Capítulo

4

Análisis de Regresión.

4.1. Etimología

El término *regresión* fue introducido por *Francis Galton* en su libro *Natural inheritance (1889)*. En un famoso artículo Galton Planteó que, a pesar de la presencia de una tendencia en la que los padres de estatura alta tenían hijos altos y los padres de estatura baja tenían hijos bajos, la estatura promedio de de los niños nacidos de padres de una estatura dada tendía a moverse o regresar"hacia la estatura promedio de la población total. En otras palabras, la estatura de los hijos inusualmente altos o de padres inusualmente bajos tendía a moverse hacia la estatura promedio de la población. La Ley de Regresión Universal de Galton fue confirmada por su amigo *Karl Pearson*, quién reunió más de mil registros de estaturas de miembros de grupos familiares. Pearson encontró que la estatura promedio de los hijos de un grupo de padres de estatura alta era menor que la estatura de sus padres y la estatura promedio de los hijos de un grupo de padres de estatura baja era mayor que la estatura de sus padres, generándose así un fenómeno mediante el cual los hijos altos e hijos bajos, regresaban"por igual hacia la estatura promedio de todos los hombres. En palabras de Galton, se trataba de una regresión hacia la mediocridad".

4.2. Análisis de Regresión

Definición 4.1 (Análisis de Regresión) *Es una técnica estadística que se usa para investigar y modelar la relación entre variables.*

Existen dos tipos de variables en un análisis de regresión.

Variable Respuesta. Como su nombre lo indica es la característica que se obtiene como respuesta en la realización de un experimento. También se conoce como variable predicha, regresada o endogena y se representa, por lo general, con la letra Y.

Variable Explicativa. Es la variable o conjunto de variables que ïnfluyen"sobre la variable respuesta. También se conoce como variable independiente, predictoras, regresoras o exogenas y se representan, por lo general, con la letra X en caso de ser una y X_i cuando son varias.

Port lo tanto, el análisis de regresión estudia la dependencia de las variables explicativas $(X_1, X_2, ...)$ sobre la variable respuesta Y

4.3. Ejemplos

1. Según la teoría de Keynes el gasto de consumo de una persona esta relacionado con su ingreso. Por lo tanto, dado el consumo y el ingreso de un grupo de personas, se puede determinar el modelo matemático que explica el comportamiento entre estas variables, y así poder determinar cuanto del ingreso de una persona destina para el consumo o estimar el cambio promedio en el gasto ante una variación de una unidad monetaria en su ingreso.

2. En la mayoría de los estudios de mercado es fundamental conocer la demanda asociada con el producto o servicio que presta una empresa. Dicha demanda puede estar representada por las personas de una población. Si la empresa desea conocer el comportamiento de la demanda en los próximos 5 años será necesario modelar el crecimiento de la población en ese período de tiempo. Por lo tanto, se debe establecer la relación existente entre el crecimiento de la población y el tiempo.

3. Bajo ciertas condiciones ideales, el tiempo que una persona dedica al estudio influye en sus notas. El análisis de regresión le ayudaría a un estudiante estimar cuanta nota podría obtener por cada hora de estudio.

4.4. Modelos de Regresión

El objetivo del Análisis de Regresión es obtener un modelo matemático que explique la relación existente entre el conjunto de variables independientes y la variable dependiente, es decir, si Y es la variable dependiente y $X_1, X_2, ..., X_n$ son las variables independientes entonces se debe encontrar una función f tal que

$$Y = f(X_1, ..., X_n)$$

A dicha función se le conoce como **Modelo de Regresión.**.

Los modelos de regresión se clasifican de acuerdo al número de variables independientes y de acuerdo a la forma de f, de la siguiente manera:

1. De acuerdo al número de variables independientes

   a) Si hay una variable independiente entonces es conocido como un **Modelo de Regresión Simple.**

b) Cuando hay más de una variable independiente se llama **Modelo de Regresión Multiple.**

2. De acuerdo a la forma de f.

   a) Si f es lineal en sus variable independientes, entonces es un **Modelo de Regresión Lineal.**

   b) Si f no es lineal en sus variable independientes, entonces es un **Modelo de Regresión No Lineal.**

Esta clasificación no es excluyente entre sí, por el contrario siempre se complementan. Por ejemplo, el caso mas sencillo del Análisis de Regresión es cuando se tiene una variable independiente y f es lineal en esa variable, se dice que es un **Modelo de Regresión Lineal Simple.**

Existen otros modelos muy particulares, entre los cuales podemos mencionar:

1. Modelo de Regresión Logística.
2. Modelo de Regresión Probit.
3. Modelo de Regresión de Cox.

4.5. Propósitos del Análisis de Regresión

Los modelos de regresión son usados con diversos propósitos, entre los cuales se tienen:

1. Describir datos.
2. Estimar parámetros.

3. Realizar pronóstico.

A continuación se analizan cada uno de ellos:

1. Describir datos. Cuando se realiza un Análisis de Regresión se obtienen ecuaciones que resumen o describen un conjunto de datos, tales ecuaciones como se vio antes son conocidas como Modelos de Regresión. Por ejemplo al estudiar el comportamiento del consumo con respecto al ingreso, dichos modelos pueden probablemente ser más convenientes de usar que una tabla o un gráfico.

2. Estimar parámetros. Los modelos de regresión pueden usarse en algunos casos para estimar parámetros. Por ejemplo, se registro la velocidad de un carro en diversos instantes de tiempo, dichos datos se graficaron en el plano y se observo que la velocidad y el tiempo están relacionados por una línea recta que pasa por el origen con pendiente a, donde a es la aceleración del carro. Por lo tanto, el Análisis de Regresión puede usarse para ajustar este modelo a los datos, produciendo así una estimación de la aceleración.

3. Realizar pronóstico. La mayoría de las investigaciones usan el Análisis de Regresión para pronosticar la variable respuesta para un valor de la variable o las variables independientes. Por ejemplo, el gerente de una fábrica puede pronosticar usando un Modelo de Regresión la s ventas en los próximos años, lo cual le permitirá planificar la producción en ese período de tiempo.

Capítulo

5

Análisis de Regresión Lineal Simple.

5.1. Introducción

En la práctica existen muy pocos casos en los que la variable respuesta esta influenciada por una variable independiente, por lo tanto el análisis de regresión lineal simple es de poca utilidad. Sin embargo es el más indicado como punto de partida en el estudio del Análisis de Regresión, ya que es el más sencillo de tratar algebraicamente. En este capítulo se va desarrollar todo lo referente al Análisis de Regresión Lineal Simple.

5.2. Modelo de Regresión Lineal Simple

En este capítulo se considera el Modelo de Regresión Lineal Simple, esto es, un modelo con una sola variable independiente X relacionada con una variable respuesta Y a través de una línea recta. El Modelo de Regresión Lineal Simple es

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{5.1}$$

donde el intercepto β_0 y la pendiente β_1 son constantes desconocidas y son los parámetros del modelo. ϵ es un componente de error aleatorio.Dichos errores en principio se asumen que tienen media cero, varianza desconocida σ^2 y que están descorrelacionados entre sí (esto significa que el valor de un error no depende del valor de otro error.

Es conveniente observar a la variable independiente X como controlada por el analista y medida sin error, es decir, que X no es una variable aleatoria. Por el contrario, para el mismo valor de la variable independiente X pueden haber distintos valores de Y. Por ejemplo, si de un grupo de personas se registran la edad y la estatura y se considera como variable independiente X la edad y como variable respuesta Y la estatura, se tiene que para una misma edad existen personas con diferentes estaturas. Por lo tanto, Y es una variable aleatoria y en consecuencia tiene una distribución de probabilidad para cada cada valor de X. La media de esta distribución es

$$E[Y/X] = \beta_0 + \beta_1 X \tag{5.2}$$

y la varianza es

$$Var[Y/X] = \sigma^2 \tag{5.3}$$

Así la media de Y es una función lineal de X y es conocida como la **Ecuación de regresión lineal** o simplemente **Recta de regresión**. La varianza de Y no depende del valor de X. Además, ya que los errores están descorrelacionados, las respuestas están descorrelacionadas.

Los parámetros β_0 y β_1 son usualmente llamados **Coeficientes de Regresión**y tienen la siguiente interpretación:

β_1: es el cambio en la media de la distribución de Y producido por un cambio en una unidad en X.

β_0: Es la media de la distribución de la respuesta Y cuando $X = 0$. Si el rango de X no incluye al cero, entonces β_0 no tiene interpretación práctica.

Cuando decimos que un modelo es lineal o no lineal, nos estamos refiriendo a la linealidad o no linealidad en los parámetros. El valor de la potencia más alta de una variable independiente en el modelo es llamado el orden del modelo. Por ejemplo,

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$$

es un modelo de regresión lineal (en los β) de segundo orden (en X). Al menos un modelo es llamado específicamente no lineal este puede ser tomado que es lineal en los parámetros, y la palabra lineal es usualmente omitida y olvidada. El orden del modelo puede ser de cualquier tamaño.

5.3. Estimadores Mínimos Cuadrados de los Parámetros

5.3.1. Estimación de β_0 y β_1.

En la sección anterior dijimos que los parámetros β_0 y β_1 son desconocidos, por lo tanto deben ser estimados a partir de una muestra. Un método para estimar dichos parámetros es el de Mínimos Cuadrados.

El método de los Minimos Cuadrados estima los valores de β_0 y β_1 de manera que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea

mínima. Supongase que tenemos n pares de datos, digamos, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Estos datos pudieron ser resultado de un experimento previamente controlado o de registros históricos. Entonces por la ecuación (7.1) podemos escribir

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{5.4}$$

por lo que la suma de cuadrados de las desviaciones con respecto a la recta verdadera es

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{5.5}$$

Seleccionamos nuestras estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ como aquellos valores, que al sustituirlos por β_0 y β_1 en la ecuación (5.5), produce el menor valor posible de S. Note que x_i y y_i son los valores que hemos observado. Podemos determinar $\hat{\beta}_0$ y $\hat{\beta}_1$ al diferenciar la ecuación (5.5) primero con respecto a β_0 y luego con respecto a β_1 e igualando los resultados a cero. Ahora,

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \tag{5.6}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \tag{5.7}$$

por lo que los estimadores de β_0 y β_1, digamos $\hat{\beta}_0$ y $\hat{\beta}_1$ satisfacen que

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{5.8}$$

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \tag{5.9}$$

donde hemos sustituido $(\hat{\beta}_0, \hat{\beta}_1)$ por (β_0, β_1), cuando igualamos las ecuaciones (5.6) y (5.7) igual a cero. De (5.8) y (5.9) tenemos

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_0 \sum_{i=1}^{n} x_i = 0 \tag{5.10}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{5.11}$$

o

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{5.12}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{5.13}$$

Estas ecuaciones son llamadas las *ecuaciones normales*.

La solución de las ecuaciones (5.12) y (5.13) para $\hat{\beta}_1$ es

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{5.14}$$

La solución de las ecuaciones (5.12) y (5.13) para $\hat{\beta}_0$ es

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{5.15}$$

Ya que el denominador de (5.14) es la suma de cuadrados corregida de los x_i y el numerador es la suma corregida de productos cruzados de x_i y y_i, podemos escribir estas cantidades en una notación más compacta como

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.16}$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) \tag{5.17}$$

Por lo tanto, una manera conveniente de escribir (5.14) es

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{5.18}$$

5.3.2. Estimación de σ^2.

Además de estimar β_0 y β_1, un estimador de σ^2 es necesario para hacer inferencias referentes al modelo de regresión. Idealmente podría pensarse que esta estimación no depende de la adecuacidad del modelo ajustado. Esto es solamente posible cuando hay varias observaciones sobre y para al menos un valor de x o cuando información a priori sobre σ^2 es posible. Cuando este método no puede usarse, el estimador de σ^2 se obtiene a partir de la suma de cuadrados del error,

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5.19}$$

cuya formula de mayor facilidad para calcularla es

$$SS_E = S_{yy} - \beta_1 S_{xy} \tag{5.20}$$

Dicha suma de cuadrados tiene $n-2$ grados de libertad, ya que dos grados de libertad están asociados con las estimaciones β_0 y β_1 envueltas para obtener $\hat{y}_i$. Ahora el valor esperado de SS_E es

$$E(SS_E) = (n-2)\sigma^2 \tag{5.21}$$

por lo tanto, un estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = MS_E \tag{5.22}$$

La cantidad MS_E es llamada el cuadrado medio del error. La raíz cuadrada de $\hat{\sigma}^2$ es algunas veces llamado el **error estándar de regresión**, y tiene las mismas unidades que la variable respuesta. Ya que depende la suma de cuadrados del error, cualquier

violación de los supuestos sobre el error en el modelo puede afectar seriamente el uso de $\hat{\sigma}^2$ como un estimador de σ^2.

5.3.3. Propiedades de los Estimadores

Los estimadores mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen diversas propiedades estadísticas importantes, las cuales se describen a continuación.

1. Son combinaciones lineales de las observaciones y_i.

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \tag{5.23}$$

   donde $c_i = \frac{x_i - \bar{x}}{S_{xx}}$ para $i = 1, 2, ..., n$

2. Son estimadores insesgados, es decir, $E(\hat{\beta}_1) = \beta_1$ y $E(\hat{\beta}_0) = \beta_0$

3. Las varianzas están dadas por

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \tag{5.24}$$

$$Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}) \tag{5.25}$$

4. Son insesgados y de mínima varianza, es decir son los mejores estimadores insesgados.

5. La suma de los residuales en cualquier modelo de regresión que contiene un intercepto β_0 es siempre cero. Es decir,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \tag{5.26}$$

6. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\hat{y}_i$, o

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \tag{5.27}$$

7. La suma de los residuos ponderados por los correspondientes valores de la variable regresora es siempre cero; esto es,

$$\sum_{i=1}^{n} x_i e_i = 0 \tag{5.28}$$

8. La suma de los residuos ponderados por los correspondientes valores ajustados es siempre cero; esto es,

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0 \tag{5.29}$$

5.4. Precisión de la Regresión Estimada

Ahora abordaremos la cuestión de buscar una medida sobre la precisión de nuestra estimación de la linea de regresión. Considere la siguiente identidad:

$$Y_i - \hat{Y}_i = Y_i - \overline{Y} + \overline{Y} - \hat{Y}_i = Y_i - \overline{Y} - (\hat{Y}_i - \overline{Y}) \tag{5.30}$$

si elevamos al cuadrado ambos lados y sumamos desde $i = 1$ hasta n, obtenemos

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 &= \sum_{i=1}^{n}[(Y_i-\overline{Y})-(\hat{Y}_i-\overline{Y})]^2\\ &= \sum_{i=1}^{n}[(Y_i-\overline{Y})^2+(\hat{Y}_i-\overline{Y})^2-2(Y_i-\overline{Y})(\hat{Y}_i-\overline{Y})]\\ &= \sum_{i=1}^{n}(Y_i-\overline{Y})^2+\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2-2\sum_{i=1}^{n}(Y_i-\overline{Y})(\hat{Y}_i-\overline{Y})\end{aligned}$$

El tercer termino puede escribirse como

$$\begin{aligned}-2\sum_{i=1}^{n}(Y_i-\overline{Y})(\hat{Y}_i-\overline{Y}) &= -2\sum_{i=1}^{n}(Y_i-\overline{Y})b_1(X_i-\overline{X})\\ &= -2b_1\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{X})\\ &= -2b_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2\\ &= -2\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\end{aligned}$$

Así

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2-\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2 \tag{5.31}$$

La ecuación (5.31) puede reescribirse como

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \tag{5.32}$$

Ahora $Y_i-\overline{Y}$ es la desviación de la i-esima observación de la media general y por lo tanto el lado izquierdo de la ecuación (5.32) es la suma de cuadrados de las desviaciones

de las observaciones con respecto a la media; es también conocida como la suma de cuadrados corregida de las observaciones (S_{yy}). Los dos componentes de S_{yy} miden, respectivamente, la cantidad de variabilidad en las observaciones y_i descrita por la recta de regresión y la variación residual no explicada por la recta de regresión. En palabras la ecuación anterior se lee como

Suma de cuadrados	=	Suma de cuadrados	+	Suma de cuadrados
sobre la media		de regresión		del error

y en símbolos se representa como

$$S_{yy} = SS_R + SS_E \tag{5.33}$$

Esto muestra que, de la variación de los $Y's$ sobre su media, algunas de las variaciones puede atribuirse a la linea de regresión y algunas, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, al hecho de que las observaciones actuales no todas caen en la recta de regresión (si todas cayeran, la suma de cuadrados del error debía ser cero). De este procedimiento podemos ver que una manera de evaluar que la recta de regresión sea usada como un predictor es ver cuánto de la SS sobre la media ha caído en la SS de la regresión y cuánto en el SS del error. Estaremos contentos si el SS de la regresión es mucho mas grande que la SS del error, o lo que es lo mismo si la razón $R^2 = \frac{\text{SS de la regresión}}{\text{SS sobre la media}}$ no está demasiado lejos de la unidad. R^2 se conoce con el nombre de *Coeficiente de determinación.*

El coeficiente de determinación calcula la proporción de la variación explicada por la variable x. Como $0 \leq SS_E \leq Syy$, se sigue que $0 \leq R^2 \leq 1$. Valores cercanos a 1 implican que gran parte de la variabilidad de Y es explicada por el modelo de regresión.

El estadístico R^2 debe usarse con precaución, ya que es siempre posible hacer R^2 grande al adicionar términos al modelo. Por ejemplo, si no hay puntos repetidos (mas de un valor de y para el mismo valor de x), un polinomio de grado $n-1$ dará un

ajuste "perfecto"($R^2 = 1$) para n puntos. Cuando hay puntos repetidos, R^2 nunca será exactamente igual a 1 debido a que el modelo no puede explicar la variabilidad debida al error "puro".

Aunque el R^2 aumente si adicionamos una nueva variable al modelo, esto no necesariamente significa que el modelo es superior al anterior. A menos que la suma de cuadrados del error en el modelo nuevo se reduzca en una cantidad igual a la suma de cuadrados del error original, el nuevo modelo tendrá un cuadrado medio del error más grande que el anterior debido a la perdida de un grado de libertad del error. Así el modelo nuevo será peor que el anterior.

La magnitud de R^2 también depende del rango de variabilidad de la variable independiente. Generalmente R^2 aumenta a medida que la dispersion de los $x's$ se incrementa y disminuye a medida que la dispersion de los $x's$ se decrementa probado el supuesto de que la forma del modelo es correcta. Hahn (1973) observó que el valor esperados de R^2 de una regresión de linea recta es aproximadamente

$$E(R^2) \simeq \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\beta}_1^2 S_{xx} + \sigma^2}$$

Claramente el valor esperado de R^2 se incrementará (decrementará) a medida de que S_{xx} (medida de dispersion de los $x's$) se incremente (decremente). Así un valor grande de R^2 puede resultar simplemente porque x ha variado sobre un rango irrealmente demasiado grande. Por otro lado, R^2 puede ser pequeño debido a que el rango de x también era pequeño como para permitir que su relación con y sea detectada.

En general, R^2 no mide la magnitud de la pendiente de la recta de regresión. Un valor grande de R^2 no implica una pendiente empinada. Además R^2 no mide si el modelo lineal es apropiado, se puede tener un R^2 alto aunque las variables no tengan

una relación lineal.

Cualquier suma de cuadrados tiene asociado a él un numero llamado sus grados de libertad. Este número indica cuantas piezas de información independientes envuelven los n numeros independientes $Y_1, Y_2, ..., Y_n$ son necesarias para calcular las sumas de cuadrados. Por ejemplo, la SS sobre la media necesita $(n-1)$ piezas independientes (de los numeros $Y_1 - \bar{Y}, Y_2 - \bar{Y}, ..., Y_n - \bar{Y}$ solamente $(n-1)$ son independientes ya que todos los n numeros deben sumar cero por la definición de la media). Podemos calcular la SS de la regresión desde una función de $Y_1, Y_2, ..., Y_n$ llamemosla b_1 (ya que $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = b_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$, y por lo tanto esta suma de cuadrados tiene un grado de libertad. por sustracción, la SS sobre la regresión tiene $(n-2)$ grados de libertad. Por lo tanto, la ecuación (5.32), muestra la relación de los grados de libertad como

$$(n-1) = (n-2) + 1 \tag{5.34}$$

Usando las ecuaciones (5.32) y (5.34) y empleando formas de cálculos alternativos para las expresiones de la ecuación (5.32) podemos construir una tabla de *análisis de varianza* en la siguiente forma:

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
Regresión	$\hat{\beta}_1 S_{xy}$	1	MS_R
Error	$S_{yy} - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E
Total	S_{yy}	$n-1$	

La columna cuadrado medio es obtenida al dividir cada suma de cuadrados entre sus correspondientes grados de libertad.

5.5. Examinando la Ecuación de Regresión

Hasta este punto no hemos hecho ningún supuesto sobre la distribución de probabilidad de los elementos del modelo. Un numero especifico de cálculos algebraicos han sido realizados y eso es todo. Ahora haremos los supuestos básicos sobre el modelo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, 2, ..., n.$

1. ϵ_i es una variable aleatoria con media cero y varianza σ^2 (desconocida), esto es, $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$.

2. ϵ_i y ϵ_j están descorrelacionadas, $i \neq j$, esto es

$$Cov(\epsilon_i, \epsilon_j) = 0$$

   Por lo tanto,

$$E(Y_i) = \beta_0 + \beta_1 X_i \quad V(Y_i) = \sigma^2$$

   y Y_i y Y_j, $i \neq j$, están descorrelacionadas.

3. ϵ_i es una variable aleatoria distribuida normal, con media cero y varianza σ^2, es decir,

$$\epsilon_i \sim N(0, \sigma^2)$$

   Bajo este último supuesto ϵ_i y ϵ_j no sólo están descorrelacionadas sino que además son independientes.

Ahora usaremos estos supuestos para examinar la ecuación de regresión. Dicho examen consiste en realizar inferencias estadísticas sobre los parámetros del modelo.

5.5.1. Inferencias sobre la pendiente, β_1

Vimos que el estimador para β_1 obtenido por el método de los mínimos cuadrados estaba dado por

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{S_{yx}}{S_{xx}}$$

y cuya varianza es

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Como el error estándar es la raíz cuadrada de la varianza, se tiene que

$$ee(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$$

o, si σ es desconocido usamos su estimador $\hat{\sigma}$ en su lugar, asumiendo que el modelo es correcto, el error estándar estimado de $\hat{\beta}_1$ está dado por

$$\widehat{ee(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \tag{5.35}$$

Prueba de hipótesis para β_1

Supongase que se desea probar la hipótesis de que la pendiente es igual a una constante, es decir de que β_1 es igual a β_{10}. Las hipótesis apropiadas son

$$H_0 \quad : \quad \beta_1 = \beta_{10}$$
$$H_1 \quad : \quad \beta_1 \neq \beta_{10}$$

donde se ha especificado una alternativa de dos colas (podría ser de una cola, el procedimiento no cambia). Ya que los errores $\epsilon_i \sim NID(0, \sigma^2)$, las observaciones $y_i \sim NID(\beta_0 + \beta_1 x_i, \sigma^2)$. Como $\hat{\beta}_1$ es una combinación lineal de la observaciones, entonces $\hat{\beta}_1 \sim NID(\beta_1, \frac{\sigma^2}{S_{xx}})$. Por lo tanto el estadístico

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \tag{5.36}$$

se distribuye $N(0, 1)$ si la hipótesis nula $H_0 : \beta_1 = \beta_{10}$ es cierta. Como σ^2 es por lo general desconocido debemos usar su estimador $\hat{\sigma}^2 = MS_E$. Y usando el hecho de que $\frac{(n-2)MS_E}{\sigma^2} \sim \chi^2_(n-2)$ y MS_E y β_1 son variables aleatorias independientes, se tiene que el estadístico

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_E}{S_{xx}}}} \tag{5.37}$$

se distribuye $t(n-2)$ si la hipótesis nula $H_0 : \beta_1 = \beta_{10}$ es cierta. El estadístico t_0 es usado para probar $H_0 : \beta_1 = \beta_{10}$ al comparar el valor de t_0 de la ecuación (5.37) con su valor crítico, el cual depende de la desigualdad de la hipótesis alternativa planteada.

Un caso de especial importancia de (5.36) es

$$H_0 \quad : \quad \beta_1 = 0$$

$$H_1 \quad : \quad \beta_1 \neq 0$$

Esta hipótesis se refiere a lo que es la **significancia de regresión**. Fallar en rechazar $H_0 : \beta_1 = 0$ implica que no hay relación lineal entre x y y.

Note que esto puede implicar que x no es suficiente para explicar la variación de y y que el mejor estimador de y es $\bar{y}$ o que la verdadera relación entre x y y no es lineal.

Alternativamente si $H_0 : \beta_1 = 0$ es rechazada, esto implica que x es de valor para explicar la variabilidad en y. Sin embargo, rechazar $H_0 : \beta_1 = 0$ puede significar que el modelo de linea recta es adecuado o que aunque hay un efecto lineal de x, mejores resultados pueden obternerse con la adición de polinomios de orden más altos en terminos de x.

El procedimiento de prueba para $H_0 : \beta_1 = 0$ puede desarrollarse de dos maneras.

1. Usando el estadístico t_o antes descrito.

2. Usando la tabla de análisis de varianza desarrollada en la sección **** . En este caso el estadístico de prueba es

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \tag{5.38}$$

Los valores esperados de estos cuadrados medios son

$$E(MS_E) = \sigma^2 \tag{5.39}$$

$$E(MS_R) = \sigma^2 + \beta_1^2 S_{xx} \tag{5.40}$$

Como MS_R y MS_E son variables aleatorias independientes, entonces si la hipótesis nula $H_0 : \beta_1 = 0$ es cierta, el estadístico F_0 sigue una distribución $F_{1,n-2}$. Los valores esperados de los cuadrados medios indican que si el valor observado de F_0 es grande, entonces es equivalente a que la pendiente $\beta_1 \neq 0$. Por lo tanto, para probar la hipótesis $H_0 : \beta_1 = 0$, se calcula el estadístico F_0 y se rechaza H_0 si

$$F_0 > F_{\alpha,1,n-2} \tag{5.41}$$

El procedimiento se resume en la tabla ***

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_0
Regresión	$\hat{\beta}_1 S_{xy}$	1	MS_R	$\frac{MS_R}{MS_E}$
Error	$S_{yy} - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	S_{yy}	$n-1$		

Intervalo de confianza para β_1

Bajo el supuesto (3), un estimador por intervalo con un nivel de confianza $1-\alpha$ para β_1 está dado por

$$\hat{\beta}_1 \pm t_{(n-2,1-\frac{\alpha}{2})} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \tag{5.42}$$

5.5.2. Inferencias sobre el intercepto, β_0

Prueba de hipótesis para β_0

Un procedimiento similar puede usarse para probar hipótesis sobre el intercepto. Para probar

$$\begin{aligned} H_0 &: \beta_0 = \beta_{00} \\ H_1 &: \beta_0 \neq \beta_{00} \end{aligned}$$

Debemos usar el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_E(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \tag{5.43}$$

y rechazar la hipótesis nula al comparar el valor de t_0 con el valor crítico obtenido de la distribución t-student.

Intervalo de confianza para β_0

Bajo el supuesto (3), un estimador por intervalo con un nivel de confianza $1 - \alpha$ para β_0 está dado por

$$\hat{\beta}_0 \pm t_{(n-2,1-\frac{\alpha}{2})}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{5.44}$$

5.5.3. Intervalo de confianza para σ^2

Si los errores son normales e independientemente distribuidos, la distribución muestral de

$$\frac{(n-2)MS_E}{\sigma^2} \tag{5.45}$$

es χ^2_{n-2}. Así,

$$P\left\{\chi^2_{1-\frac{\alpha}{2},n-2} \leq \frac{(n-2)MS_E}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-2}\right\} = 1-\alpha \tag{5.46}$$

En consecuencia un intervalo de confianza de $100(1-\alpha)\,\%$ para σ^2 es

$$\frac{(n-2)MS_E}{\chi^2_{\frac{\alpha}{2},n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_E}{\chi^2_{1-\frac{\alpha}{2},n-2}} \tag{5.47}$$

5.6. Predicción

Un importante uso de los modelos de regresión es estimar tanto la respuesta media $E(y)$ como la respuesta individual para un valor particular de la variable independiente x.

5.6.1. Predicción Media

Sea x_0 un valor de la variable independiente para el cual se desea estimar la respuesta media, es decir $E(y/x_0)$. Asumamos que x_0 es cualquier valor de la variable independiente dentro del rango de valores de x usados para ajustar el modelo. Un estimador puntual insesgado de $E(y/x_0)$ es obtenido del modelo ajustado como

$$\widehat{E(y/x_0)} \equiv \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{5.48}$$

Para obtener un intervalo de confianza de $(1-\alpha)\,\%$ para $E(y/x_0)$, primero note que $\hat{y}_0$ es una variable aleatoria que se distribuye normal ya que esta es una combinación

lineal de las observaciones y_i. La varianza de $\hat{y}_0$ es

$$\begin{aligned} V(\hat{y}_0) &= V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = V[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

ya que $Cov(\bar{y}, \hat{\beta}_1) = 0$. Así la distribución muestral de

$$\frac{\hat{y}_0 - E(y/x_0)}{\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \tag{5.49}$$

es t_{n-2}. En consecuencia un intervalo de confianza de $(1-\alpha)\,\%$ para $E(y/x_0)$ es

$$\hat{y_0} \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \tag{5.50}$$

Note que el ancho del intervalo es una función de x_0. La amplitud del intervalo es minima para $x_0 = \bar{x}$ y se ensancha a medida que $|x_0 - \bar{x}|$ se incrementa. Intuitivamente esto es razonable, debe esperarse mejores estimaciones de y hechas con valores de x cercanos al centro de los datos que con valores en los bordes.

5.6.2. Predicción Individual

Una importante aplicación del modelo de regresión es la predicción de nuevas observaciones y correspondientes a un nivel especifico de la variable independiente x. Si x_0 es el valor de la variable independiente de interés, entonces

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{5.51}$$

es la estimación puntual del nuevo valor de la respuesta y_0.

Ahora consideremos obtener un intervalo de confianza de $(1-\alpha)\,\%$ de esta observación futura y_0. El intervalo de confianza sobre la respuesta media en $x = x_0$, ecuación (5.50), es inapropiado para este problema ya que esta es una estimación por intervalo sobre la media de y (un parámetro), no una afirmación probabilística acerca de observaciones futuras de esa distribución. Desarrollaremos un intervalo de predicción para futuras observaciones y_0. Note que la variable aleatoria

$$\psi = y_0 - \hat{y}_0$$

esta normalmente distribuida con media cero y varianza

$$\begin{aligned} V(\psi) &= V(y_0 - \hat{y}_0) \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

ya que la observación futura y_0 es independiente de $\hat{y}_0$. Si usamos $\hat{y}_0$ para predecir y_0, entonces el error estándar de $\psi = y_0 - \hat{y}_0$ es parte del estadístico apropiado para una predicción por intervalo. Así el intervalo de predicción de $100(1-\alpha)$ porciento sobre una observación futura en x_0 es

$$\hat{y_0} \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \tag{5.52}$$

La amplitud del intervalo de predicción es mínima cuando $x_0 = \bar{x}$ y se ensancha a medida de que $|x_0 - \bar{x}|$ se incrementa. Al comparar (5.50)

5.7. Correlación entre X e Y.

Si X y Y son ambas variables aleatorias que siguen una distribución de probabilidad bivariada (desconocida), entonces podemos definir el coeficiente de correlación entre X y Y como

$$\rho_{XY} = \frac{Cov(X,Y)}{[Var(X)Var(Y)]^{\frac{1}{2}}} \tag{5.53}$$

Puede demostrarse que $-1 \leq \rho_{XY} \leq 1$. La cantidad ρ_{XY} es una medida de la asociación entre las variables aleatorias X y Y. Por ejemplo,

- si $\rho_{XY} = 1$, X y Y están correlacionadas positivamente perfectamente y todos los posibles valores de X y Y caen en una línea recta con pendiente positiva en el plano (X, Y).

- si $\rho_{XY} = 0$, las variables se dicen estar descorrelacionadas, es decir, no están asociadas linealmente una de la otra. Esto no significa que X y Y sean estadísticamente independientes.

- si $\rho_{XY} = -1$, X y Y están correlacionadas negativamente perfectamente y todos los posibles valores de X y Y caen en una línea recta con pendiente negativa en el plano (X, Y).

El estimador de ρ es el coeficiente de correlación muestral dado por

$$r = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} = \frac{S_{xy}}{[S_{xx}S_{yy}]^{\frac{1}{2}}} \tag{5.54}$$

Note que

$$\hat{\beta}_1 = \left(\frac{S_{yy}}{S_{xx}}\right)^{\frac{1}{2}} r \tag{5.55}$$

pues la pendiente $\hat{\beta}_1$ es justamente el coeficiente de correlación muestral r multiplicado por un factor de escala que es la raíz cuadrada de la dispersion de los $y's$ divididos por la dispersion de los $x's$. Así $\hat{\beta}_1$ y r están íntimamente relacionados, aunque ellos proveen información diferente. El coeficiente de correlación muestral es una medida de la asociación entre y y x, mientras que $\hat{\beta}_1$ mide el cambio predicho en y por un cambio en una unidad de x. En el caso de una variable controlable x, r no tiene significado debido a que la magnitud de r depende de la selección del espacio de x. También puede demostrarse que $r^2 = R^2$, es decir, el coeficiente de determinación R^2 es el cuadrado del coeficiente de correlación entre y y x.

Mientras que la regresión y la correlación están íntimamente relacionadas, la regresión es una herramienta más poderosa en muchas situaciones. La correlación es sólo una medida de asociación y es de poco uso en la predicción. Sin embargo, los métodos de regresión son usados en desarrollar relaciones cuantitativas entre variables, la cual puede ser usada en predicción.

A menudo se quiere probar la hipótesis de que el coeficiente de correlación sea igual a cero; esto es,

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

El estadístico de prueba apropiado para esta hipótesis es

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{5.56}$$

la cual sigue una distribución t con $n-2$ grados de libertad si $H_0 : \rho = 0$ es cierta. Por lo tanto se rechaza la hipótesis nula si $|t_0| > t_{\alpha/2,n-2}$. Esta prueba es equivalente a la prueba t para $H_0 : \beta_1 = 0$, esta relación se sigue directamente de (5.55).

5.8. Ejercicios

1. Determine si los siguientes modelos son lineales en los parámetros, o en las variables, o en ambos. ¿Cuáles de estos modelos son de regresión lineal?.

   a) $y_i = \beta_0 + \beta_1 \left(\frac{1}{x_i}\right) + \epsilon_i$

   b) $y_i = \beta_0 + \beta_1 \ln(x_i) + \epsilon_i$

   c) $\ln y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

   d) $\ln y_i = \ln \beta_0 +$
   $beta_1 \ln x_i + \epsilon_i$

   e) $\ln y_i = \beta_0 - \beta_1 \left(\frac{1}{x_i}\right) + \epsilon_i$

2. Los siguientes, ¿son modelos de regresión lineal? ¿Por qué razón?.

   a) $y_i = \exp \beta_0 + \beta_1 x_i + \epsilon_i$

   b) $y_i = \frac{1}{1+\exp \beta_0 + \beta_1 x_i + \epsilon_i}$

   c) $y_i = \beta_0 + (0{,}75 - \beta_0) \exp -\beta_1 (x_i - 2) + \epsilon_i$

   d) $y_i = \beta_0 + \beta_1^3 x_i + \epsilon_i$

3. Probar que $SS_E = \sum_{i=1}^{n} e_i^2$ puede escribirse como $_E = S_{yy} - \beta_1 S_{xy}$

4. Las siguientes variables representan datos sobre el rendimiento del kilometraje de la gasolina de 32 automóviles diferentes.

   a) Ajustar un modelo de regresión lineal simple que relacione el kilometraje de la gasolina (y) con el desplazamiento del motor (pulgadas cubicas) (x_1).

   b) Construir la tabla de análisis de varianza y probar la significancia del modelo.

   c) ¿Qué porcentaje del total de la variabilidad en el kilometraje de la gasolina es explicada por la relación lineal con el desplazamiento del motor?.

   d) Hallar un intervalo del 95 % sobre la media del kilometraje de la gasolina si el desplazamiento del motor es 275 $pulg^3$.

   e) Suponga que se desea predecir el kilometraje de la gasolina de un carro con un motor de 275 $pulg^3$. De una estimación puntual del kilometraje. Hallar un intervalo de predicción del 95 % del kilometraje.

   f) Compare los dos intervalos obtenidos en las partes d y e. Explique la diferencia entre ellos. ¿Cuál es más ancho, y por qué?.

5. Considere los datos del ejercicio anterior. Repita las partes a, b y c usando el ancho del vehículo (x_{10}) como la variable regresora. Basado en una comparación de los dos modelos, se puede concluir que (x_1) es mejor regresor que (x_{10}).

6. El peso y la presión sistólica sanguínea de 26 hombres seleccionados aleatoriamente con edades entre 25 y 30 se muestran a continuación. Asumiendo que el peso y la presión sanguínea (BP) se distribuyen normal conjuntamente.

a) Hallar una recta de regresión lineal que relacione la presión sanguínea con el peso.

b) Estimar el coeficiente de correlación.

c) Probar la hipótesis de que $\rho = 0$.

d) Probar la hipótesis de que $\rho = 0{,}6$.

e) Hallar un intervalo de confianza del 95 % para ρ.

7. Considere el modelo de regresión lineal simple $y = \beta_0 + \beta_1 x + \epsilon$, con $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$, y ϵ descorrelacionados.

   a) Demostrar que $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{barx\sigma^2}{S_{xx}}$

   b) Demostrar que $Cov(\overline{y}_0, \hat{\beta}_1) = 0$

8. Considere el modelo de regresión lineal simple $y = \beta_0 + \beta_1 x + \epsilon$, con $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$, y ϵ descorrelacionados.

   a) Demostrar que $E(MS_E) = \sigma^2$

   b) Demostrar que $E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$

9. Probar que el valor máximo de R^2 es menor que 1 si los datos contienen observaciones repetidas de y en el mismo valor de x.

10. Considere el modelo de regresión lineal simple $y = \beta_0 + \beta_1 x + \epsilon$, donde el intercepto β_0 es conocido.

    a) Hallar el estimador de mínimos cuadrados de β_1 para este modelo. ¿Es la respuesta razonable?.

b) ¿Cuál es la $Var(\hat{\beta}_1)$ para el estimador de minimos cuadrados hallado en la parte a?.

c) Hallar un intervalo de confianza de $100(1-\alpha)\,\%$ para β_1. ¿Es este intervalo más estrecho que el estimador para el caso donde ambos, el intercepto y la pendiente, son desconocidos?.

Capítulo

6

Regresión usando R

6.1. Funciones que se usan en el análisis de regresión

1. **lm()**

   La función lm() determina las estimaciones de los parámetros de un modelo de regresión lineal. La sintaxis más simple de dicha función es:

   lm(formula,data)

   donde

   - *formula* es un objeto que representa el modelo planteado y se representa en la forma *respuesta regresoras* donde respuesta es la variable dependiente o respuesta y regresoras es el conjunto de variables independientes que en el caso de regresión múltiple van separadas por el signo +. Por ejemplo si y es la variable respuesta y x_1 y x_2 son las variables independientes entonces la formula es $y \sim x_1 + x_2$.

   - *data* es el conjunto de datos que se están estudiando.

Se recomienda asignar todos los resultados de la función **lm()** a un objeto. Recuerde que esto se hace simplemente con la siguiente instrucción

M1<-lm(formula,data)

Para imprimir los resultados obtenidos con la función **lm()** se coloca el nombre del objeto, es decir

M1

2. **summary()**

   El argumento de la función summary es un objeto de **lm()**. Esta función despliega información más abundante sobre el análisis de regresión que la impresa directamente por el objeto de la función **lm()**. La información aportada por **summary()** es

   - Algunos Estadísticos descriptivos sobre los residuos,
   - las estimaciones de los parámetros, así como la desviación estándar, el valor del estadístico de prueba para evaluar la significancia de dicho parámetro y su respectivo p-valor.
   - Otros estadísticos que premiten evaluar la bondad del ajuste del modelo tales como: error estándar del residual con su respectivo grados de libertad, el valor del R^2 y el R^2 ajustado, el estadístico F con sus respectivos grados de libertad y p-valor los cuales se usan para determinar la significancia del modelo.

3. **anova()**

   El argumento de la función **anova()** es un objeto de **lm()**. Esta función proporciona el análisis de varianza que se usa para evaluar la significancia del modelo de regresión lineal.

4. **fitted()**

   El argumento de la función **fitted()** es un objeto de **lm()**. Devuelve los valores ajustados por el modelo de regresión lineal.

5. **residuals()**

   El argumento de la función **residuals()** es un objeto de **lm()**. Devuelve los residuales del modelo de regresión lineal.

6. **rstudent()**

   El argumento de la función **rstudent()** es un objeto de **lm()**. Devuelve los residuales estudentizados del modelo de regresión lineal.

7. **rstandard()**

   El argumento de la función **rstandard()** es un objeto de **lm()**. Devuelve los residuales estandarizados del modelo de regresión lineal.

8. **plot()**

   La función **plot()** se usa para obtener los gráficos de los residuos excepto el histograma y el gráfico Q-Q. La sintaxis mas simple es:
   plot(variablex,variabley,xlab="Nombre del eje x",ylab="Nombre del eje y",main="Titulo del gráfico")

9. **hist()**

   La función **hist()** se usa obtener el histograma de los residuales. Tiene la misma sintaxis del **plot()** pero se coloca solo una variable

10. **qqnorm()**

    La función **qqnorm()** se usa para obtener el gráfico de probabilidad normal Q-Q y tiene la misma sintaxis del histograma.

6.2. Ejemplo: Tiempo de Entrega

Este es un ejemplo tomado de Montgomery(2002):Un embotellador de bebidas gaseosas analiza las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Le interesa predecir el tiempo necesario para que el representante de ruta atienda las máquinas expendedoras en una tienda. Esta actividad de servicio consiste en abastecer la máquina con productos embotellados, y algo de mantenimiento o limpieza. El ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan el tiempo de entrega y son la cantidad de cajas de producto abastecido, x_1, y la distancia caminada por el representante, x_2. El ingeniero ha reunido 25 observaciones de tiempo de entrega que se ven en la tabla 7.1. Se ajustará el modelo de regresión lineal simple siguiente

$$y = \beta_0 + \beta x_1 + \varepsilon$$

Para realizar el análisis de regresión usando R, primero su importan y se cargan lo datos usando las siguientes instrucciones

```
> Datos <- read.table("tiempo de entrega.txt",header = TRUE)
> attach(Datos)
```

Estimación de los Parámetros: Para obtener la estimación de los parámetros se usa la función lm() como se muestra en la siguiente instrucción:

```
> MRL1 <- lm(y ~ x1, data = Datos)
```

Con lo cual se obtiene la siguiente tabla
Para obtener resultados mas detallados sobre el análisis de regresión se usa la siguiente instrucción

Tabla 6.1: Datos de tiempo de entrega

Observación	y	x_1	x_2
1	16,68	7	560
2	11,50	3	220
3	12,03	3	340
4	14,88	4	80
5	13,75	6	150
6	18,11	7	330
7	8,00	2	110
8	17,83	7	210
9	79,24	30	1460
10	21,50	5	605
11	40,33	16	688
12	21,00	10	215
13	13,50	4	255
14	19,75	6	462
15	24,00	9	448
16	29,00	10	776
17	15,35	6	200
18	19,00	7	132
19	9,50	3	36
20	35,10	17	770
21	17,90	10	140
22	52,32	26	810
23	18,75	9	450
24	19,83	8	635
25	10,75	4	150

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

Tabla 6.2: Estimacin y Significancia de los Parmetros

```
> summary(MRL1)


Call:

lm(formula = y ~ x1, data = Datos)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.5811 -1.8739 -0.3493  2.1807 10.6342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371   2.422   0.0237 *
x1             2.176      0.124  17.546 8.22e-15 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared: 0.9305,        Adjusted R-squared: 0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

Tabla de Análisis de Varianza: Para obtener la tabla de análisis de varianza se usa la siguiente instrucción

```
> anova(MRL1)
```

Con lo cual se obtiene la tabla 6.3

Tabla 6.3: Anlisis de Varianza

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	5382.41	5382.41	307.85	0.0000
Residuals	23	402.13	17.48		

Análisis de los Residuos

En la tabla 6.4 se muestran los valores observados, ajustados, residuales y residuales estudentizados de los datos en estudio.

Tabla 6.4: Observaciones, valores ajustados, residuales y residuales estudentizados

Observación	y_i	$\hat{y}_i$	e_i	r_i
1	16.68	18.5539	-1.8739	-0.4500
2	11.50	9.8493	1.6507	0.4017
3	12.03	9.8493	2.1807	0.5321
4	14.88	12.0254	2.8546	0.6962
5	13.75	16.3778	-2.6278	-0.6353
6	18.11	18.5539	-0.4439	-0.1062
7	8.00	7.6731	0.3269	0.0797
8	17.83	18.5539	-0.7239	-0.1732
9	79.24	68.6058	10.6342	4.6851
10	21.50	14.2016	7.2984	1.8908
11	40.33	38.1394	2.1906	0.5395
12	21.00	25.0824	-4.0824	-0.9970
13	13.50	12.0254	1.4746	0.3567
14	19.75	16.3778	3.3722	0.8201
15	24.00	22.9063	1.0937	0.2615
16	29.00	25.0824	3.9176	0.9551
17	15.35	16.3778	-1.0278	-0.2466
18	19.00	18.5539	0.4461	0.1067
19	9.50	9.8493	-0.3493	-0.0847
20	35.10	40.3156	-5.2156	-1.3369
21	17.90	25.0824	-7.1824	-1.8436
22	52.32	59.9011	-7.5811	-2.3790
23	18.75	22.9063	-4.1563	-1.0152
24	19.83	20.7301	-0.9001	-0.2152
25	10.75	12.0254	-1.2754	-0.3084

El calculo de los ajustados, residuales y residuales estudentizados se obtuvieron con las siguientes instrucciones respectivamente

```
> fitted(MRL1)
>residuals(MRL1)
>rstudent(MRL1)
```

En la figura 6.2 se muestran diversos gráficos que permiten evaluar los supuestos y la adecuacidad del modelo. Dichos gráficos se obtuvieron con las siguientes instrucciones:

- Para el histograma

```
> hist(residuals(MRL1), main = "", xlab = "Residuales", ylab =
  "Frecuencia")
```

- Para el gráfico Q-Q

```
> qqnorm(rstudent(MRL1), main = "", pch = 19, xlab =
  "Cuantiles Te\'{o}ricos", ylab = "Cuantiles Muestrales")
> abline(0, 1)
```

- Para el gráfico de los residuales vs los valores ajustados

```
> plot(fitted(MRL1), residuals(MRL1), xlab = expression(hat(y)[i]),
  ylab = expression(e[i]))
```

- Para el gráfico de los residuales vs la variable independiente

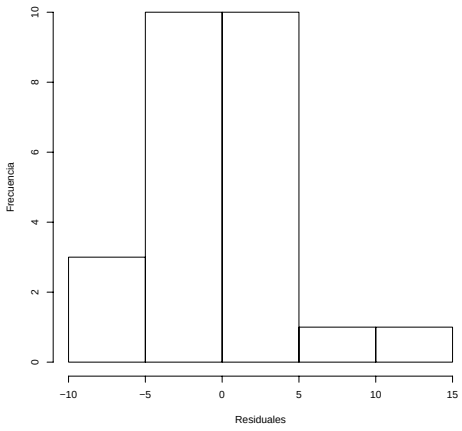
```
> plot(x1, rstudent(MRL1), xlab = "Cajas", ylab = expression(r[i]))
```

- Para el gráfico de e_{i+1} vs e_i

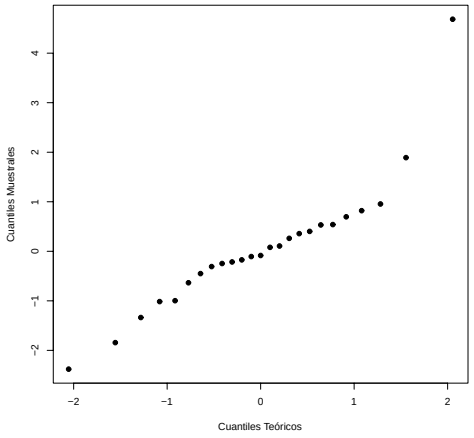
```
> resi <- rstudent(MRL1)
> plot(resi[-25], resi[-1], xlab = expression(e[i]), ylab =
  expression(e[i+1]))
```

- Para el gráfico de los residuales vs tiempo

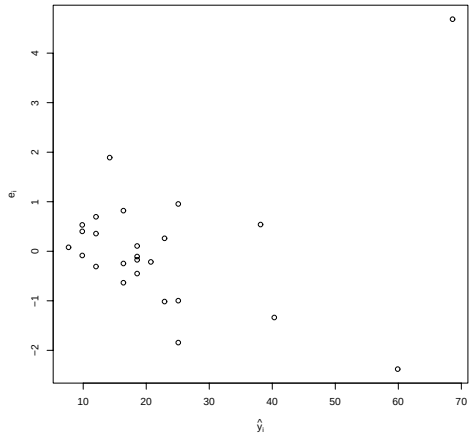
```
> resi <- residuals(MRL1)
> plot(resi, xlab = "Observaci\'{o}n", ylab = expression(e[i]))
```



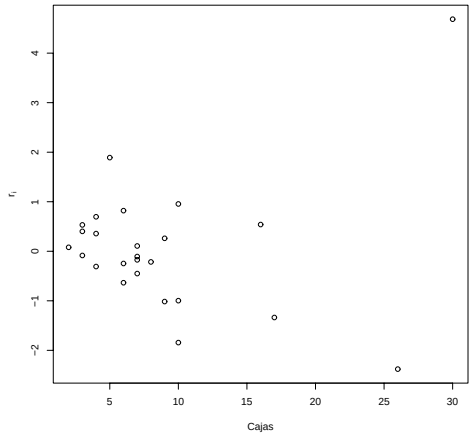
(a) Histograma de residuales



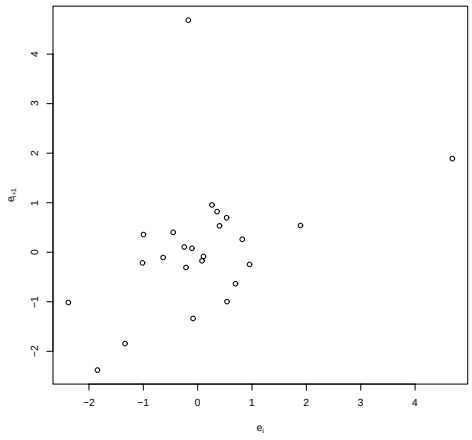
(b) Gráfico Q-Q



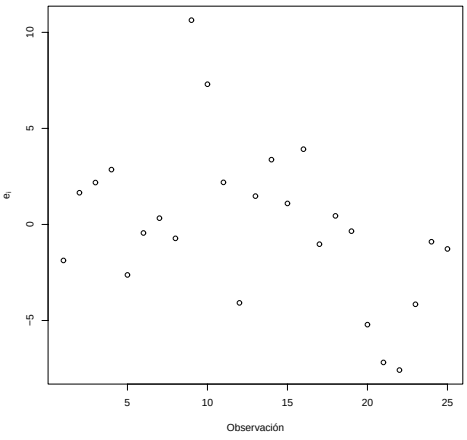
(c) Residuales vs Ajustados



(d) Residuales vs Variable independiente



(e) e_{i+1} vs e_i



(f) Residuales versus tiempo

Figura 6.1: Grficos de residuales