CAPÍTULO 1

REGRESIÓN LINEAL SIMPLE

1.1. Introducción

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

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

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

Ejemplo 1. En la siguiente tabla se muestra la tasa de mortalidad infantil (muertes de niños de 5 años o menos por cada 1,000 nacidos vivos) y el porcentaje de vacunación en veinte países del mundo. Los datos fueron tomados de un reporte de la UNICEF del año 1994.

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

El objetivo es hallar una ecuación que represente lo más preciso posible la relación entre la variable independiente: el porcentaje de inmunización, y la variable dependiente:la tasa de mortalidad. La figura 1.1 muestra el plot de los datos, obtenido usando el programa R, el comando udsado aparece en el laboratorio 1(ver apéndice del texto). El plot suguiere que hay una acceptable relación lineal entre las variables. Además, la tasa de mortalidad tiende a bajar a medida que aumenta el porcentaje de inmunización.

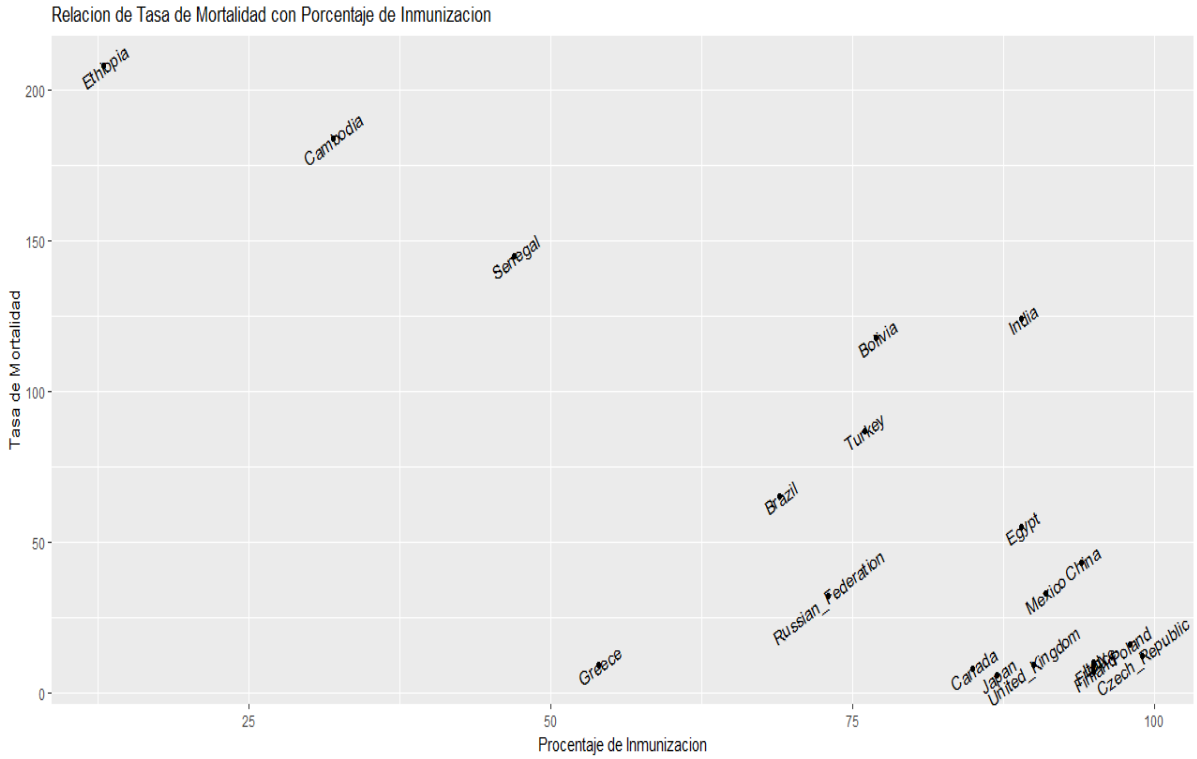

Figura 1.1 Plot que relaciona la tasa de mortalidad con el porcentaje de inmunización en cada país

De la figura 1.1 se puede ver que los países Ethiopia(8), Cambodia(3) y Senegal(18) parecen estar algo alejados de la mayoría de los datos. Igualmente, Greece(11) e India(12) aparecen algo fuera de la tendencia. No es muy obvio concluir que hay una relación lineal entre las variables. La figura 1.2 muestra la línea de regresión obtenida usando el programa R. Los comandos aparecen en el laboratorio 1 que aparece en el apéndice del texto.

La salida obtenida en R para la regresión lineal correspondiente es como sigue:

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

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

X -2.1359 0.3893 -5.4864 0

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

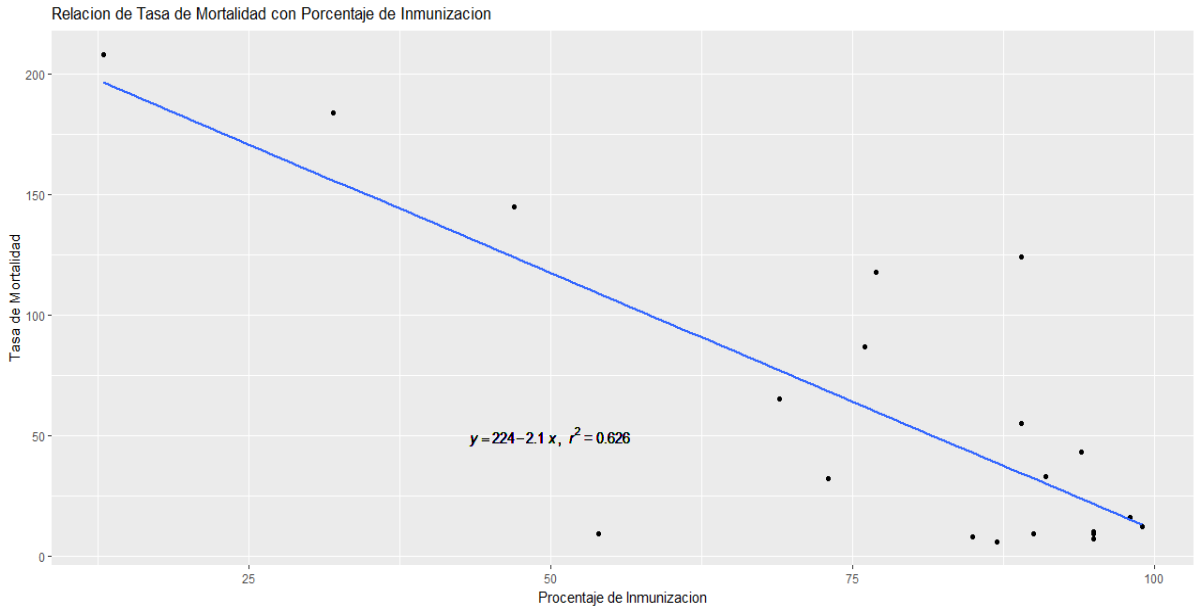

Figura 1.2 Linea de Regresión para los datos del ejemplo 1.

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

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

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

Se observa que el R^2 subió a un 86.2%, que es bastante aceptable. Asimismo en la figura 1.3 muestra la nueva linea de regresión que ajusta a mejor a los datos.

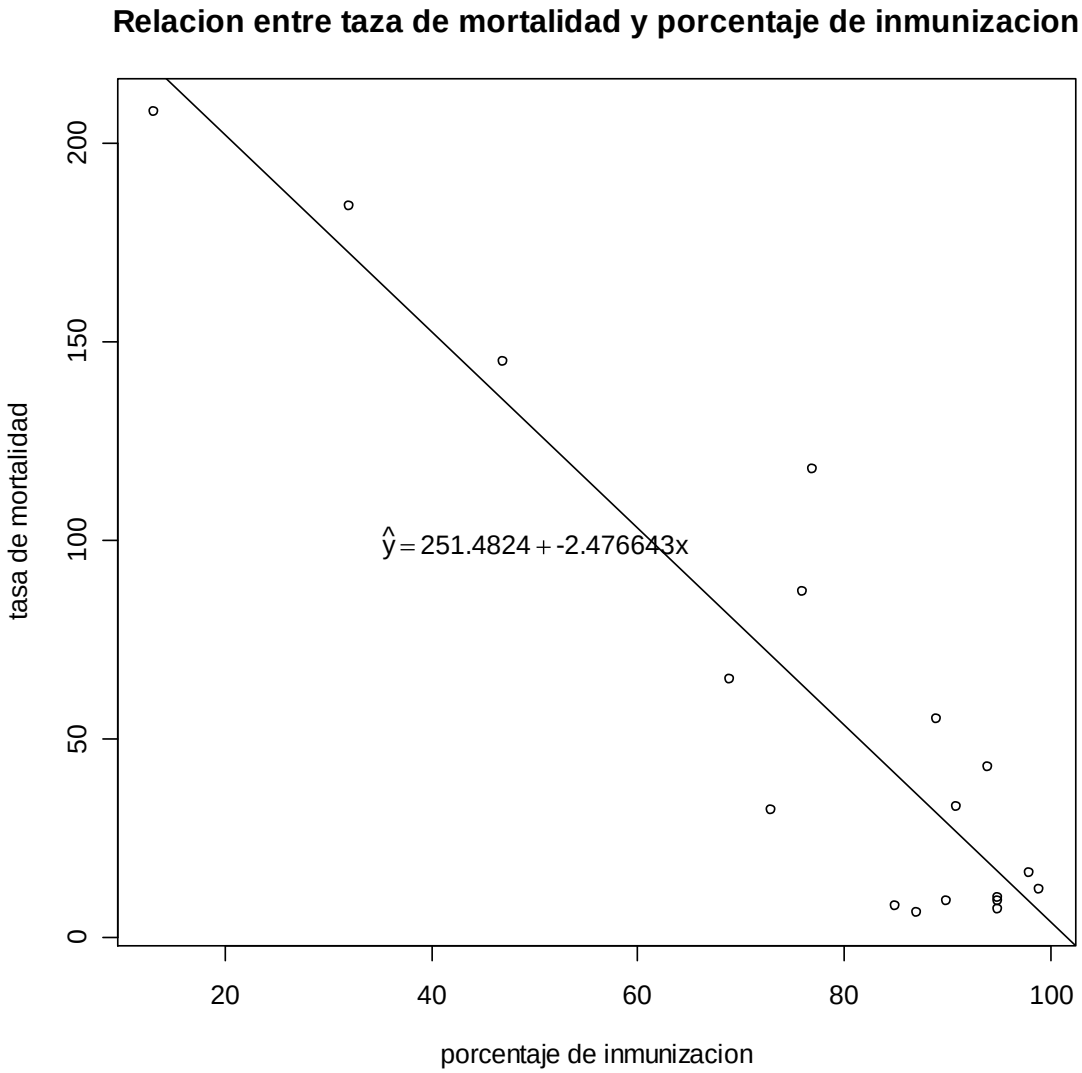

Figura 1.3 Linea de regresión despues de eliminar las observaciones atipícas 11 y 12

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

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

1.1.1 Usos del análisis de regresión:

Los siguientes son los principales usos de un modelo de regresión, aunque frecuentemente estos se dan al mismo tiempo en el análisis de un conjunto de datos:

a) **Predicción:** El objetivo aquí es pronosticar valores de la variable de respuesta para valores futuros de la variables predictoras, es decir para valores más allá del rango de valores de las variables

predictoras presentes en la muestra de entrenamiento. Tal vez ésta sea la razón principal para usar regresión en el análisis estadístico.
b) **Descripción:** La idea es establecer una ecuación lineal o linealizable que describa la relación entre la variable dependiente y las variables predictoras.
c) **Control:** Se busca controlar el comportamiento o variación de la variable de respuesta de acuerdo a los valores que asumen las variables predictoras. Por ejemplo, cuantas horas debería estudiar como mínimo un estudiante para sacar 90 puntos o más en un examen.
d) **Selección de variables:** Inicialmente se pueden haber considerado muchas variables para explicar el comportamiento de la variable de respuesta a través de un modelo lineal, pero la presencia de muchas variables predictoras puede afectar el rendimiento del modelo además de que la computación del mismo se puede volver lenta. Por lo tanto, hay que usar técnicas para escoger solo las variables predictoras que sean más relevantes y aquellas que no sean redundantes en explicar la variación de la variable de respuesta.

1.2 El modelo de Regresión Lineal simple

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

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

Aqui α y β son el intercepto y la pendiente del modelo de regresión respectivamente y ε es un error aleatorio. El modelo es lineal porque la variable predictora no esta elevado a ninguna potencia o no es usada como argumento de otra function. Por otro lado, si se toma una muestra, que es representada por los n pares ordenados (X_i,Y_i) entonces el modelo se puede escribir como

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

Las constantes α y β son los parámetros del modelo, e_i para i=1,2, …n, es una muestra aleatoria del error aleatorio ε, al igual que Y_i (i=1,2,…,n) es una muestra aletoria de la variable aletoria Y. Los parámetros α y β son estimados en base a la muestra estimada y a la ecuación lineal

$$Y = \hat{\alpha} + \hat{\beta} X$$

es llamada la línea de regresión estimada.
En la figura 1.4 se muestra la línea de regresión estimada y los errores estimados para algunas de las observaciones de la muestra.

Suposiciones del modelo:

a) La variable predictora X es no aleatoria y se supone que ha sido medida con la mejor precisión posible. Sin embargo hay algunas situaciones donde también se supone que X es aleatoria .
b) Los errores e_i son variables aleatorias con media 0 y varianza constante σ^2. Por ahora no se requerirá que los errores tengan una distribución normal.
c) Los errores e_i y e_j (i≠j=1…,n) son independientes entre si. Es decir, $Cov(e_i, e_j) = 0$

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

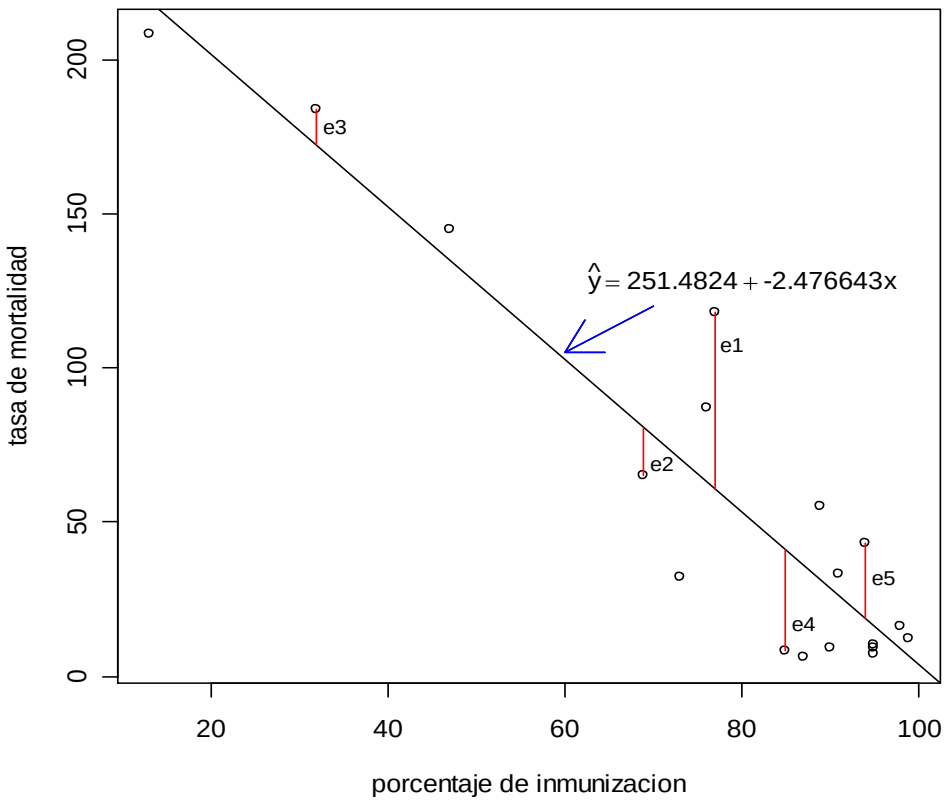

Figura 1.4 Errores con repecto a la linea de regresión para algunas de las observaciones del ejemplo 1

1.2.1 Estimación de la línea de regresión usando Mínimos Cuadrados

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

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

O más especificamente que

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

Es decir, la esperanza (o media) condicional de y dado que la variable predictora asume el valor de x es una ecuación lineal en x. La notación anterior es mas adecuada cunado se considera que x también es aletaoria. La ecuación (1.4) es llamada la línea de regresión poblacional.

Los parámetros α y β deben ser estimados en base a la muestra tomada. El método usual para estimarlos es el de los cuadrados mínimos. La idea es minimizar la suma de los cuadrados de los errores e_i, con respecto a α y β. Es decir,

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

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

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

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

simplificando ambas ecuaciones se obtiene

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

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

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

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

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

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

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

De la primera ecuación normal es fácil ver que:

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

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

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

como $\frac{\partial^2 Q(\alpha,\beta)}{\partial^2 \alpha} = 2n > 0$ y

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

Finalmente, la línea de regresión estimada, llamada también la línea ajustada, por cuadrados mínimos será:

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

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

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

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

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

También se puede usar el método de Máxima verosimilitud para estimar los coeficientes de la línea de regresión, pero se necesita considerar la suposición de que los errores aleatorios e_i se distribuyen normalmente con media cero y varianza σ^2. En este caso las estimaciones se obtienen maximizando la función de verosimilitud $L(\alpha,\beta,\sigma^2) = \prod_{i=1}^{n} f(y_i - \alpha - \beta x_i)$, donde f representa la función de densidad de una normal $N(0,\sigma^2)$. Sin embargo, es más frecuente trabajar con el logaritmo de la función de verosimilitud. Luego, se deriva log L con respecto a α, β, σ^2 y se iguala a cero. Al resolver el sistema de ecuaciones se obtiene las estimaciones de los coeficientes α y β, asi como el estimador de la varianza σ^2.

Las propiedades de los estimadores mínimo cuadráticos de los coeficientes de regresión se discuten en las secciones 1.2.3 y 1.2.4.

1.2.2 Interpretación de los coeficientes de regresión estimados

La pendiente $\widehat{\beta}$ indica el cambio promedio en la variable de respuesta cuando la variable predictora aumenta en una unidad adicional. El intercepto $\widehat{\alpha}$ indica el valor promedio de la variable de respuesta cuando la variable predictora vale 0. Sin embargo, carece de interpretación práctica si es irrazonable pensar que el rango de valores de x incluye a cero. Cuando se tiene evidencia de que la variable de respuesta assume el valor 0 cuando la predictora es cero, entonces es mas razonable ajustar una linea de regresión sin intercepto, veáse ejercicio 4.

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

Tasa_mort = 224.316 – 2.13587%inmuniz,

lo que significa que en promedio la tasa de mortalidad de niños menores de 5 años disminuirá en promedio en 2.13 cuando el % de inmunizacion aumenta en uno porciento.

Por otro lado la tasa de mortalidad promedio de los paises donde no hay inmunización será de 224.316. Aunque es difícil pensar que exista un país donde no se vacunen a los niños, ya que muchas veces la UNICEF dona las vacunas.

1.2.3 Propiedades de los estimadores mínimos cuadráticos de regresión

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

Recordar que:

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

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

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

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

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

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

Luego,

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

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

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

Usando la propiedad que Var(cy)=c²Var(y) y el hecho que la suposición de que Cov(e_i,e_j)=0 es equivalente a Cov(y_i,y_j)=0, se tiene que Var($\sum_{i=1}^{n} c_i y_i$)= $\sum_{i=1}^{n} c_i^2 Var(y_i)$. En consecuencia,

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

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

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

Luego,

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

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

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

Hay otra forma de calcular la varianza de $\hat{\alpha}$, usando el hecho que $Cov(\bar{y}, \hat{\beta}) = 0$, veáse ejercicio 1. Las propiedades discutidas en esta sección serán usadas cuando se haga inferencia estadística para el modelo de regresión.

1.2.4 Distribución de los estimadores mínimos cuadráticos

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

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

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

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

1.2.5 Propiedades de los residuales

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

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

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

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

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

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

1.2.6 Estimación de la varianza del error

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

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

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

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

En esta verificación consideraremos que $(y_i - \hat{y}_i)$
es aleatoria,

Notar que

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

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

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

Por otro lado,

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

Asímismo,

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

Sustituyendo (1.26) en (1.25) se obtiene que

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

Reemplazando (1.27) en (1.24) se llega a

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

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

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

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

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

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

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

Por lo tanto,

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

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

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

Una vez que se fija la línea de regresión y se estiman los errores por los residuales se tiene que un estiamdo de la varianza del error es

$$s^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

1.2.7 Descomposición de la suma de cuadrados total

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

En la figura 1.5 se puede ver que la desviación de un valor observado y_i con respecto a la media $\bar{y}$ se puede escribir como

$$(y_i - \bar{y}) = (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y}) \qquad (1.31)$$

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

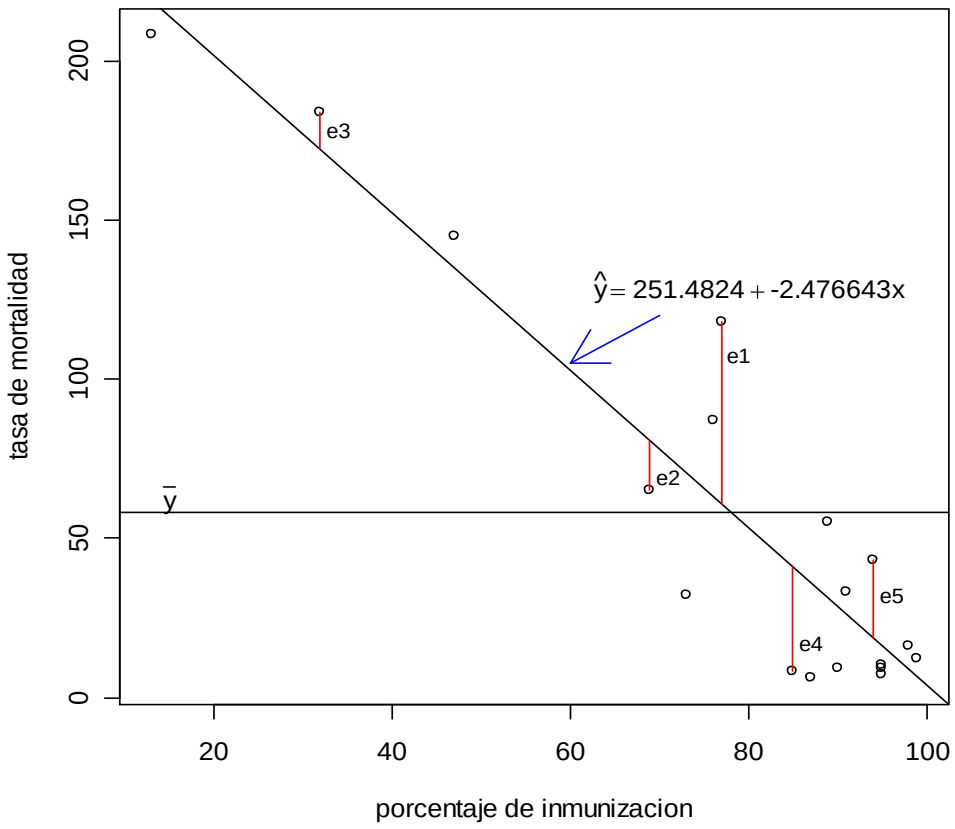

Figura 1.5 Diagrama para descomponer la deviación total en desviacion debido a la regresion mas desviacion debido al error.

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

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

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

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

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

donde,

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

y's.

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

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

Las sumas de cuadrados definidas anteriormente son variables aleatorias, pues dependen de y, la cual es aleatoria. Asi, si en *SSR* se sustituye $\hat{y}_i$ por $\hat{\alpha} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$ se tiene que

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

Por otro lado, considerando las sumas de cuadrados como variables aleatorias y tomando valor esperado en cada lado de relación (1.35) se tiene

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

Luego,

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

Asimismo, E(SST)=E(SSR)+E(S SSE). Asi que,

$$E(SST) = \sigma^2 + \beta^2 S_{xx} + (n-2)\sigma^2 = (n-1)\sigma^2 + \beta^2 S_{xx} \qquad (1.37)$$

Las sumas de cuadrados juegan un papel muy importante cuando se hace inferencia en regresión, por eso es importante saber como es su distribución. Por teoría de modelos lineales, se puede establecer que las sumas de cuadrados son formas cuadráticas de las variables y_i y por lo tanto se distribuyen como una Ji-cuadrado. Más especificamente, se pueden establecer los siguientes resultados:

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

ii). $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$ (Ji-cuadrado con n-2 grados de libertad). Para calcular *SSE* se usan *n* datos

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

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

1.2.8 El Coeficiente de Determinación R^2

Es una medida de la bondad de ajuste del modelo. Está definido por

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

También indica que porcentaje de la variación de la variable de respuesta es explicada por su relación lineal con la variable predictora. Un modelo de regresión con R^2 mayor o igual a 75% se puede considerar bastante aceptable. Aunque se puede ser un poco flexible dependiendo del tipo de datos y de la cantidad de datos disponible.

En el ejemplo 1, sólo un 62.6% de la variacion de la mortalidad infantil de niños menores de 5 años es explicada por su relación lineal con el porcentaje de inmunización, lo cual no es muy alto y hace poco confiable las predicciones.

Lamentablemente, el valor de R^2 es afectado por la presencia de valores anormales. Asi, un valor de R^2 bien cercano al 100% no necesariamente garantiza una buena predicción del modelo. Pero si se puede decir que un modelo con R^2 bajo es inadecuado para hacer predicciones.

1.3 Inferencia en Regresión Lineal Simple

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

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

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

equivalente a decir, que el estadístico $z = \dfrac{\hat{\beta} - \beta}{\dfrac{\sigma}{\sqrt{Sxx}}}$ se distribuye como una normal estándar, N(0,1).

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

$$t = \frac{z}{\sqrt{\dfrac{\chi^2_{(n-2)}}{n-2}}} = \frac{\hat{\beta} - \beta}{\dfrac{s}{\sqrt{Sxx}}} \qquad (1.38)$$

se distribuye como una t de Student con n-2 grados de libertad. El estadístico *t* dado en 1.38 es usado para hacer prueba de hipótesis y calcular intervalos de confianza acerca de β.

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

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

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

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

```
> summary(l2)

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

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

Coefficients:
```

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

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

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

```
> # Hallando el intervalo de confianza del 95% para la pendiente Beta
> bint<-c(beta-qt(.975,18)*eebeta,beta+qt(.975,18)*eebeta)
> bint
[1] -2.95290 -1.31890
```

Luego, el Intervalo de confianza del 95% para β será

(-2.95290, -1.31890)

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

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

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

Ahora se considerará prueba de hipótesis para los coeficientes de la linea de regresión. Desde el punto de vista clásico, la siguiente tabla muestra la manera de hacer pruebas de hipótesis para la pendiente β, asumiendo que su valor es β* y con un nivel de significación α.

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

$t_{cal}<-t_{(\alpha,n-2)}$	$\|t_{cal}\|>t_{(\alpha/2,n-2)}$	$t_{cal}>t_{(\alpha,n-2)}$

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

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

Solución: Usando los resultados del laboratorio 2 de R.

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

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

Las hipótesis serán:
Ho: $\beta=0$ (Es decir, no hay relación lineal entre las variables)
Ha: $\beta\neq0$ (Hay relación lineal entre las variables)

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

Similarmente, se pueden hacer pruebas de hipótesis para el intercepto.

Las hipótesis serán:
Ho: $\alpha=0$ (La línea de regresión poblacional pasa por el origen)
Ha: $\alpha\neq0$ (La línea de regresión poblacional no pasa por el origen)

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

1.3.2 El análisis de varianza para regresión lineal simple

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

Al dividir las suma de cuadrados por sus grados de libertad se obtienen los **cuadrado medio**.
Así se tienen tres cuadrados medios

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

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

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

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

siempre que la hipótesis nula Ho:β=0 es cierta. Aquí el numerador tiene 1 grado de libertad y el denominador tiene n-2. La independencia descansa en el hecho que $COV(\hat{Y}_i - \bar{Y}, \hat{Y}_i - Y_i) = 0$, veáse ejercicio 10.

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

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

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

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

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

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

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

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

Talvez el uso más frecuente que se le da a una línea de regresión es para hacer predicciones acerca de la variable de respuesta *Y* para un valor dado de x. Supongamos que queremos predecir el valor medio de las *Y* para un valor x_0 de la variable predictora x. Es decir, $E(Y/x=x_0)=\alpha+\beta x_0$. Es natural pensar que el estimado puntual sera $\hat{y}_0=\hat{\alpha}+\hat{\beta}x_0$. Sin embargo, es muy riesgoso predecir basado en un solo valor y es más conveniente usar un intervalo donde se espera que caiga el valor de *Y* con un cierto grado de confianza.Como $\hat{\alpha}$ y $\hat{\beta}$ se distribuyen normalmente, entonces $\hat{y}_0$ también se distribuye normalmente con media $\alpha+\beta x_0$ y varianza igual a:

$$Var(\hat{Y}_0)=Var(\hat{\alpha}+\hat{\beta}x_0)=Var(\hat{\alpha})+x_0^2Var(\hat{\beta})+2x_0Cov(\hat{\alpha},\hat{\beta})$$

Sustituyendo expresiones halladas en la sección 1.2.3 y el hecho que $Cov(\hat{\alpha},\hat{\beta})=-\frac{\bar{x}\sigma^2}{Sxx}$ (ver ejercicio 5), se tiene:

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

de donde resulta

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

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

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

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

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

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

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

Luego,

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

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

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

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

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2,n-2)} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}} \qquad (1.40)$$

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

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

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

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

$se.fit
[1] 9.032315

$df
[1] 18

$residual.scale
[1] 40.13931

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

$se.fit
[1] 9.032315
```

```
$df
[1] 18

$residual.scale
[1] 40.13931
```

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

La figura1.6 muestra las bandas de confianza y predicción del 95 porciento para los datos del ejemplo 1.

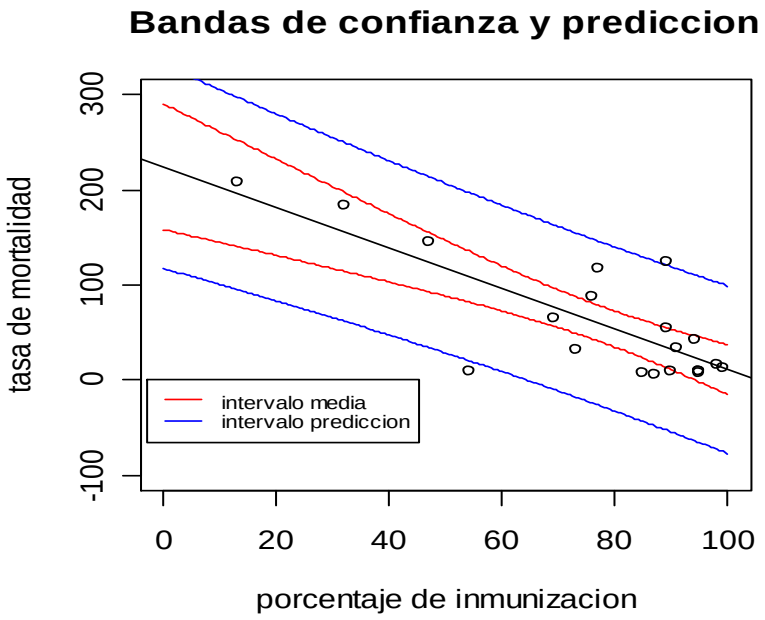

Figura 1.6 Bandas de confianza y predicción para los datos del ejemplo 1.

1.4 El Coeficiente de Correlación

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

El coeficente de correlación, es un valor que mide el grado de asociacón lineal entre las variables aleatorias X y Y y se define como

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{1.41}$$

Se puede mostrar que

a) $-1 \le \rho \le 1$
b) Si $\rho^2=1$ entonces Y=A + BX, con probabilidad 1. Si Y=A + BX, donde A y B son constantes, entonces si A>0, ρ=1 y si A<0, ρ=-1.
c) Si la regresión de Y sobre X es lineal, esto es la media condicional de Y dado X es $E(Y/X) = \alpha + \beta x$. Entonces, $\beta = \rho \frac{\sigma_y}{\sigma_x}$, y $\alpha = \mu_y - \beta \mu_x$. Notar que si la pendente de la línea de regresión es cero entonces la correlación es 0, y que β y ρ varian en la misma dirección.
d) Si (X,Y) se distribuye como una normal bivariada, entonces la varianza condicional de Y dado X, está dado por $\sigma^2_{y/x} = \sigma^2_y(1-\rho^2)$. Luego, si $\rho = \pm 1$, entonces $\sigma^2_{y/x} = 0$, implicando que hay una perfecta relación lineal entre Y y X. Más especificamente, si $\rho = 1$, entonces X y Y crecen en la misma dirección y si $\rho = -1$, Y decrece cuando X crece.

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

$$r = \frac{Sxy}{\sqrt{SxxSyy}} \tag{1.42}$$

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

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

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

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

Advertencia: ***Correlación alta no implica necesariamente una relación causa efecto entre las variables. Usando la fórmula de correlación entre dos variables que en la vida real no tiene***

ninguna relación entre si, (por ejemplo X: peso de los profesores y Y=salario del profesor) se puede obtener un r bastante alto cercano a 1 o –1 y eso no implica necesariamente que X explique el comportamiento de Y (podria darse el caso que mientras menos pesa un profesor gana menos).

La figura 1.7 muestra varios diagramas de puntos y sus respectivas correlaciones. Los datos y comandos correspondientes aparecen el laboratorio 3 del texto.

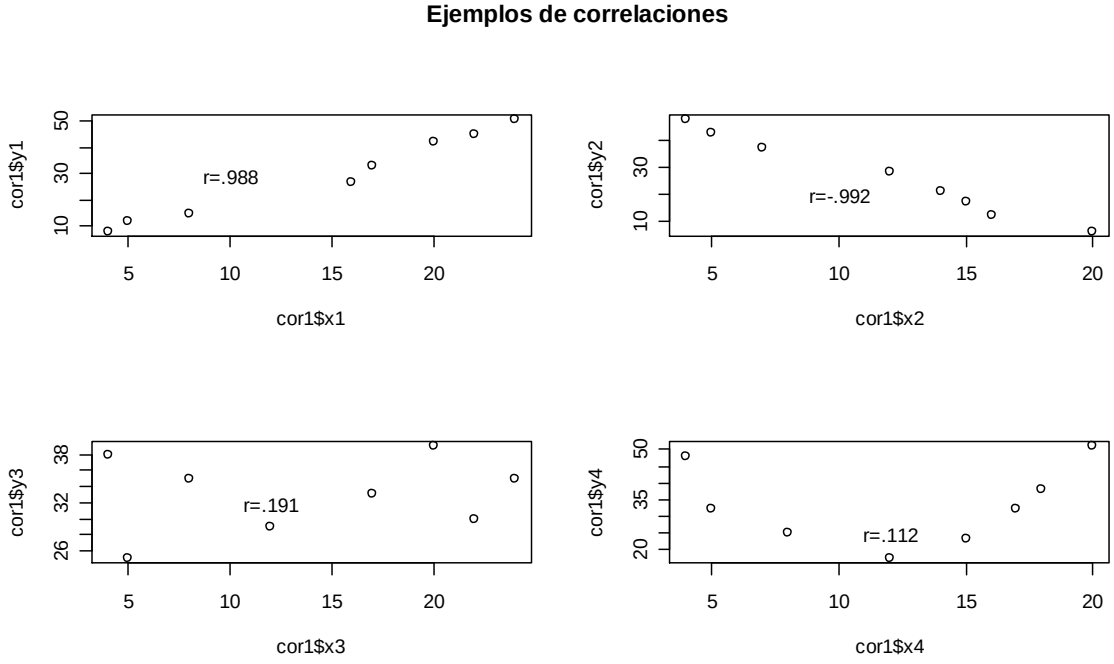

Figura 1.7 Cuatro distintos patrones de datos y sus correspondientes coeficientes de correlación

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

El valor de correlación es afectado por la presencia de valores anormales, en la figura 1.8 se puede ver el efecto de los valores anormales en el valor de la correlación para 4 diferentes relaciones. Los datos y comandos correspondientes aparecen el laboratorio 3 del texto.

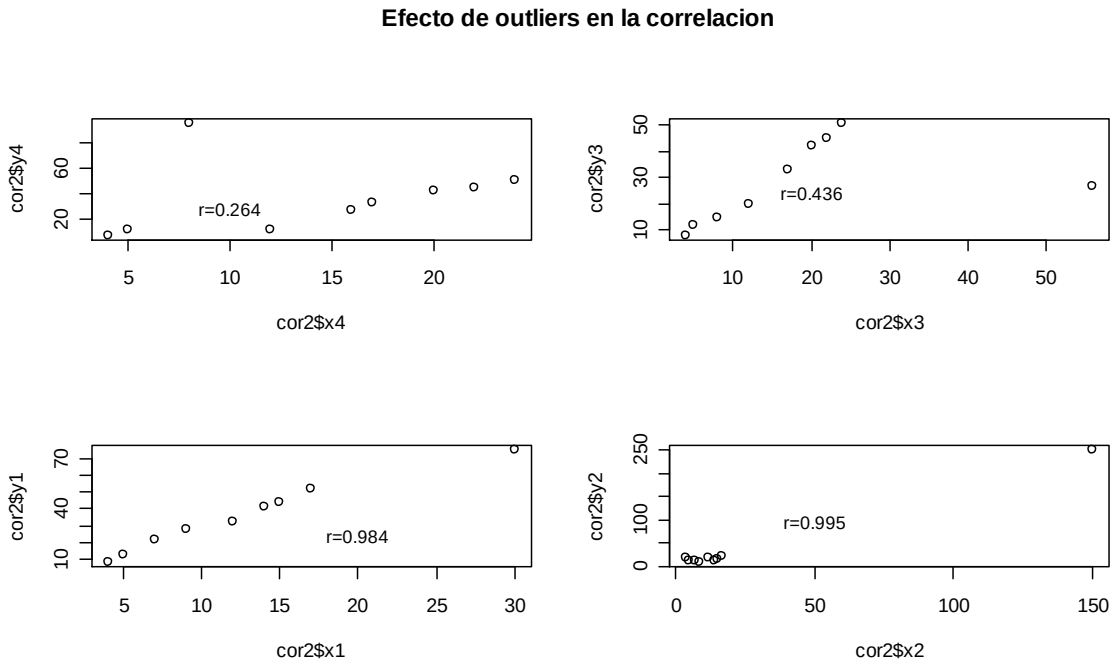

Figura 1.8 Cuatro distintos patrones de datos para mostrar el efecto de outliers en el coeficiente de correlación

Interpretación de la figura 8: *En el primer caso existe un valor bastante anormal en la dirección vertical que hace que la correlación sea bastante bajo a pesar de que los otros valores parecen estar bastante alineados.*

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

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

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

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

$$t = \frac{\hat{\beta}}{\frac{s}{\sqrt{Sxx}}} = \frac{r\sqrt{\frac{Syy}{Sxx}}}{\sqrt{\frac{\frac{Syy(1-r^2)}{n-2}}{Sxx}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)} \qquad (1.43)$$

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

$$z = \frac{1}{2}\log(\frac{1+r}{1-r}) = \tanh^{-1}(r) \qquad (1.44)$$

la cual tiene una distribución aproximadamente normal con media

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

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

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

1.5 Análisis de residuales

Los residuales, $r_i = y_i - \hat{y}_i$ que son estimaciones de los errores aleatorios del modelo $\hat{e}_i = y_i - \hat{y}_i = y_i - \alpha - \beta x_i$, son importantes para establecer si las suposiciones del modelo se cumplen y para explorar el porqué de un mal ajuste del modelo. La manera más fácil de examinar los residuales es mediante plots los cuales permiten cotejar:

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

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

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

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

ii) Residual Estudentizado: En el residual estandarizado se está considerando de antemano que cada residual tiene la misma varianza, pero en realidad cada uno de ellos tiene su propia varianza como se muestra a continuación.

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

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

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

calculando la covarianza, se obtiene

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

En consecuencia,

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

Por lo tanto, usando el correspondiente estimado para σ^2 se tiene que:

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

el cual es llamado el i-ésimo **residual estudentizado**. En algunos programas estadísticos, los r_i^* son llamados residuales estandarizados. También son llamados residuales estudentizados internamente (ver más adelante, la sección 3.1).

1.5.1 Cotejando normalidad de los errores y detectando outliers

Normalidad de los errores es un requisito indispensable para que tengan validez las pruebas estadísticas de t y F que se usan en regresión. Existen varios métodos gráficos y pruebas estadísticas tanto paramétricas cono noparamétricas para cotejar la normalidad de un conjunto de datos. La manera más fácil es usando gráficas tales como histogramas, "stem-and-leaf" o "Boxplots".
Una gráfica más especializada es el plot de Normalidad. Aquí se plotea los residuales versus los valores que se esperarían si existiera normalidad, estos valores son llamados los scores normales. Dado el i-ésimo residual, su score normal se encontrará determinando primero a que percentil le corresponde en la distribución de los datos, se han propuesto varias maneras de hacer esto. Luego de determinar el percentil se halla el valor que le corresponde a dicha percentil en la distribución normal estándar. El i-ésimo score normal es aproximado en forma bastante precisa por

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

donde Φ representa la función de distribución acumulada de una normal estándar y n es el número de observaciones en la muestra.
Habrá normalidad si los puntos del plot se alinean cerca de una linea que pasa por el origen. Si se usan los residuales estudentizados la línea además de pasar por el origen deberia tener pendiente cercana a 1.

Ejemplo 5. Cotejar si existe normalidad para los datos del ejemplo 1.

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

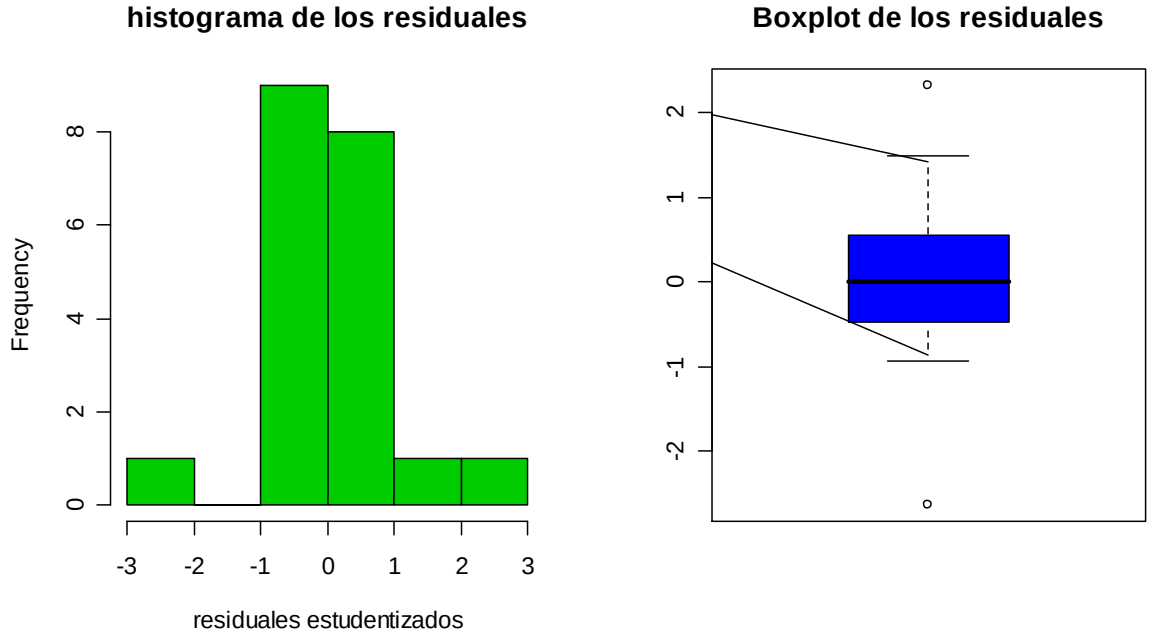

Figura 1.9 Histograma y boxplot de residuales de la regresion del ejemplo 1.

El histograma no parece ser de forma acampanada, es decir no hay normalidad, además parece haber un "outlier" inferior. El boxplot indica bastante simetría en el centro pero no así en los extremos de la distribución. Además se identifican dos outliers, uno superior y el otro inferior.
En R, el plot de normailidad se usa usando los comandos **qqnorm** y **qqline**. El plor de Normalidad correspondiente al ejemplo 1 se muestra en la figura 1.10.

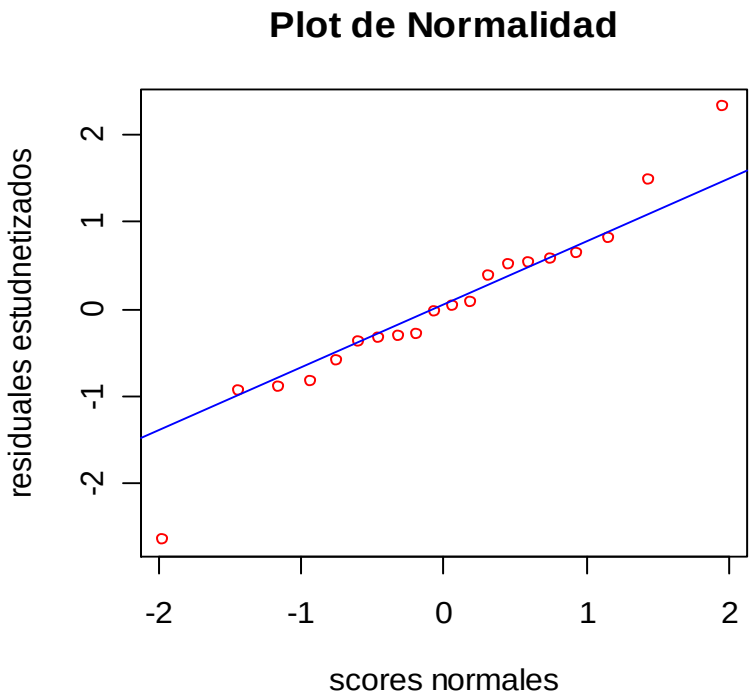

Figura 1.10 Plot de Normalidad para los residuales de la regresion del ejemplo 1.

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

1.5.2 Cotejando que la varianza sea constante

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

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

Los commandos en R para obtener estas graficas aparecen el laboratorio 4 del texto. Para los datos del ejemplo 1 se obtienen los plots que aprecen en las figuras 1.11 y 1. 12.

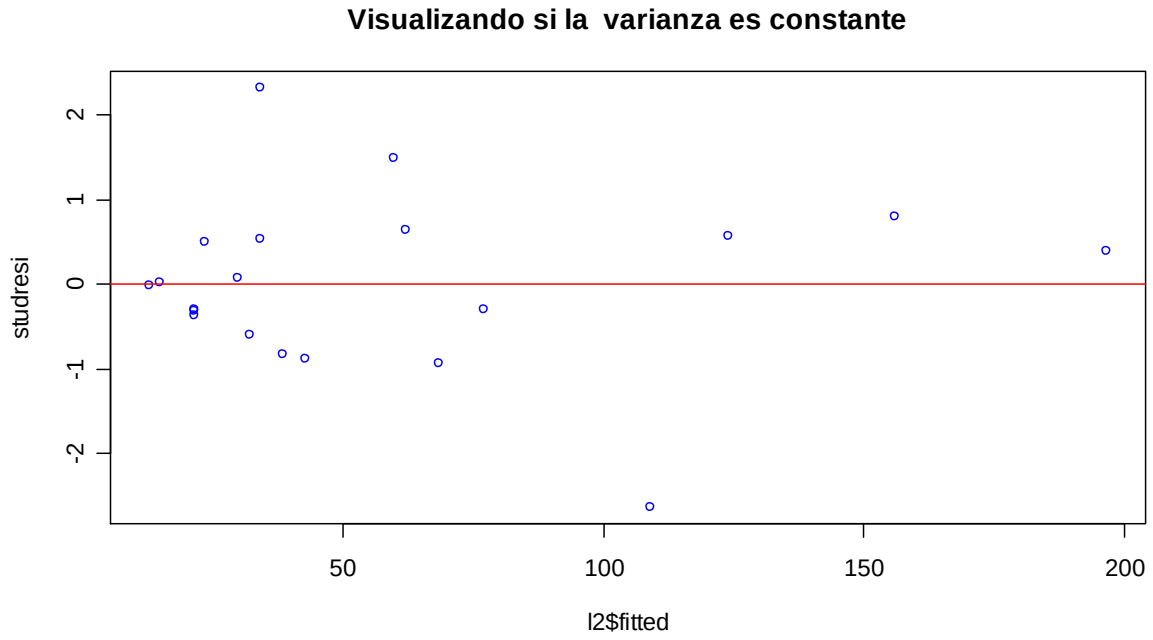

Figura 1.11 Plot de residuales para detectar si la varianza constante con respecto a los valores ajustados

En la Figura 1.11 se puede notar que los puntos se reparten equitativamente alrededor de la línea horizontal. Nuevamente lo que llama más la atención es la presencia del "outlier". Por lo tanto, la varianza parece ser constante. Si ploteamos los residuales versus la variable predictora en lugar de los valores ajustados, se obtiene la siguiente gráfica

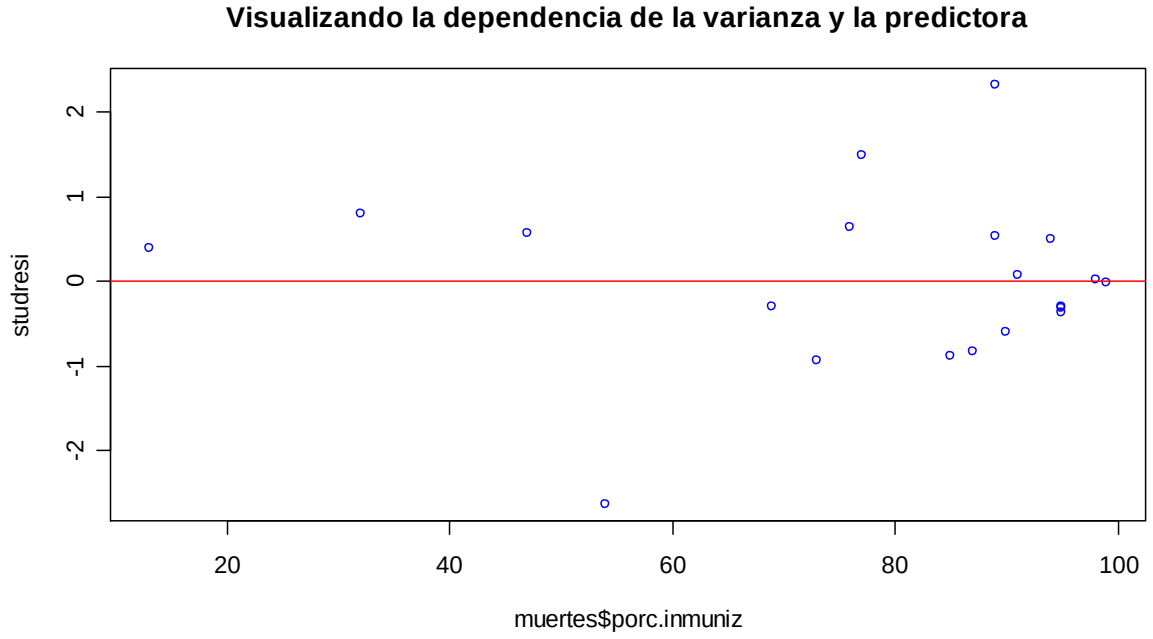

Figura 1.12 Plot de residuales para detectar si la varianza constante con respecto a la variable predictora

Al igual que en la figura anterior, se puede ver en la figura 1.12 que, excluyendo los dos "outliers", los puntos parecen estar en una franja horizontal, por lo tanto se podría considerar que la varianza es constante con respecto a la predictora. Notar que también hay cuatro puntos alejados en la dirección horizontal. Estas observaciones también pueden tener influencia en los cálculos de la línea de regresión.
Si se observa algún patrón en el plot se puede hacer transformaciones en una o en ambas variables para estabilizar la varianza. Otra alternativa es usar ***mínimos cuadrados ponderados***. Nuevamente esto será discutido más detalladamente en el capítulo 3 del texto cuando se haga análisis de residuales en regresión múltiple.

En R se puede hacer un plot simultáneo de los residuales. Usando el laboratorio 4 de R para el ejemplo 1 se obtiene la siguiente gráfica.

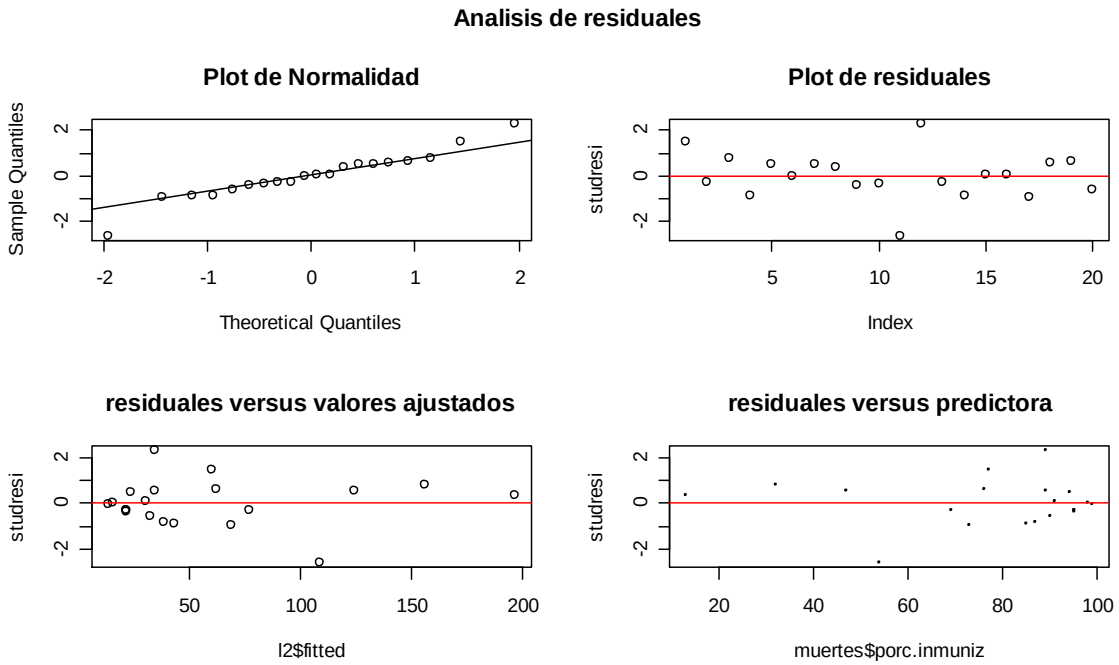

Figura 1.13 Plots para hacer análisis de residuales

1.5.3 Cotejando si los errores estan correlacionados.

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

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

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

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

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

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

EJERCICIOS

1. Considerando un modelo de regresión lineal simple, calcular $Cov(\bar{Y}, \hat{\beta})$
2. Probar que la línea de regresión estimada pasa por $(\bar{X}, \bar{Y})$
3. En un modelo de regresión lineal simple calcular E[SST]=E[$\sum_{i=1}^{n}(y_i - \bar{y})^2$]

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

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

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

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

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

7. En un pueblo se eligen 15 personas al azar y se anota su salario mensual (X) , y la cantidad que ahorran mensualmente (Y). Ambas cantidades están expresadas en dólares. (ver datos **salarios** en la página de internet del curso).

Salario	Ahorro
800	150
850	100

900	280
1200	400
1500	350
1700	500
1900	635
2000	600
2300	750
2500	680
2700	900
3000	800
3200	300
3500	1200
5000	1000

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

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

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

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

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

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

10. Considerando un modelo de regresión lineal simple, calcular

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

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

12. Suponga que en el modelo de regresión lineal simple los valores x_i y y_i son reemplazados por ax_i+b y cy_i+d respectivamente donde a,b,c y d son constantes tales que $a\neq 0$ y $c\neq 0$. Cuál es el efecto

de estas transformaciones en $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$, R^2 y la prueba estadística para probar la hipótesis nula Ho:β=0.?

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

14. Los siguientes datos fueron recolectados por el físico James Forbes para estimar indirectamente la altura sobre el nivel del mar de acuerdo a las mediciones del punto de ebullición (boiling point) del agua. (ver datos **Forbes** en la página de internet del curso).

Columna 1: Numero de la observación
Columna 2: Boiling Point(ºF)
Columna 3: Pressure (in. Hg)
Columna 3: Log(Pressure)
Columna 4: 100*Log(Pressure)

1 194.5 20.79 1.3179 131.79
2 194.3 20.79 1.3179 131.79
3 197.9 22.40 1.3502 135.02
4 198.4 22.67 1.3555 135.55
5 199.4 23.15 1.3646 136.46
6 199.9 23.35 1.3683 136.83
7 200.9 23.89 1.3782 137.82
8 201.1 23.99 1.3800 138.00
9 201.4 24.02 1.3806 138.06
10 201.3 24.01 1.3804 138.04
11 203.6 25.14 1.4004 140.04
12 204.6 26.57 1.4244 142.44
13 209.5 28.49 1.4547 145.47
14 208.6 27.76 1.4434 144.34
15 210.7 29.04 1.4630 146.30
16 211.9 29.88 1.4754 147.54
17 212.2 30.06 1.4780 147.80

a) Hacer un diagrama de puntos de Pressure versus Boiling point. Piensa Ud. que hay una tendencia lineal.
b) Hacer un diagrama de puntos de 100*log(Pressure) versus Boiling point. Piensa Ud. que se observa mejor la tendencia lineal que en a)
c) Ajustar la linea de regresión de 100*log(Pressure) versus Boiling point. Trazar la linea sobre el plot hallado en b). Comentar los coeficientes de regresión. Interpretar los "p-values" de la prueba t y el de la prueba F.
d) Interpretar el Coeficiente de determinación R^2
e) Obtener un intervalo de confianza del 99% para β. Interpretar su resultado
f) Obtener un intervalo de confianza del 99% para el valor predicho y un intervalo de confianza para el valor medio de 100*log(Pressure) cuando el Boiling Point es de 195 ºF.

15. Los siguientes datos fueron recolectado para tratar de pronosticar el nivel del agua del rio Snake en Wyoming. (ver datos **River** en la página de internet del curso).

Columna 1: Contenido de agua en la nieve caida hasta Abril 1, desde 1919 hast 1935
Colunna 2: Produccion de agua (en pulgadas) en el rio Snake, entre los meses de abril y julio.

23.1 10.5
32.8 16.7
31.8 18.2
32.0 17.0
30.4 16.3
24.0 10.5
39.5 23.1
24.2 12.4
52.5 24.9
37.9 22.8
30.5 14.1
25.1 12.9
12.4 8.8
35.1 17.4
31.5 14.9
21.1 10.5
27.616.1

a) Hacer un diagrama de puntos de la producción de agua versus cantidad de agua en la nieve. Piensa Ud. que hay una tendencia lineal.
b) Ajustar la linea de regresión producción de agua versus cantidad de agua en la nieve. Trazar la linea sobre el plot hallado en a). Comentar los coeficientes de regresión. Interpretar los "p-values" de la prueba t y el de la prueba F.
c) Interpretar el Coeficiente de determinación R^2.
d) Hallar un intervalo de confianza del 95% para la pendiente. Interpretar su resultado.
e) Obtener un intervalo de confianza del 95% para el valor predicho y un intervalo de confianza para el valor medio de la producción de agua cuando la cantida de agua en la nieve es de un 35%.

16. Los siguientes datos fueron registrados en el río Amazonas (Iquitos, Perú) para observar si la deforestación afecta los niveles del agua del río. (Ver datos **Amazonas** en la página de internet del curso).

columna 1: año
columna 2: nivel de agua máximo (en metros) en el rio Amazonas en Iquitos
columna 3: nivel de agua mínimo (en metros) en el rio Amazonas en Iquitos

1962 25.82 18.24
1963 25.35 16.50
1964 24.29 20.26
1965 24.05 20.97
1966 24.89 19.43
1967 25.35 19.31
1968 25.23 20.85
1969 25.06 19.54
1970 27.13 20.49
1971 27.36 21.91
1972 26.65 22.51

1973 27.13 18.81
1974 27.49 19.42
1975 27.08 19.10
1976 27.51 18.80
1977 27.54 18.80
1978 26.21 17.57

a) Hacer un diagrama de puntos de Nivel máximo versus año, Nivel mínimo versus año y de Nivel Máximo versus Nivel mínimo. Piensa Ud. que hay una tendencia lineal?.
b) Obtener la linea de regresion de Nivel máximo versus año, Nivel mínimo versus año y de Nivel Máximo versus Nivel mínimo. Piensa Ud. que hay una tendencia lineal. Interpretar los coeficientes y los "p-values" de la pruebas t y F.
c) Interpretar el Coeficiente de determinación R^2 para cada una de las 3 regresiones.
d) Obtener un intervalo de confianza del 95% para el valor predicho y un intervalo de confianza para el valor medio del nivel máximo del agua para el año 1980.

17. Los siguientes datos se han recolectado para explicar el rendimiento en millas por gallon de varios modelos de carros.

Columna 1: Modelo de carro
Columna 2.VOL: Volumen de la cabina del carro.
Columna 3.HP: caballos de potencia del motor
Columna 4.MPG: millas promedio por galón
Columna 5 .SP: Velocidad máxima (mph)
Columna 6.WT: Peso de vehiculo (100 lb)

Modelo de carro	VOL	HP	MPG	SP	WT
GM/GeoMetroXF1	89	49	65.4	96	17.5
GM/GeoMetro	92	55	56.0	97	20.0
GM/GeoMetroLSI	92	55	55.9	97	20.0
SuzukiSwift	92	70	49.0	105	20.0
DaihatsuCharade	92	53	46.5	96	20.0
GM/GeoSprintTurbo	89	70	46.2	105	20.0
GM/GeoSprint	92	55	45.4	97	20.0
HondaCivicCRXHF	50	62	59.2	98	22.5
HondaCivicCRXHF	50	62	53.3	98	22.5
DaihatsuCharade	94	80	43.4	107	22.5
SubaruJusty	89	73	41.1	103	22.5
HondaCivicCRX	50	92	40.9	113	22.5
HondaCivic	99	92	40.9	113	22.5
SubaruJusty	89	73	40.4	103	22.5
SubaruJusty	89	66	39.6	100	22.5
SubaruJusty4wd	89	73	39.3	103	22.5
ToyotaTercel	91	78	38.9	106	22.5
HondaCivicCRX	50	92	38.8	113	22.5
ToyotaTercel	91	78	38.2	106	22.5
FordEscort	103	90	42.2	109	25.0
HondaCivic	99	92	40.9	110	25.0
PontiacLeMans	107	74	40.7	101	25.0
IsuzuStylus	101	95	40.0	111	25.0
DodgeColt	96	81	39.3	105	25.0
GM/GeoStorm	89	95	38.8	111	25.0
HondaCivicCRX	50	92	38.4	110	25.0

HondaCivicWagon	117	92	38.4	110	25.0
HondaCivic	99	92	38.4	110	25.0
Subaru Loyale	102	90	29.5	109	25.0
VolksJettaDiesel	104	52	46.9	90	27.5
Mazda323Protege	107	103	36.3	112	27.5
FordEscortWagon	114	84	36.1	103	27.5
FordEscort	101	84	36.1	103	27.5
GM/GeoPrism	97	102	35.4	111	27.5
ToyotaCorolla	113	102	35.3	111	27.5
EagleSummit	101	81	35.1	102	27.5
NissanCentraCoupe	98	90	35.1	106	27.5
NissanCentraWagon	88	90	35.0	106	27.5
ToyotaCelica	86	102	33.2	109	30.0
ToyotaCelica	86	102	32.9	109	30.0
ToyotaCorolla	92	130	32.3	120	30.0
ChevroletCorsica	113	95	32.2	106	30.0
ChevroletBeretta	106	95	32.2	106	30.0
ToyotaCorolla	92	102	32.2	109	30.0
PontiacSunbirdConv	88	95	32.2	106	30.0
DodgeShadow	102	93	31.5	105	30.0
DodgeDaytona	99	100	31.5	108	30.0
EagleSpirit	111	100	31.4	108	30.0
FordTempo	103	98	31.4	107	30.0
ToyotaCelica	86	130	31.2	120	30.0
ToyotaCamry	101	115	33.7	109	35.0
ToyotaCamry	101	115	32.6	109	35.0
ToyotaCamry	101	115	31.3	109	35.0
ToyotaCamryWagon	124	115	31.3	109	35.0
OldsCutlassSup	113	180	30.4	133	35.0
OldsCutlassSup	113	160	28.9	125	35.0
Saab9000	124	130	28.0	115	35.0
FordMustang	92	96	28.0	102	35.0
ToyotaCamry	101	115	28.0	109	35.0
ChryslerLebaronConv	94	100	28.0	104	35.0
DodgeDynasty	115	100	28.0	105	35.0
Volvo740	111	145	27.7	120	35.0
FordThunderbird	116	120	25.6	107	40.0
ChevroletCaprice	131	140	25.3	114	40.0
LincolnContinental	123	140	23.9	114	40.0
ChryslerNewYorker	121	150	23.6	117	40.0
BuickReatta	50	165	23.6	122	40.0
OldsTrof/Toronado	114	165	23.6	122	40.0
Oldsmobile98	127	165	23.6	122	40.0
PontiacBonneville	123	165	23.6	122	40.0
LexusLS400	112	245	23.5	148	40.0
Nissan300ZX	50	280	23.4	160	40.0
Volvo760Wagon	135	162	23.4	121	40.0
Audi200QuatroWag	132	162	23.1	121	40.0
BuickElectraWagon	160	140	22.9	110	45.0
CadillacBrougham	129	140	22.9	110	45.0
CadillacBrougham	129	175	19.5	121	45.0
Mercedes500SL	50	322	18.1	165	45.0
Mercedes560SEL	115	238	17.2	140	45.0
JaguarXJSConvert	50	263	17.0	147	45.0
BMW750IL	119	295	16.7	157	45.0
Rolls-RoyceVarious	107	236	13.2	130	55.0

a) Hacer un diagrama de puntos de MPG versus HP. Piensa Ud. que hay una tendencia lineal.
b) Ajustar la linea de regresión de MPG versus HP. Trazar la linea sobre el plot hallado en a). Comentar los coeficientes de regresión. Interpretar los "p-values" de la prueba t y el de la prueba F.
c) Interpretar el Coeficiente de determinación R^2
d) Obtener un intervalo de confianza del 99% para β. Interpretar su resultado
e) Obtener un intervalo de confianza del 90% para el valor predicho y un intervalo de confianza para el valor medio de MPG cuando HP=100.
f) Hacer un diagrama de puntos de MPG versus WT. Piensa Ud. que hay una tendencia lineal.
g) Ajustar la linea de regresión de MPG versus WT. Trazar la linea sobre el plot hallado en a). Comentar los coeficientes de regresión. Interpretar los "p-values" de la prueba t y el de la prueba F.
h) Interpretar el Coeficiente de determinación R^2
i) Obtener un intervalo de confianza del 99% para β. Interpretar su resultado
j) Obtener un intervalo de confianza del 90% para el valor predicho y un intervalo de confianza para el valor omedio de MPG cuando WT=35.

18.
a) Si Y=3.5-1.5X , SST=219 y SSE=59, hallar e interpretar el valor de la correlación entre X y Y
b) Considerando los datos dados en a) y que la muestra de entrenamiento consiste de 36 datos, hallar el valor de la prueba estadística para probar que la pendiente de regresión es cero.

19.Considere que (X,Y) tiene una distribución normal bivariada con parámetros μ_x, μ_y, σ_x , σ_y y el coeficiente de correlación ρ. Probar que

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

b)La varianza condicional de las Y dado X, está dado por $\sigma^2_{y/x}=\sigma^2_y(1-\rho^2)$. Luego, si $\rho=\pm1$, entonces $\sigma^2_{y/x}=0$, implicando que hay una perfecta relación lineal entre Y y X. Más especificamente, si $\rho=1$, entonces X y Y crecen en la misma dirección y si $\rho=-1$, Y decrece cuando X crece.