan variables a un modelo. (Ver 2.13.2)

2. Lo mismo del anterior cuando se borra o altera una observación.

3. Pruebas de t no significativos para los β_k individuales cuando las de F parecen mostrar lo contrario. (Ver 2.13.4)

4. Cambios de signos en los β_k al agregar o suprimir variables.

5. Altos valores en los coeficientes de la matriz de correlación entre las variables independientes, como los mostrados en la Tabla 2.14, para el modelo con 3 variables ya analizado. Para algunos autores, valores de $R > 0.5$ ya la presumen, como se ve entre d y h.

6. Intervalos de confianza muy amplios (inestabilidad para variables muy importantes del modelo).

Tabla 2.14. Matriz de coeficientes de correlación estimados: modelo con 3 variables independientes.

	Constante	*d*	*h*	*ed*
Constante	1.0000	0.5150	-0.8502	-0.1737
d	0.5150	1.0000	-0.8252	-0.8470
h	-0.8502	-0.8252	1.0000	0.4310
ed	-0.8470	-0.8470	0.4310	1.0000

2.14.2 Cuantificación de la multicolinealidad

Los diagnósticos anteriores no la cuantifican. Existen algunas pruebas para ello como la expuesta en las lecturas complementarias de la sección 2.19, para detectarla con base en los factores de inflación de la varianza *VIF* que permiten medir incrementos en las varianzas de los coeficientes estimados de regresión comparados con los de variables no linealmente relacionadas. Se debe intentar entonces reducir los problemas de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ que muestra gran sensibilidad cuando su determinante es cercano a cero o cuando los valores de las variables difieren enormemente en el orden de sus magnitudes, para lo cual se recurrirá a la transformación de correlaciones.