Universidad Nacional del Altiplano

Resumen de regresión lineal simple

Aplicaciones en R.

(PARTE 1)

Docente: Dr. Edgar Carpio Vargas

1. Modelo lineal simple

(Calcina, 2019) El modelo lineal simple (mls) es un modelo uniecuacional (ecuación de una recta) que permite explicar la relación lineal entre dos variables, donde Y es la variable dependiente (explicada, respuesta) y la X es la variable independiente (predictora, explicativa).

El modelo de regresión lineal simple se describe de acuerdo a la ecuación:

Función Lineal. $Y = \beta_0 + \beta_1 X$

β_0: ordenada en el origen (valor de Y cuando X = 0)

β_1: pendiente (cambio de Y al aumentar X en 1)

Modelo de regresión lineal simple: Una relación lineal se expresa como: $E(Y/X) = \beta_0 + \beta_1$, donde la Y observada puede diferir aleatoriamente de E(Y/X) y esta diferencia también es denotada por u_i.

$$Y = \beta_0 + \beta_1 X_i + \mu_i \quad o \quad Y = E(Y/X) + \mu_i \quad y \quad \mu_i = Y - E(Y/X)$$

Modelo matemático o teórico. es la expresión matemática de una determinada teoría. $Y = \beta_0 + \beta_1 X_i$, donde: $\beta_0 \, y \, \beta_1$ son parámetros a encontrar.

Modelo estadístico. Es la expresión matemática de una determinada teoría que incluye en su expresión el término de perturbación o error.

$$Y = \beta_0 + \beta_1 X_i + \mu_i$$

En este modelo no se consideran los errores de observación, pero si se considera el error en la especificación del modelo.

μ_i= término de perturbación o error. es una variable aleatoria (estocástica) con propiedades probabilísticas bien definidas. Es aquí donde están involucradas todos los errores que no han sido consideradas en el modelo, que pueden provenir de (Trujillo, 2017):

- De variables explicativas relevantes que no se consideraron en el modelo.

- De errores de especificación en la relación de correspondencia entre las variables (se considera una relación lineal, cuando en realidad es una función no lineal) (Morocho, 2015)
- De errores de medida sobre las variables endógenas, considerándose dichos errores como aleatorios y se les incorpora en la variable estocástica de cada ecuación de un modelo.

Las relaciones exactas como la indicada no se dan a menudo en la práctica, por cuanto si muestreamos un par de valores (X, Y) no se distribuirán exactamente a lo largo de la recta, ello se debe a que el modelo es teórico y no es más que una simplificación de la realidad. En estadística al realizar tipos de muestreo, expresar un modelo etc. se cometen ciertos errores o perturbadores aleatorios que, para tapar esta brecha, entre la realidad y la práctica, en nuestro modelo introducimos la "u". Entonces el modelo quedara de la siguiente manera:

Si el estudio se está realizando con muestras el modelo quedara especificado como:

$y = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$ (modelo estadístico muestral)

2. Relaciones entre variables

a) Una variable X puede influir en otra variable Y, esto es X→ Y.

Ejemplo: La edad influye sobre la actividad mental del niño, la harina influye en el volumen del pan, la lluvia influye en la cosecha, el peso de un animal vivo influye en el peso de la carcasa, la temperatura influye en la intensidad del ataque de los insectos, etc.

b) Dos variables pueden estar influenciadas entre sí; esto es X ↔ Y.

Ejemplo: Precio y producción de un artículo, peso y volumen del pan, peso y altura de los individuos, nubosidad y horas del sol, uso de tabaco y afecciones cardiacas, etc.

c) Dos variables sin estar influenciadas, pueden estar relacionadas entre sí (concomitantes), por estar ambas influenciadas por una tercera variable.

   Ejemplo: El peso de los hermanos y el peso de las hermanas, el peso del pan y el precio de las papas (influencia del aumento de costo de vida), las notas de química y de bioquímica relacionadas por la afición de los alumnos a los cursos de ciencias. (correlación).

En resumen, la regresión lineal simple consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependiente o respuesta se le identifica como Y y a la variable predictora o independiente como X (Amat, 2016).

El modelo de regresión lineal simple se describe de acuerdo a la ecuación:

$$Y = \beta_0 + \beta_1 X_1 + u_i$$

Modelo muestral:

$$y = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

siendo $\hat{\beta}_0$ la ordenada en el origen, $\hat{\beta}_1$ la pendiente y e_i el error aleatorio. Este último representa la diferencia entre el valor ajustado por la recta y el valor real. Recoge el efecto de todas aquellas variables que influyen en Y pero que no se incluyen en el modelo como predictores. Al error aleatorio también se le conoce como residuo (Amat, 2016).

En la gran mayoría de casos, los valores β_0 y β_1 poblacionales son desconocidos, por lo que, a partir de una muestra, se obtienen sus estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$. Estas estimaciones se conocen como coeficientes de regresión o *least square coefficient estimates*, ya que toman aquellos valores que minimizan la suma de cuadrados residuales, dando lugar a la recta que pasa más cerca de todos los puntos. (Existen alternativas al método de mínimos cuadrados para obtener las estimaciones de los coeficientes).

(existen diferentes ecuaciones para calcular $\hat{\beta}_0$ y $\hat{\beta}_1$, veamos algunas.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_y}{S_x} R$$

$$\hat{\beta}_0 = \bar{Y} = +\hat{\beta}_1 \bar{X}$$

$\hat{\beta}_0$ es el valor esperado de la variable Y cuando $X = 0$, es decir, la intersección de la recta con el eje Y. Es un dato necesario para generar la recta, pero, en ocasiones, no tiene interpretación práctica (situaciones en las que X no puede adquirir el valor 0).

Una recta de regresión puede emplearse para diferentes propósitos y dependiendo de ellos es necesario satisfacer distintas condiciones. En caso de querer medir la relación lineal entre dos variables, la recta de regresión lo va a indicar de forma directa (ya que calcula la correlación). Sin embargo, en caso de querer predecir el valor de una variable en función de la otra, no solo se necesita calcular la recta, sino que además hay que asegurar que el modelo sea bueno (Amat, 2016).

3. Supuestos (hipótesis) fundamentales en los que se basa

La estimación de parámetros requiere de un conjunto de hipótesis que deben de cumplirse en un modelo de regresión lineal simple.

3.1 Hipótesis relativas a los perturbadores

Los perturbadores constituyen un conjunto de variables individualmente poco relevantes, y si nos basamos en el supuesto de que estas variables son independientes entre sí, la perturbación tendrá en virtud del teorema del límite central una distribución aproximadamente normal; además, es lógico suponer que estas variables irrelevantes actúan en dirección positiva o negativa y por lo tanto su media será cero; finalmente parece verosímil que la distribución muestral de las perturbaciones tendrá una dispersión constante (varianza constante) e independiente del valor de X (homocedasticidad) de lo expresado se desprende lo siguiente:

a) El valor esperado del término de error es igual a cero; es decir:

$$E(\mu_i) = 0: \quad i = 1,2,..,n$$

El hecho de aceptar esta hipótesis significa que las perturbaciones van a tener valores positivos y negativos.

b) Todas las perturbaciones aleatorias tienen varianza constante, es decir $E(\mu_i^2) = \sigma^2; 1,2, \dots, n$

Esta propiedad se conoce como homocedasticidad, o sea que la variable aleatoria error tiene varianza finita, constante e independiente de X_i.

c) Las perturbaciones (μ_i) son independientes entre sí (incorrelacionadas en sentido estadístico), es decir la correlación de los errores correspondientes a observaciones distintas es igual a cero: $E(\mu_i\mu_j) = cov(\mu_i\mu_j) = 0 \; para \; i \neq j$

Esta propiedad se conoce como no autocorrelación (no están auto correlacionadas), Los valores que asume ui para cada i son completamente independientes de todos los valores precedentes.

d) El termino de error sigue una distribución normal con media cero y varianza σ^2. μ_i= N (0, σ^2). Si e_i se distribuye normalmente, entonces Y_i también se distribuyen normalmente, esto se deduce puesto que Y es una combinación lineal de los errores los cuales son todos normales.

e) Asunción adicional. μ_i, μ_j no son solamente correlacionadas sino necesariamente independientes.

3.2 Hipótesis relativas a las variables.

El modelo en estudio es de dos variables. La estimación suele hacerse suponiendo que X es una variable fija, es decir no aleatoria e independiente del muestreo; En cuanto a Y es una variable aleatoria que puede descomponerse en dos partes:

a) La variable X no es aleatoria, es decir es fija, los valores vienen fijados.

b) La relación entre X e Y es una relación lineal.

c) La variable endógena Y es evidentemente una variable aleatoria, cuyos parámetros serán:

$Y = \hat{\beta}_0 + \hat{\beta}_1 x + e_i$

$E(Y) = E(\hat{\beta}_0 + \hat{\beta}_1 x + e_i)$ tomando valor esperado

$E(Y) = E(\hat{\beta}_0) + E(\hat{\beta}_1 X_i) + E(e_i)$ empleando propiedades y conociendo que E (e_i) = 0

$E(Y) = E(\hat{\beta}_0) + \hat{\beta}_1 E(X_i)$ Xi constante

$E(Y) = \hat{\beta}_0 + \hat{\beta}_1 X_i$ es la media

Por otro lado, por definición de varianza tenemos:

$\sigma_y^2 = E[Y_i - E[Y_i]]^2 = E[(\hat{\beta}_0 + \hat{\beta}_1 X_i + e_i)]^2$

$E[(\hat{\beta}_0 + \hat{\beta}_1 X_i + e_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)]^2$

$E[(e_i)]^2 = \sigma^2$ es la varianza.

d) Las variables Y e X se obtienen sin errores de observación.

3.3 Hipótesis relativas a los parámetros

El valor de los parámetros especifica una determinada estructura entre las variables, que se supone constante para todos los elementos de la población y de la muestra.

Los parámetros estructurales son constantes para todas las unidades de la muestra y no existe ninguna restricción para ellos.

4. Inferencia

4.1 Significancia e intervalo de confianza para β_0 y β_1

En la mayoría de casos, aunque el estudio de regresión se aplica a una muestra, el objetivo último es obtener un modelo lineal que explique la

relación entre las dos variables en toda la población. Esto significa que el modelo generado es una estimación de la relación poblacional a partir de la relación que se observa en la muestra y, por lo tanto, está sujeta a variaciones. Para cada uno de los parámetros de la ecuación de regresión lineal simple (β_0 y β_1) se puede calcular su significancia (p-value) y su intervalo de confianza. El test estadístico más empleado es el t-test (existen alternativas no paramétricas) (Amat, 2016).

Existe un uso diferenciado entre la z y la t, se usa la Z cuando el tamaño de muestra es grande es decir mayor a 30 y t cuando es menor a 30.

DISTRIBUCION Z.- es una distribución normal con media 0 y varianza 1, se emplea para probar hipótesis ya que los parámetros deben distribuirse normalmente. Dentro de esta curva normal se halla el 100% de los valores que toman los parámetros.

4.2 Inferencia (Hipótesis) para $\beta 1$.

1) Hipótesis estadística.

   $H0$: No hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es cero. Y es independiente de X, no hay coherencia, la pendiente es cero, X no influye en Y. $\beta_1 = 0$.

   Ha: Hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es distinta de cero. X influye en Y, existe relación lineal, Y es dependiente de X. $\beta_1 \neq 0$.

2) nivel de significancia = α

3) prueba estadística:

   Cálculo del estadístico y del p-value:

$$z = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}}; \text{muestra grande}$$

$$t = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}}; \text{muestra pequeña}$$

   Varianza del estimador:

$$S^2_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta\sqrt{\sum x_i^2}}{S_e} = \frac{S_e^2}{\sum x_i^2} = S_e^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Desviación estándar del estimador:

$$S_{\hat{\beta}_1} = \sqrt{S^2_{\hat{\beta}_1}}$$

4) región critica, decisión:

Utilizando valor puntual, se rechaza Ho si:

Muestra grande $Z > Z_{\alpha/2}$ y muestra pequeña: $t > t_{(n-2),\alpha/2}$

Utilizando probabilidades: $p < \alpha$

4.3 Inferencia (Hipótesis) para β_0.

1) Hipótesis estadística.

$Ho: \beta_0 = 0$

$Ha: \beta_0 \neq 0$

2) nivel de significancia = α
3) prueba estadística:

$$Z = \frac{\hat{\beta}_0 - 0}{S_{\hat{\beta}_0}} \text{ ; muestra grande}$$

$$t = \frac{\hat{\beta}_0 - 0}{S_{\hat{\beta}_0}} \text{; muestra pequeña}$$

Varianza del estimador:

$$S^2_{\hat{\beta}_0} = \frac{S_e^2 \sum X_i^2}{n \sum x_i^2} = S_e^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

$$S_{\hat{\beta}_0} = \sqrt{S^2_{\hat{\beta}_0}}$$

4) región critica, decisión:

Utilizando valor puntual, se rechaza Ho si:

Muestra grande $Z > Z_{\alpha/2}$ y muestra pequeña: $t > t_{(n-2),\alpha/2}$

Utilizando probabilidades: $p < \alpha$

4.4 Cálculo de la Varianza residual (varianza no explicada)

La varianza residual σ^2es desconocida, siendo su estimador insesgado, entonces:

$$MSE = S_e^2 = \frac{\sum(Y_i - \hat{\beta}_0 X_i - \hat{\beta}_1)^2}{n-2}$$

$$S_e^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum X_i Y_i}{n-2} = \frac{(n-1)(S_y^2 - \hat{\beta}_1{}^2 S_x^2)}{n-2}$$

$$= \frac{\sum y_i^2 - \hat{\beta}_1{}^2 \sum x_i^2}{n-2}$$

La varianza del error σ^2 se estima a partir del Residual Standard Error (RSE), que puede entenderse como la diferencia promedio que se desvía la variable respuesta de la verdadera línea de regresión. En el caso de regresión lineal simple, RSE equivale a (Amat, 2016):

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)}$$

Grados de libertad (df) = número observaciones - 2 = número observaciones -número predictores – 1, suma del cuadrado de cada residuo (RSS).

4.5 Los intervalos de confianza en regresión

Los intervalos de confianza se utilizan para evaluar la precisión de las estimaciones de los parámetros de la regresión y para hacer inferencias sobre la relación entre las variables independientes y la variable dependiente. Los intervalos de confianza son un rango de valores dentro del cual se espera que se encuentre el verdadero valor del parámetro con un cierto nivel de confianza.

En la regresión lineal simple, los intervalos de confianza se pueden calcular para los coeficientes de la regresión (intercepto y pendiente) y para la predicción de valores individuales de la variable dependiente.

En resumen, los intervalos de confianza en regresión son una herramienta útil para evaluar la precisión de las estimaciones de los parámetros de la regresión y para hacer inferencias sobre la relación entre las variables independientes y la variable dependiente.

4.6 Intervalos de confianza para β_0

$$p(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

$$p(-z_{\alpha/2} < \frac{\hat{\beta}_0 - \beta_0\sqrt{n\sum x_i^2}}{\sigma_e\sqrt{\sum X_i^2}} < z_{\alpha/2}) = 1 - \alpha$$

$$p(\hat{\beta}_0 - z_{\alpha/2}\frac{\sigma_e\sqrt{\sum X_i^2}}{\sqrt{n\sum x_i^2}} < \beta_0 < \hat{\beta}_0 + z_{\alpha/2}\frac{\sigma_e\sqrt{\sum X_i^2}}{\sqrt{n\sum x_i^2}}) = 1 - \alpha$$

Muestra grande:

$$IC = p(\hat{\beta}_0 - z_{\alpha/2}S_{\hat{\beta}_0} < \beta_0 < \hat{\beta}_0 + z_{\alpha/2}S_{\hat{\beta}_0}) = 1 - \alpha$$

Muestra pequeña

$$IC = p(\hat{\beta}_0 - t_{(n-2),\alpha/2}S_{\hat{\beta}_0} < \beta_0 < \hat{\beta}_0 + t_{(n-2),\alpha/2}S_{\hat{\beta}_0}) = 1 - \alpha$$

4.7 Intervalo de confianza para el parámetro β_1

Muestra grande:

$$p(\hat{\beta}_1 - z_{\alpha/2}S_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + z_{\alpha/2}S_{\hat{\beta}_1}) = 1 - \alpha$$

Muestra pequeña:

$$p(\hat{\beta}_1 - t_{(n-2),\alpha/2}S_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{(n-2),\alpha/2}S_{\hat{\beta}_1}) = 1 - \alpha$$

Cuanto menor es el número de observaciones n, menor la capacidad para calcular el error estándar del modelo. Como consecuencia, la exactitud de

los coeficientes de regresión estimados se reduce. Esto tiene importancia sobre todo en la regresión múltiple (Amat, 2016).

En R, cuando se genera el modelo de regresión lineal, se devuelve junto con el valor de la pendiente y la ordenada en el origen el valor del estadístico *t* obtenido para cada uno y los p-value correspondientes. Esto permite saber, además de la estimación de β_0 y β_1, si son significativamente distintos de 0.

También es posible que se quiera determinar una región de confianza para la estimación simultánea de los parámetros, sabiendo que:

$$Q = \frac{\left[n(\hat{\beta}_0 - \beta_0)^2 - 2n\overline{X}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + (\hat{\beta}_1 + \beta_1)^2 \sum X^2\right]}{\sigma_e^2} \rightarrow X_2^2$$

$$\frac{(n-2)S_e^2}{\sigma_e^2} \rightarrow X_{(n-2)}^2$$

Se puede ver que:

$$F = \frac{\left[n(\hat{\beta}_0 - \beta_0)^2 - 2n\overline{X}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + (\hat{\beta}_1 + \beta_1)^2 \sum X^2\right]}{2S_e^2} = F_{2.(n-2)}$$

5. Residuos del modelo

El residuo de una estimación se define como la diferencia entre el valor observado y el valor esperado acorde al modelo. A la hora de sumar el conjunto de residuos hay dos posibilidades:

- La suma del valor absoluto de cada residuo.
- La suma del cuadrado de cada residuo (RSS). Esta es la aproximación más empleada (mínimos cuadrados) ya que magnifica las desviaciones más extremas. En R, cuando se genera un modelo los residuos también se calculan automáticamente y se almacenados dentro del modelo.

Cuanto mayor es la sumatoria del cuadrado de los residuos, menor es la precisión con la que el modelo puede predecir el valor de la variable

dependiente a partir de la variable predictora. Los residuos son muy importantes, puesto que, en ellos, se basan las diferentes medidas de la bondad de ajuste del modelo (Amat, 2016).

6. Análisis de varianza en R.L.S. (ANVA, ANOVA)

El análisis de varianza (ANOVA) en regresión se utiliza para determinar si la regresión lineal simple es estadísticamente significativa y para evaluar la importancia relativa de las variables independientes en la explicación de la variabilidad de la variable dependiente. El ANOVA en regresión se basa en la comparación de la varianza explicada por la regresión con la varianza no explicada (residual)

El ANOVA en regresión se divide en dos partes: el análisis de varianza global y el análisis de varianza individual. El análisis de varianza global se utiliza para determinar si la regresión en su conjunto es estadísticamente significativa. El análisis de varianza individual se utiliza para determinar la importancia relativa de cada variable independiente en la explicación de la variabilidad de la variable dependiente.

El análisis de varianza global se realiza mediante la prueba F de Fisher. El estadístico F se calcula dividiendo la varianza explicada por la regresión entre la varianza no explicada (residual). Si el valor de F es mayor que el valor crítico correspondiente para un nivel de significancia determinado, se rechaza la hipótesis nula de que la regresión no es significativa.

Es importante tener en cuenta que el ANOVA en regresión asume que los residuos siguen una distribución normal con media cero y varianza constante, y que los residuos son independientes. Si estos supuestos no se cumplen, los resultados del ANOVA pueden ser sesgados y poco confiables. Por lo tanto, es importante realizar una evaluación cuidadosa de los supuestos de la regresión antes de realizar cualquier análisis.

Tomemos la desviación $Y_i - \overline{Y}$ (variación total), sumando y restando $\hat{Y}_i$ para la media de Y_i.

$$Y_i - \overline{Y} = \underbrace{\hat{Y}_i - \overline{Y}} + \underbrace{Y_i - \hat{Y}_i}$$

Variación explicada + variación no explicada

Elevando al cuadrado y aplicando sumatorias

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$
$$\sum y_i^2 = \sum \hat{Y}_i^2 + \sum e_i^2$$
$$SCT = SCR + SCE$$

(Canavos, 1995) Si todas las observaciones se encuentran sobre la recta estimada, el valor de todos los residuos es cero y SCE=0. Entre más grande es el valor de SCE mayor es la contribución de la componente de error a la variación de las observaciones, o mayor es la incertidumbre. La SCR representa la variación de la observación que es atribuible al efecto lineal de X sobre Y. Si la pendiente es cero, entonces SCR= 0. de esta forma entre más grande es la proporción de SCR con respecto a SCT, mayor será la cantidad de la variación en las observaciones que puede explicarse mediante el termino lineal BX. Las SC son asociados con números llamados grados de libertad (g.l.). Este número indica la fracción de información independiente que existe en n números dependientes $y_1,\ldots,y_n$.

Para la **SCT** existen *(n-1)* g.l. ya que se pierde uno por causa de la restricción lineal $\sum(Y_i - \overline{Y}) = 0$ entre las observaciones Y_i.

Para la **SCE** existen *(n-2)* g.l. Ya que se pierden dos g.l. a causa de dos restricciones lineales dadas (se estima dos parámetros).

Para la **SCR** debido a que es aditivo, entonces, g.l.SCR= g.l.SCT - g.l.SCE, esto es; (n-1) -(n-2) = 1.

A la ecuación fundamental dividimos sobre n, se obtiene varianzas poblacionales sesgadas, esto es:

$$\frac{\sum y_i^2}{n} = \frac{\sum \hat{Y}_i^2}{n} + \frac{\sum e_i^2}{n}$$

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2$$

Para corregir se usa los g.l. y obtenemos varianzas insesgadas.

$$\sigma_y^2 = S_y^2 = \frac{\sum y_i^2}{n-1} \qquad \sigma_{\hat{y}}^2 = S_{\hat{y}}^2 = \frac{\sum \hat{Y}_i^2}{1} \qquad \sigma_e^2 = S_e^2 = \frac{\sum e_i^2}{n-2}$$

Con todos estos datos podemos consolidar un cuadro ANVA

Tabla 1.

ANVA para el modelo lineal simple

Fuente de variación	**Grados de libertad**	**Suma de cuadrados (SC)**	**Cuadrados medios (CM)**	**F**
Regresión	1	SCR	CMR	CMR/CME
Error	n-2	SCE	CME	-
Total	n-1	SCT	-	-

Donde:

$Suma\ de\ cuadrados\ de\ la\ regresion: SCR = \sum \hat{y}_i^2 = \hat{\beta}_1^2 \sum x_i^2 = (n)\hat{\beta}_1^2 S_x^2$

$Suma\ de\ cuadrados\ del\ error: SCE = \sum e_i^2 = \sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum X_i Y_i$

$Suma\ de\ cuadrados\ del\ total: SCT = \sum y_i^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = \sum Y_i^2 - n\overline{Y}^2$

$Cuadrado\ medio\ de\ la\ regresion: CMR = SCR/1$

$Cuadrado\ medio\ del\ error: CME = SCE/(n-2) = S_e^2$

Tabla 2.

ANVA más general corregida por la media

Fuente de variación	**Grados de libertad**	**Suma de cuadrados (SC)**	**Cuadrados medios (CM)**	**F**
Regresión	1	SCR	CMR	CMR/CME
Error	n-2	SCE	CME	-

Total	n-1	SCT	-	-
debido a b*	1	SCC	CMC=SCC/1	CMC/CME
Total	n	SCTC		

$$SCC = (\sum y_i)^2/n = n\overline{Y}^2$$

$$SCTC = \sum Y_i^2$$

Prueba de hipótesis de F para la significación de la regresión.

1) Hipótesis estadística:

   Ho: $\beta_0 = \beta_1 = 0$; no existe relación lineal entre X e Y

   Ha: $\beta_0 \neq \beta_1 \neq 0$; existe relación lineal entre X e Y

2) nivel de significancia = α

3) prueba estadística, estadístico de contraste

$$F = \frac{CMR}{CME}$$

4) región critica, decisión:

   Se rechaza Ho si:

   Muestra grande $F > F_{(1,n-2),\alpha/2}$

7. Bondad de ajuste del modelo

Una vez que se ha ajustado un modelo es necesario verificar su eficiencia, ya que aun siendo la línea que mejor se ajusta a las observaciones de entre todas las posibles, el modelo puede ser malo. Las medidas más utilizadas para medir la calidad del ajuste son: error estándar de los residuos, el test F y el coeficiente de determinación R^2.

Error estándar de los residuos (Residual Standard Error, RSE): Mide la desviación promedio de cualquier punto estimado por el modelo respecto de la verdadera recta de regresión poblacional. Tiene las mismas unidades que

la variable dependiente Y. Una forma de saber si el valor del RSE es grande consiste en dividirlo entre el valor medio de la variable respuesta, obteniendo así un % de la desviación.

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

En modelos lineales simples, dado que hay un único predictor

$$(n-p-1) = (n-2)$$

Coeficiente de determinación R^2:

Describe la proporción de variabilidad observada en la variable dependiente Y explicada por el modelo y relativa a la variabilidad total. Su valor está acotado entre 0 y 1. Al ser adimensional presenta la ventaja frente al RSE de ser más fácil de interpretar.

$$R^2 = \frac{Suma\ de\ cuadrados\ totales - Suma\ de\ cuadrados\ residuales}{Suma\ de\ cuadrados\ totales} =$$

$$1 - \frac{SCE}{SCT} =$$

$$1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$$

En los modelos de regresión lineal simple el valor de R^2 se corresponde con el cuadrado del coeficiente de correlación de Pearson (r) entre X e Y, no siendo así en regresión múltiple. Existe una modificación de R^2 conocida como $R2{-}ajustado$ que se emplea principalmente en los modelos de regresión múltiple. Introduce una penalización cuantos más predictores se incorporan al modelo. En los modelos lineales simples no se emplea.

Test F: El test F es un test de hipótesis que considera como hipótesis nula que todos los coeficientes de correlación estimados son cero, frente a la hipótesis alternativa de que al menos uno de ellos no lo es. Se emplea en modelos de regresión múltiple para saber si al menos alguno de los predictores introducidos en el modelo contribuye de forma significativa. En modelos

lineales simples, dado que solo hay un predictor, el p-value del test F es igual al p-value del t-test del predictor (Amat, 2016).

Desarrollo:

Los residuos pueden contribuir a proporcionar una medida útil de hasta qué punto la recta de regresión estimada se ajusta a los datos o que porcentaje de la variación es explicada por el modelo.

Tomemos la variación total de Y con respecto a su media.

$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$ Como desviaciones de puede representar:

$y_i = \hat{y}_i + e_i$ Elevando al cuadrado y aplicando sumatorias

$$\sum y_i^2 = \sum (\hat{y}_i + e_i)^2$$

$\sum y_i^2 = \sum \hat{y}_i^2 + 2\sum \hat{y}_i e_i + \sum e_i^2$ Como $\sum \hat{y}_i e_i = 0$ queda

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

SCT = SCR + SCE

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

$$R^2 = \left(\frac{S_{XY}}{S_X S_Y}\right)^2$$

Coeficiente de no determinación: 1- R^2

Coeficiente de alejamiento: $\sqrt{1 - R^2}$

Ejemplo: R^2 = 90%. La regresión mínimo cuadrática explica el 90% de la varianza de Y. La proporción de variación total alrededor de la media explicada por la regresión es 90%. La reducción de la suma de cuadrados del error total fue del 90%. Entre más cercano se encuentre R^2 a 100%, mayor será el grado de asociación lineal entre X e Y.

8. Condiciones para el uso de la regresión lineal.

La regresión lineal simple es una técnica estadística que se utiliza para examinar la relación entre dos variables continuas. Sin embargo, la validez de los resultados de la regresión depende de ciertos supuestos. Estos supuestos son los siguientes:

1. Linealidad: (Amat, 2016) La relación entre ambas variables debe ser lineal, lo que significa que los cambios en una variable deben estar asociados con cambios proporcionales en la otra variable. Si la relación es no lineal, la regresión lineal simple no es apropiada.

   Para comprobarlo se puede recurrir a:

   - Graficar ambas variables a la vez (scatterplot o diagrama de dispersión), superponiendo la recta del modelo generado por regresión lineal.
   - Calcular los residuos para cada observación acorde al modelo generado y graficarlos (scatterplot). Deben distribuirse de forma aleatoria en torno al valor 0.

2. Distribución Normal de los residuos: Los errores de la regresión deben seguir una distribución normal, con media 0. Si los errores no siguen una distribución normal, los intervalos de confianza y las pruebas de hipótesis pueden no ser precisos. La normalidad se puede comprobar con un histograma, con la distribución de cuantiles (*qqnorm()* + *qqline()*) o con un test de hipótesis de normalidad. Los valores extremos suelen ser una causa frecuente por la que se viola la condición de normalidad (Amat, 2016).

3. Varianza de residuos constante (homocedasticidad): La varianza de los errores debe ser constante en todo el rango de valores de la variable independiente (X). Si la varianza no es constante, se dice que hay heterocedasticidad. La heterocedasticidad puede afectar la precisión de los coeficientes de la regresión y los intervalos de confianza

Se puede comprobar mediante gráficos (scatterplot) de los residuos de cada observación (formas cónicas son un claro indicio de falta de homocedasticidad) o mediante contraste de hipótesis mediante el test de Breusch-Pagan (Amat, 2016).

4. Independencia, Autocorrelación: Las observaciones deben ser independientes entre sí. Esto significa que el valor de la variable independiente no debe estar relacionado con el valor de la variable dependiente en ninguna otra observación. La violación de este supuesto puede conducir a resultados sesgados y poco confiables.
5. Valores atípicos y de alta influencia: Los valores atípicos pueden tener un efecto significativo en los resultados de la regresión (pueden generar una falsa correlación que realmente no existe, u ocultar una existente.). Si hay valores atípicos en los datos, deben ser identificados y evaluados para determinar si deben ser excluidos de la regresión.

Es importante tener en cuenta que la violación de estos supuestos no siempre invalida los resultados de la regresión lineal simple. Sin embargo, la validez de los resultados se verá afectada y puede ser necesario utilizar técnicas más avanzadas para analizar los datos. Por lo tanto, es importante evaluar cuidadosamente los supuestos de la regresión lineal simple antes de realizar cualquier análisis.

9. Predicción de valores

(Amat, 2016) Una vez generado un modelo que se pueda considerar válido, es posible predecir el valor de la variable dependiente Y para nuevos valores de la variable predictora X. Es importante tener en cuenta que las predicciones deben, a priori, limitarse al rango de valores dentro del que se encuentran las observaciones con las que se ha generado el modelo. Esto es importante puesto que solo en esta región se tiene certeza de que se cumplen

las condiciones para que el modelo sea válido. Para calcular las predicciones se emplea la ecuación generada por regresión.

Dado que el modelo generado se ha obtenido a partir de una muestra y por lo tanto las estimaciones de los coeficientes de regresión tienen un error asociado, también lo tienen los valores de las predicciones. Existen dos formas de medir la incertidumbre asociada con una predicción:

- Intervalos de confianza (predicción puntual): Responden a la pregunta ¿Cuál es el intervalo de confianza del valor promedio de la variable respuesta Y para un determinado valor del predictor X? (Amat, 2016)

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{MSE\left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{(x_i - \bar{x})^2}\right)}$$

Obtención del intervalo.

a) error de predicción: (fuente de error) se muestra por la ecuación.

$$\mu_i = Y_p - \hat{Y}_i$$

b) valor medio esperado de μ_i:

$$E(\mu_i) = E(Y_p - \hat{Y}_i)$$
$$E(\mu_i) = E(Y_p - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$
$$E(\mu_i) = E(Y_p) - E(\hat{\beta}_0) - X_i E(\hat{\beta}_1)$$
$$E(\mu_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = 0$$

c) Varianzas de $\hat{Y}_p$:

Se sabe que $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Tomemos:

$$V(\hat{Y}_i) = V(\hat{\beta}_0 + \hat{\beta}_1 X_p)$$
$$V(\hat{Y}_i) = V(\overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_p)$$
$$V(\hat{Y}_i) = V(\overline{Y}) + (X_p - \overline{X})^2 V(\hat{\beta}_1)$$

$$V(\hat{Y}_i) = \frac{\sigma_e^2}{n} + (X_p - \overline{X})^2 \frac{\sigma_e^2}{\sum(X - \overline{X})^2}$$

$$V(\hat{Y}_i) = \sigma_e^2 \left[\frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum(X - \overline{X})^2}\right]$$

Estimando σ_e^2 *por* S_e^2

$$V(\hat{Y}_i) = S^2_{\hat{Y}_p} = S_e^2 \left[\frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum(X - \overline{X})^2}\right]$$

La desviación estándar es:

$$S_{\hat{Y}_p} = S_e \sqrt{\frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum(X - \overline{X})^2}} = S_e \sqrt{\frac{1}{n} + \frac{(X_p - \overline{X})^2}{(n-1)S_X^2}}$$

con:

$$t = \frac{\hat{Y}_p - Y_p}{S_{\hat{Y}_p}} \rightarrow t_{(n-2)}$$

Intervalo de confianza: para E(y/x)

$$\left[\hat{Y}_p - t_{\alpha/2(n-2)} S_{\hat{Y}_p} < E(Y/X) < \hat{Y}_p + t_{\alpha/2(n-2)} S_{\hat{Y}_p}\right] = 1 - \alpha$$

- Intervalos de predicción: Responden a la pregunta ¿Dentro de que intervalo se espera que esté el valor de la variable respuesta Y para un determinado valor del predictor X?

$$\hat{y} \pm t_{\alpha/_{2,n-2}} \sqrt{MSE(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{(x_i - \bar{x})^2})}$$

Desarrollo:

a) **valor medio ordenado**:

$$E(\hat{Y}_p - Y_p) = E(\hat{\beta}_0 + \hat{\beta}_1 X_p - (\alpha + \beta X_p + \mu_p)$$

$$E(\hat{Y}_p - Y_p) = E(\hat{\beta}_0 + \hat{\beta}_1 X_p - \alpha - \beta X_p - \mu_p)$$

$$E(\hat{Y}_p - Y_p) = E(\hat{\beta}_0) + X_p E(\hat{\beta}_1) - E(\alpha) + X_p E(\beta) + E(\mu_p)$$

$$E(\hat{Y}_p - Y_p) = \alpha + X_p \beta - \alpha - X_p \beta - 0$$

$$E(\hat{Y}_p - Y_p) = 0$$

b) **Varianza estimada** de una E individual predicha para una X dada:

$$\sigma^2_{\underset{(Y_p - Y_p)}{\wedge}} = V(\hat{\beta}_0 + \hat{\beta}_1 X_p - e_p)$$

$$\sigma^2_{\underset{(Y_p - Y_p)}{\wedge}} = V(\overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_p - e_p)$$

$$\sigma^2_{\underset{(Y_p - Y_p)}{\wedge}} = V(\overline{Y}) + (X_p - \overline{X})^2 V(\hat{\beta}_1) + V(e_p)$$

$$\sigma^2_{\underset{(Y_p - Y_p)}{\wedge}} = \frac{\sigma_e^2}{n} + (X_p - \overline{X})^2 \frac{\sigma_e^2}{\sum(Xi - \overline{X})^2} + \sigma_e^2$$

$$\sigma^2_{\underset{(Y_p - Y_p)}{\wedge}} = \sigma_e^2 \left[1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum(Xi - \overline{X})^2}\right] \quad entonces:$$

$$S^2_{\underset{(Y_p - Y_p)}{\wedge}} = S_e^2 \left[1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum(Xi - \overline{X})^2}\right]$$

Intervalo de predicción

$$I.C = \left[\hat{Y}_p - t_{\alpha/2,(n-2)} S_{(\hat{Y}_p - Y_p)} < Y_o < \hat{Y}_p + t_{\alpha/2,(n-2)} S_{(\hat{Y}_p - Y_p)}\right] = 1 - \alpha$$

Si bien ambas parecen similares, la diferencia se encuentra en que los intervalos de confianza se aplican al valor promedio que se espera de Y para un determinado valor de X, mientras que los intervalos de predicción no se aplican al promedio. Por esta razón los segundos siempre son más amplios que los primeros. En R se puede emplear la función *predict*() que recibe como argumento el modelo calculado, un dataframe con los nuevos valores del predictor X y el tipo de intervalo (confidence o prediction).

Una característica que deriva de la forma en que se calcula el margen de error en los intervalos de confianza y predicción, es que el intervalo se ensancha a medida que los valores de X se aproximan a los extremos el rango observado (Amat, 2016).

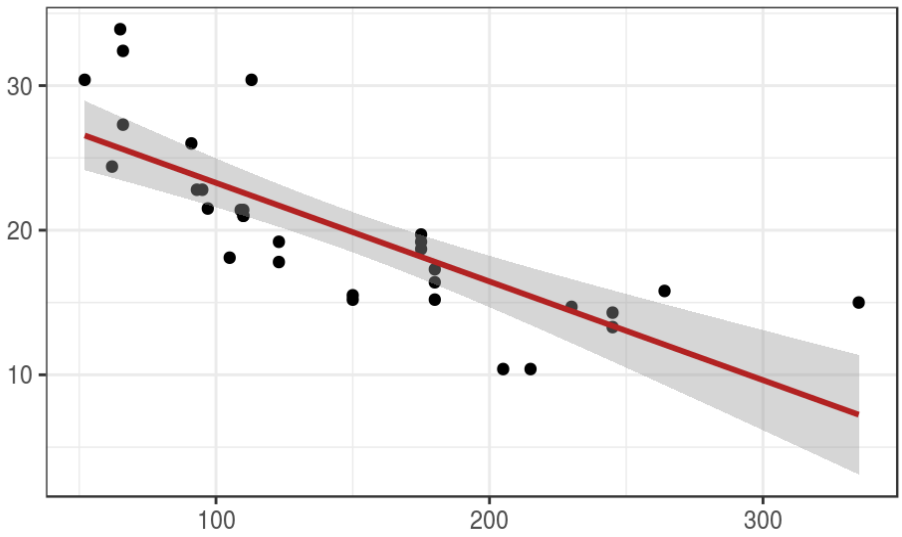


Figura 1. *Bandas de confianza para la línea de regresión*

¿Por qué ocurre esto? Prestando atención a la ecuación del error estándar del intervalo de confianza, el numerador contiene el término $(x_k - x)^2$ (lo mismo ocurre para el intervalo de predicción) (Amat, 2016).

$$\sqrt{MSE\left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{(x_i - \bar{x})^2}\right)}$$

Este término se corresponde con la diferencia al cuadrado entre el valor xk para el que se hace la predicción y la media x de los valores observados del predictor X. Cuanto más se aleje xk de x mayor es el numerador y por lo tanto el error estándar.

10. Ejemplo 1 en R.

Regresión lineal simple con variable independiente numérica.

```
# regresión lineal simple
# Y: volumen de ventas
# X: gastos en publicidad
Y <- c10,15,20,22,30,32,12,16,23,29,31,14,17,27,28)
X <- c16,32,48,56,64,80,20,22,50,52,75,35,20,55,35)
```

Creando un *data.frame* y observando los primeros 6 datos

```
datos <- data.frame(X,Y)
head(datos)
```

La función head muestra los primeros 6 datos, sea de un data.frame o un vector

```
##    X  Y
## 1 16 10
## 2 32 15
## 3 48 20
## 4 56 22
## 5 64 30
## 6 80 32
```

La función *str* (estructura) sirve para observar la estructura de los datos (cuantitativos o cualitativos)

```
str(datos)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ X: num  16 32 48 56 64 80 20 22 50 52 ...
##  $ Y: num  10 15 20 22 30 32 12 16 23 29 ...
```

Observamos que la data set está conformada por 15 observaciones y 2 variables. Así mismo, indica que:

- X e Y son variables numéricas (*num*), en este caso son datos cuantitativos enteros)

El volumen de ventas (Y) estará en función al gasto en publicidad (X), dicho de otra manera, el gasto en publicidad influye el volumen de ventas de la empresa.

a) Observar las variables descriptivamente.

```
# representación gráfica
# histograma de frecuencias para las variables X e Y
require(ggplot2)
par(mfrow = c(1,2)) # dos figuras en una fila
hist(datos$X, breaks = 10, main = "", xlab = "Gastos en publicidad",
     border = "darkred")
hist(datos$Y, breaks = 10, main = "", xlab = "Volúmen de ventas",
     border = "blue")
```

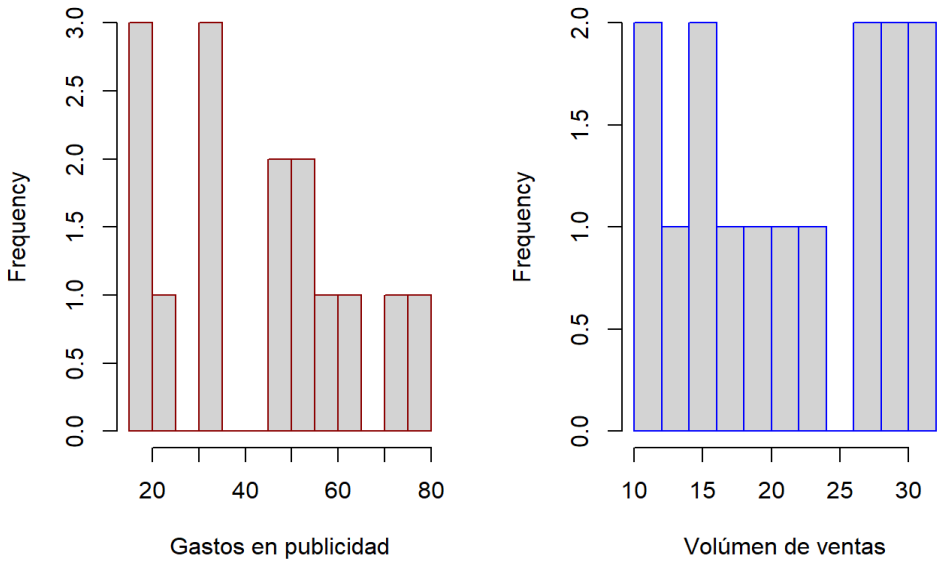


Figura 7. *Histograma de frecuencias para el número de bateos y numero de corridas*

La figura muestra una tendencia a valores bajos en gastos en publicidad y valores altos en volumen de venta, al parecer sin tendencia a una distribución normal.

```
# estadísticos descriptivos
summary(datos)
```

```
##        X               Y
##  Min.   :16.0   Min.   :10.00
##  1st Qu.:27.0   1st Qu.:15.50
##  Median :48.0   Median :22.00
##  Mean   :44.0   Mean   :21.73
```

```
##  3rd Qu.:55.5    3rd Qu.:28.50
##  Max.   :80.0    Max.   :32.00
```

El promedio de volumen de ventas (Y) es 21.73 millones de pesetas, con ventas mínimas de 10 millones y máximo de 32 millones de pesetas. El promedio de gastos en publicidad (X) es de 48 mil pesetas, con valores mínimo y máximo de 16 y 80 mil pesetas respectivamente. La mediana del volumen de ventas es 22 millones de pesetas y la mediana del gasto en publicidad es 48 mil pesetas.

b) Representación gráfica de las observaciones

Es necesario graficar diagramas de dispersión, para ver la tendencia de los datos.

```
# diagrama de dispersión
require(ggplot2)
ggplot(data = datos, mapping = aes(x = X, y = Y)) +
  geom_point(color = "firebrick", size = 2) +
  (labs(title = "Diagrama de dispersión", x="Gastos en publicidad",
        y= "Volúmen de ventas"))+
  theme_bw() +
  theme(plot.title = element_text(hjust =0.5))
```

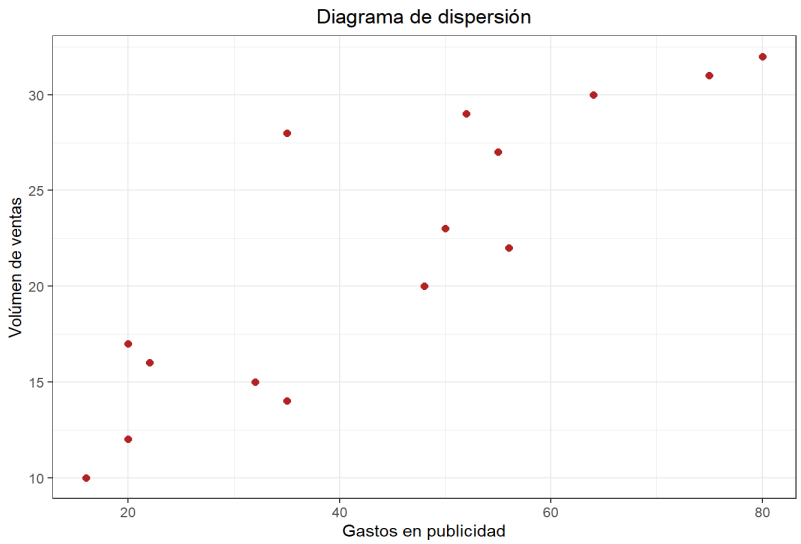


Figura 8. *Diagrama de dispersión*

La figura muestra un comportamiento (tendencia) positivo que puede ser ajustada a un modelo lineal, este hecho, podría interpretarse como: a más gastos en publicidad se incrementa el volumen de ventas. En la figura igualmente no se observa presencia de datos atípicos (outliers) o valores influyentes.

c) comportamiento del coeficiente de correlación

Otro indicador de una posible relación lineal entre las variables X e Y es el coeficiente de correlación.

Si no se observa en el diagrama de dispersión una tendencia lineal, y además se observa un valor bajo del coeficiente de correlación, no tiene sentido seguir adelante generando un modelo lineal, sería mejor buscar alternativas no lineales. Por tener datos cuantitativos, se usará el coeficiente de Pearson.

Cuando los datos son cuantitativos, el coeficiente recomendado es el coeficiente de correlación de Pearson , siempre y cuando se cumpla los supuestos del coeficiente.

```
# correlación de Pearson
cor.test(x = datos$X, y = datos$Y, method = "pearson")
```

```
##  Pearson's product-moment correlation
##
## data:  datos$X and datos$Y
## t = 6.3186, df = 13, p-value = 2.665e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6418579 0.9556266
## sample estimates:
##       cor
## 0.8685436
```

Resumen:

Coeficiente de correlación	Test de significancia	Intervalo de confianza de r	p-value
r = 0.8685	t = 6.3186	0.6419 - 0.9556	0.00002665

1) hipótesis estadística:

   Ho: la correlación es igual a 0

   Ha: la correlación es diferente a 0

2) nivel de significancia: $\alpha = 0.05$

3) coeficiente de correlación $r = 0.868$

   test de significancia: $t = 6.3186$

   p-value: $p = 0.00002665$

   intervalo de confianza: $IC = (0.6419 - 0.9556)$

4) decisión: El test de correlación muestra una relación lineal significativa $p(0.00002665) < \alpha(0.05)$, de intensidad considerable (r=0.868). Tiene sentido intentar generar un modelo de regresión lineal que permita predecir el volumen de ventas. Se espera que el coeficiente de correlación oscile entre 0.64 y 0.95 utilizando el IC al 95%.

d) Cálculo del modelo de regresión lineal simple

La regresión es una herramienta estadística que se utiliza para analizar la relación entre dos variables , así como para predecir los valores de la variable dependiente a partir de los valores de la variable independiente. El modelo es:

$$Y = a + bX + e$$

Y es la variable dependiente, X es la variable independiente, a es el intercepto (valor de Y cuando X es igual a cero), b es la pendiente de la recta y e es el error aleatorio.

La pendiente (b) de la recta indica el cambio en Y, por unidad de cambio en X. La magnitud de la pendiente indica la fuerza de la relación entre las dos variables.

Para obtener los estimadores de la ecuación, utilizamos el siguiente script

```
# Calculo del modelo de regresión lineal simple
modelo_lineal <- lm(Y ~ X, data=datos) # selección de variables y data
summary(modelo_lineal)
```

```
## Call:
## lm(formula = Y ~ X, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8635 -2.8558 -0.6466  1.8241  9.1365
##
## Coefficients:
##             Estimate  Std.Error  t value   Pr(>|t|)
## (Intercept)  7.70280    2.43091   3.169    0.0074 **
## X            0.31888    0.05047   6.319    2.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.831 on 13 degrees of freedom
## Multiple R-squared:  0.7544, Adjusted R-squared:  0.7355
## F-statistic: 39.92 on 1 and 13 DF,  p-value: 2.665e-05
```

Tanto el intercepto como la variable X (gastos en publicidad) son estadísticamente significativas al 5%, en ambos casos $[p(0.0074) < \alpha(0.05) y\ p(0.0000266) < \alpha(0.05)]$, con tendencia positiva. El modelo tiene un coeficiente de determinación de 75.44% (el modelo calculado explica el 75.44% de la variabilidad presente en la variable respuesta volumen de ventas mediante la variable independiente (gastos en publicidad). La ecuación de regresión se expresa como:

$$Y = 7.70280 + 0.31888(X)$$

Por cada unidad (mil pesetas) que se invierte en gastos de publicidad, el volumen de ventas se incrementa en promedio 0.3188 millones de pesetas.

El modelo en su conjunto es también significativo (ANVA) con F= 39.92 y $p(0.000026665) < \alpha(0.05)$. El error residual (RSS) = 3.831.

Cálculo del error cuadrático medio RMSE:

```
# Cálculo del error cuadrático medio RMSE:
sqrt(mean(modelo_lineal$residuals^2))
```

```
## [1] 3.566881
```

El modelo está prediciendo con 3.56688 de error.

Un modelo completo de análisis de varianza se puede obtener con el siguiente script.

```
anova(modelo_lineal)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1 586.09  586.09  39.925 2.665e-05 ***
## Residuals 13 190.84   14.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hasta aquí se puede concluir que el modelo es bueno y existe relación entre las variables.

e) Intervalos de confianza para los parámetros del modelo

```
# Intervalos de confianza para los parámetros del modelo
confint(modelo_lineal)
```

```
##                 2.5 %     97.5 %
## (Intercept) 2.4511286 12.9544694
## X           0.2098502  0.4279014
```

Los intervalos de confianza contienen a los coeficientes obtenidos. Por cada unidad (mil pesetas) que se invierte en gastos de publicidad, el volumen de ventas aumenta en promedio entre 0.2098 y 0.4279 millones de pesetas.

f) Representación gráfica del modelo

Es una línea recta que se ajusta a los datos de la muestra utilizando el modelo obtenido.

La recta de regresión se calcula utilizando la técnica de mínimos cuadrados, que implica encontrar la línea recta que minimiza la distancia entre los puntos de datos y la línea recta.

```
# Representación gráfica del modelo
ggplot(data = datos, mapping = aes(x = X, y = Y)) +
  geom_point(size=3) +
  (labs(title = "Diagrama de dispersión", x="Gastos en publicidad",
        y= "Volúmen de ventas"))+
  geom_smooth(method = "lm", se = FALSE, color = "firebrick") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.2))
```

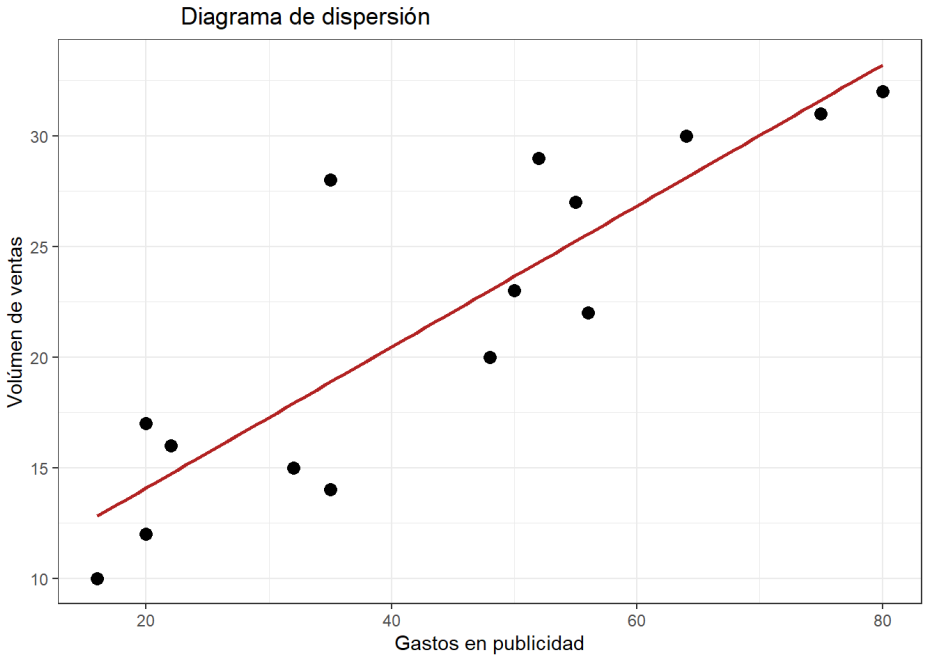


Figura 9. *Representación de la línea de regresión*

```
names(modelo_lineal)
```

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

La opción *names* guarda los resultados generados por el modelo (ejemplo: modelo_lineal$coefficients) guarda los coeficientes del modelo.

Además de la línea de mínimos cuadrados es recomendable incluir los límites superior e inferior del intervalo de confianza. Esto permite identificar la región en la que, según el modelo generado y para un determinado nivel de confianza, se encuentra el valor promedio de la variable dependiente.

Para poder representar el intervalo de confianza a lo largo de todo el modelo se recurre a la función predict() para predecir valores que abarquen todo el eje *X*. Se añaden al gráfico líneas formadas por los

límites superiores e inferiores calculados para cada predicción (Amat, 2016).

Obtenemos los valores predichos de Y para cada valor de X.

```
# predecir
nuevos_datos <- data.frame(X= seq(min(X),max(X)))
predict_value <- predict(modelo_lineal)
head(predict_value)# muestra 6 valores predichos
```

```
##        1        2        3        4        5        6
## 12.80481 17.90682 23.00884 25.55984 28.11085 33.21286
```

Graficando las bandas de confianza:

Intervalos de confianza de la respuesta media

```
# solo una banda
par(mfrow = c(1, 1))
puntos <- seq(from = min(datos$X),
              to = max(datos$X), length.out = 100)
limites_intervalo <- predict(object = modelo_lineal,
                             newdata = data.frame(X = puntos),
                             interval = "confidence", level = 0.95)
head(limites_intervalo, 3)
```

```
##        fit      lwr      upr
## 1 12.80481 9.078325 16.53130
## 2 13.01095 9.341982 16.67992
## 3 13.21710 9.605180 16.82901
```

```
plot(datos$X, datos$Y, col = "firebrick", pch = 19,
     ylab = "Gastos en publicidad", xlab = "Volúmen de ventas",
     main = "Volúmen de ventas ~ Gastos en publicidad")
abline(modelo_lineal, col = 1)
lines(x = puntos, y = limites_intervalo[, 2], type = "l", col = 2, lty
= 3)
```

```
lines(x = puntos, y = limites_intervalo[, 3], type = "l", col = 3, lty
= 3)
```

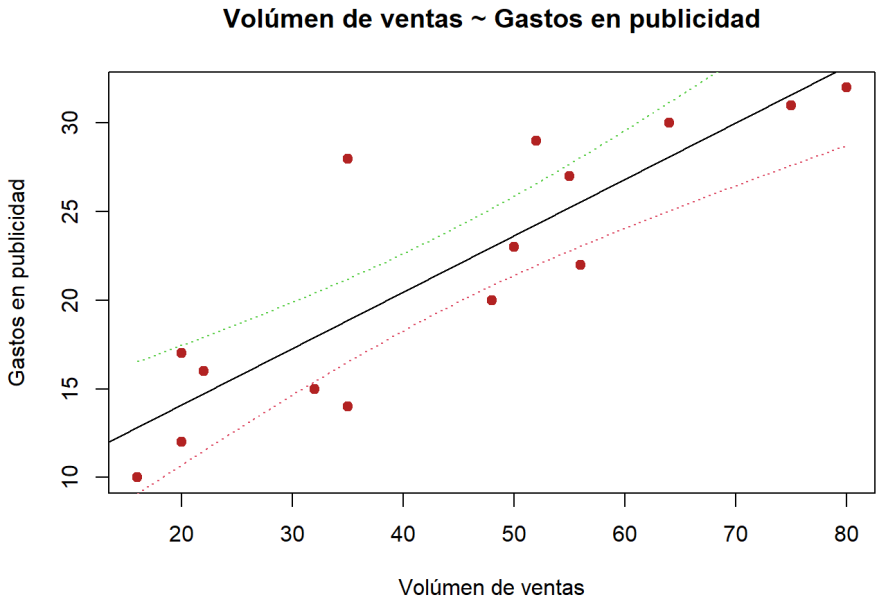


Figura 10. *Bandas de confianza para el ajuste de la regresión*

La función geom_smooth() del paquete ggplot2 genera la regresión y su intervalo de forma directa.

```
# con geom plot
ggplot(data = datos, mapping = aes(x = X, y = Y)) +
  geom_point(color = "firebrick", size = 2) +
  labs(title = "Diagrama de dispersion", x = "numero  de bateos") +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

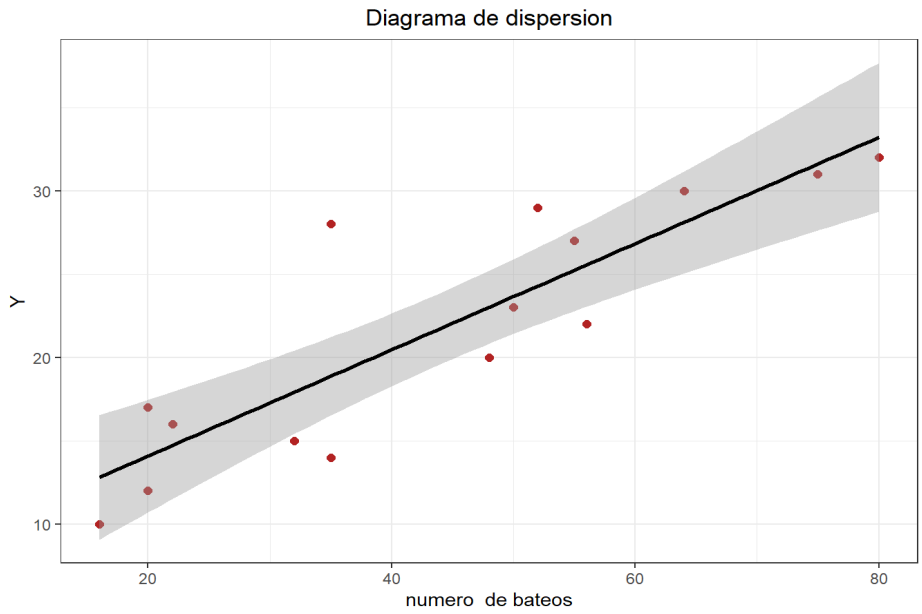


Figura 11. *Bandas de confianza para la respuesta media*

Intervalo de confianza para la respuesta media y para la predicción:

```
# dos bandas de confianza
par(mfrow = c(1, 1))
# Grafico dispersion y recta
plot(datos$X, datos$Y, col = "firebrick", pch = 19,
     ylab = "Y", xlab = "X",
     main = "")
abline(modelo_lineal, col = 1)

# Intervalos de confianza de la respuesta media:
# valores medios

ic <- predict(modelo_lineal, nuevos_datos, interval = 'confidence')
lines(nuevos_datos$X, ic[, 2], lty = 2)
lines(nuevos_datos$X, ic[, 3], lty = 2)

# Intervalos de predicción
# para cualquier valor
ic <- predict(modelo_lineal, nuevos_datos, interval = 'prediction')
lines(nuevos_datos$X, ic[, 2], lty = 2, col = 'red')
lines(nuevos_datos$X, ic[, 3], lty = 2, col = 'red')
```

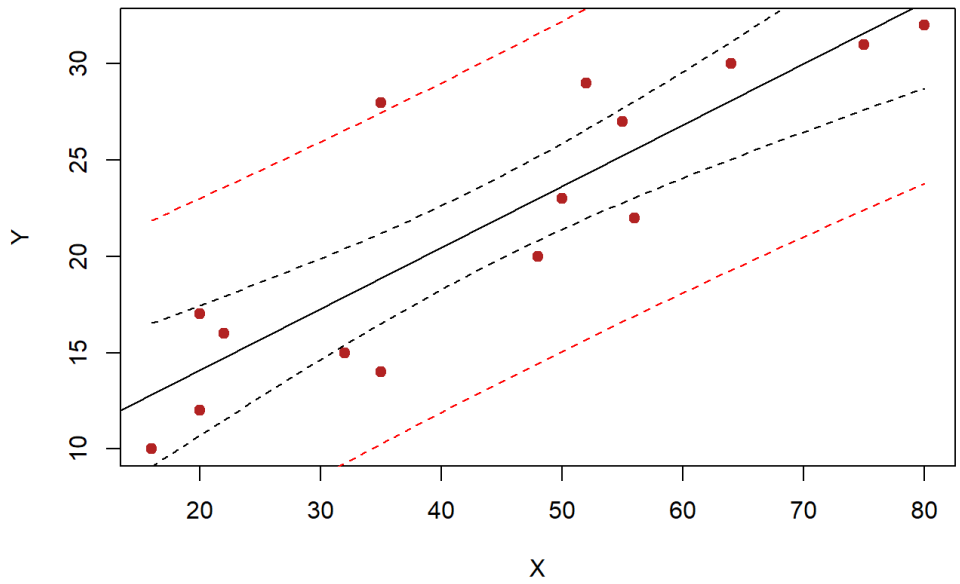


Figura 12. *Bandas de confianza para la respuesta media y la predicción*

Observamos un punto fuera de la banda de confianza de la predicción, posible punto outlier.

El siguiente script muestra los datos reales, los datos predichos y los errores calculados.

```
datos$prediccion <- modelo_lineal$fitted.values # valores predichos
datos$residuos <- modelo_lineal$residuals    # residuales
head(datos)
```

```
##    X  Y prediccion  residuos
## 1 16 10   12.80481 -2.804811
## 2 32 15   17.90682 -2.906824
## 3 48 20   23.00884 -3.008836
## 4 56 22   25.55984 -3.559843
## 5 64 30   28.11085  1.889151
## 6 80 32   33.21286 -1.212861
```

g) Verificar los supuestos del modelo lineal

Linealidad. Se calculan los residuos para cada observación y se representan gráficamente (scatterplot). Si las observaciones siguen la línea del modelo, los residuos se deben distribuir aleatoriamente entorno al valor 0 (Amat, 2016).

```
# linealidad
ggplot(data = datos, aes(x = prediccion, y = residuos)) +
  geom_point(aes( color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_hline(yintercept = 0) +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  labs(title = "Distribucion de los residuos", x = "prediccion
modelo",
       y = "residuo") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position =
"none")
```

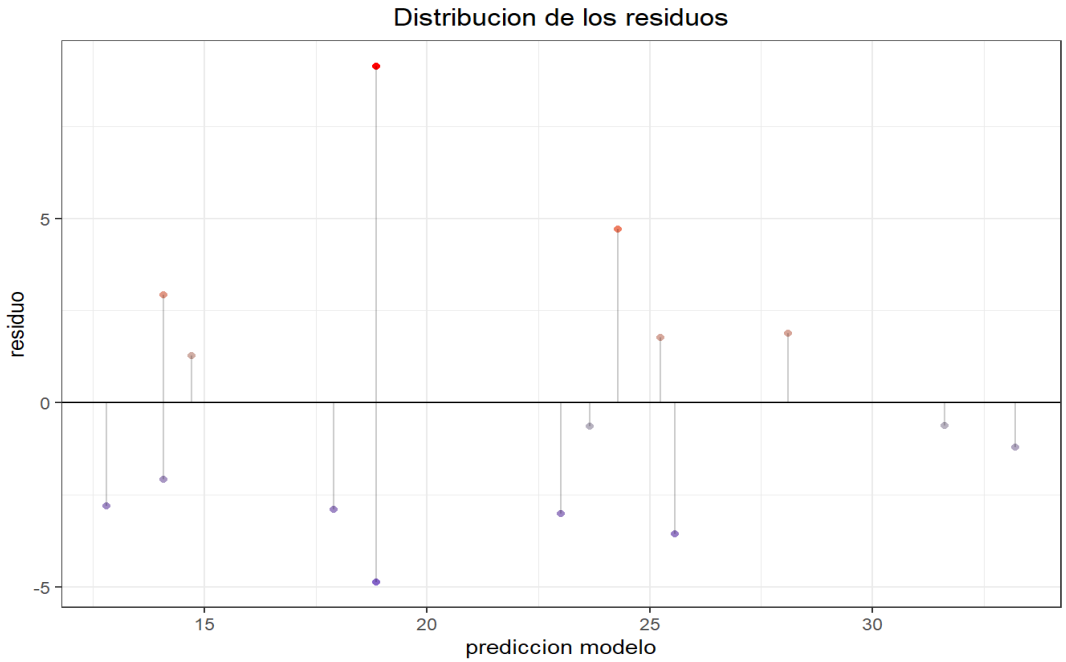


Figura 13. *Distribución de los residuos de modelo.*

Los residuos se distribuyen de forma aleatoria entorno al 0 por lo que se acepta la linealidad.

Normalidad de los residuos:

Los residuos se deben distribuir de forma normal con media 0. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a un test de contraste de normalidad.

Histograma de residuos.

```
# Distribución normal de los residuos:
ggplot(data = datos, aes(x = residuos)) +
  geom_histogram(aes(y = ..density..)) +
  labs(title = "Histograma de los residuos") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

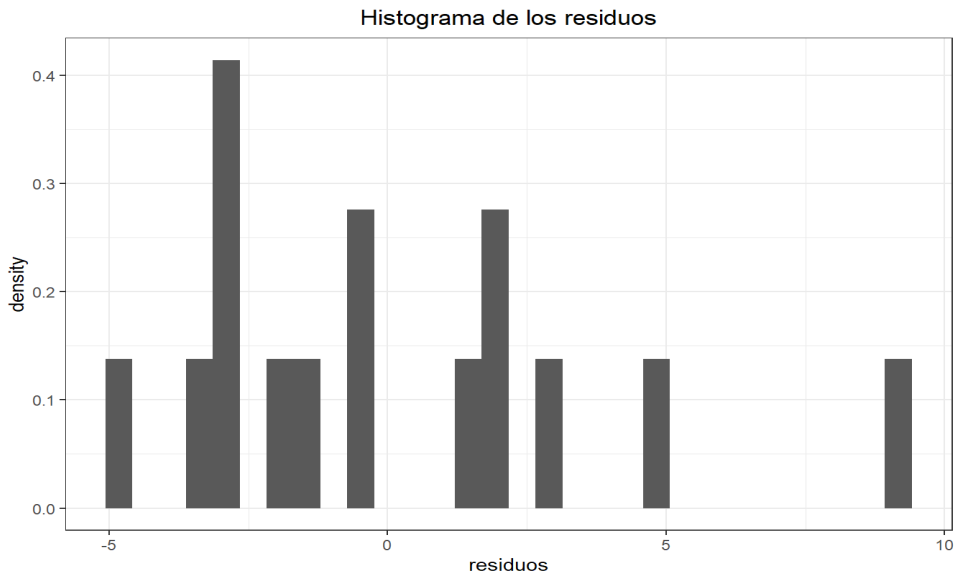


Figura 14. *Histograma de densidad de los residuos*

No se evidencia una clara distribución normal de los residuos.

Gráfico de cuantiles.

```
# grafico de cuantiles
qqnorm(modelo_lineal$residuals)
qqline(modelo_lineal$residuals)
```

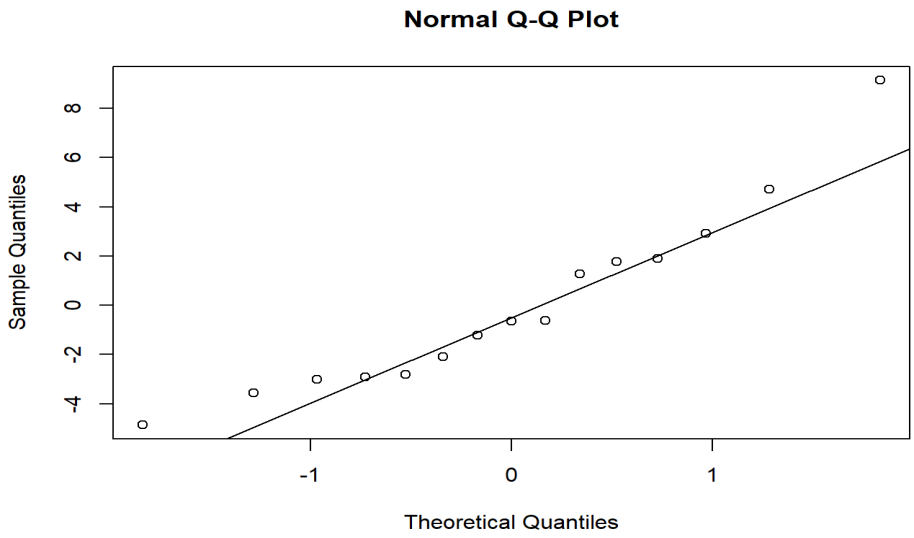


Figura 15. *Gráfico de cuantiles de los residuos*

Los puntos no se muestran alrededor de la línea, algunos se muestran alejados, posible no normalidad en los errores.

Finalmente realizamos una prueba de hipótesis de normalidad, por la cantidad de datos, el estadístico más adecuado sería el test de Shapiro.

```
# test de normalidad
shapiro.test(modelo_lineal$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  modelo_lineal$residuals
## W = 0.92214, p-value = 0.2077
```

1) Prueba de hipótesis

   Ho: existe normalidad en los residuos

   Ha: no existe normalidad en los residuos

2) nivel de significancia 0.05

3) Prueba estadística. Shapiro y Wilks

W = 0.92214 con p=0.2077

4) decisión: $p(0.2077) > \alpha(0.05)$, se acepta la Ho de normalidad de los residuos.

Podemos obtener también el test de kolmogorov (muestras grandes).

```
# Kolmogorov test
ks.test(modelo_lineal$residuals, "pnorm",
        mean = mean(modelo_lineal$residuals),
        sd = sd(modelo_lineal$residuals))
```

```
##  One-sample Kolmogorov-Smirnov test
##
## data:  modelo_lineal$residuals
## D = 0.16652, p-value = 0.7405
## alternative hypothesis: two-sided
```

también muestra normalidad de los residuos $p(0.7405) > \alpha(0.05)$.

Varianza constante de los residuos (Homocedasticidad):

La variabilidad de los residuos debe de ser constante a lo largo del eje X.

```
# Varianza constante de los residuos (Homocedasticidad):
ggplot(data = datos, aes(x = prediccion, y = residuos)) +
geom_point(aes( color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  geom_smooth(se = FALSE, color = "firebrick") +
  labs(title = "Distribucion de los residuos", x = "prediccion
modelo", y = "residuo") +
  geom_hline(yintercept = 0) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position =
"none")
```

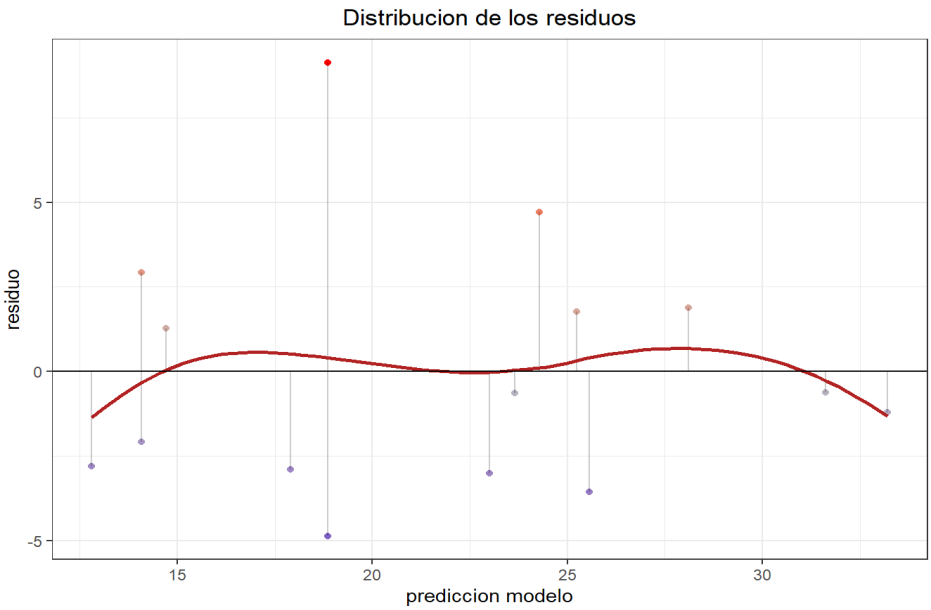


Figura 16. *Distribución de los residuos.*

Los residuos se comportan de manera aleatoria, no existe un patrón, existe homocedasticidad. Verifiquemos con un test de hipótesis.

Prueba de hipótesis de Breush Pagan

```
# Test de Breush-Pagan
library(lmtest)
bptest(modelo_lineal)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo_lineal
## BP = 0.5092, df = 1, p-value = 0.4755
```

1) Prueba de hipótesis

   Ho: existe homocedasticidad.

   Ha: falta de homocedasticidad

2) nivel de significancia 0.05

3) Prueba estadística. Breuch Pagan

BP = 0.5092 con p=0.4755

4) decisión: $p(0.4755) > \alpha(0.05)$, se acepta la Ho de existencia de homocedasticidad.

Ni la representación gráfica ni el contraste de hipótesis muestran evidencias que haga sospechar falta de homocedasticidad.

Autocorrelación de residuos:

Cuando se trabaja con intervalos de tiempo, es muy importante comprobar que no existe autocorrelación de los residuos, es decir que son independientes. Esto puede hacerse detectando visualmente patrones en la distribución de los residuos cuando se ordenan según han registrado o con el test de Durbin-Watson dwt() del paquete Car (Amat, 2016).

```
# Autocorrelación de residuos:
ggplot(data = datos, aes(x = seq_along(residuos), y = residuos)) +
geom_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
geom_line(size = 0.3) +
  labs(title = "Distribución de los residuos", x = "index", y =
"residuo")+
  geom_hline(yintercept = 0) +
  theme(plot.title = element_text(hjust = 0.5), legend.position =
"none")
```

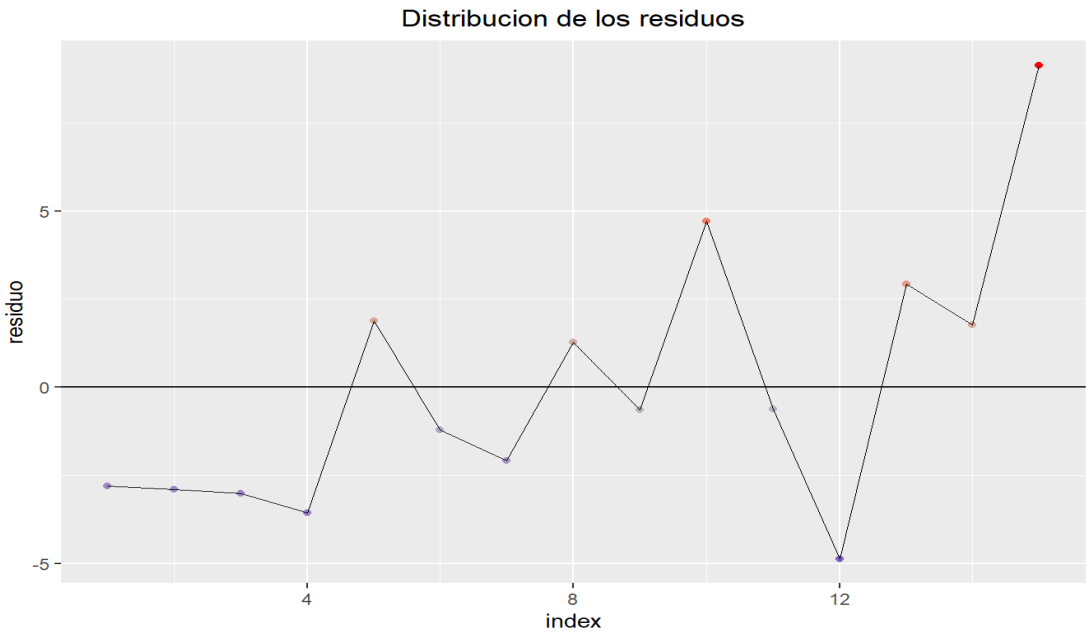


Figura 17. *Distribución de los residuos.*

En este caso, la representación de los residuos no muestra ninguna tendencia, indicaría la no existencia de autocorrelación.

Test de Durbin Watson

```
#test de Durwin Watson
library(lmtest)
dwtest(modelo_lineal)
```

```
##  Durbin-Watson test
##
## data:  modelo_lineal
## DW = 1.2942, p-value = 0.06938
## alternative hypothesis: true autocorrelation is greater than 0
```

1) Prueba de hipótesis

   Ho: no existe autocorrelación.

   Ha: existe autocorrelación

2) nivel de significancia 0.05

3) Prueba estadística. Durwin Watson

DW = 1.2942 con p=0.06938

4) decisión: $p(0.06938) > \alpha(0.05)$, se acepta la Ho de no existencia de autocorrelación.

h) Identificación de valores atípicos: *outliers*, *leverage* y observaciones influyentes.

- **Outlier u observación atípica:** Observaciones que no se ajustan bien al modelo. El valor real se aleja mucho del valor predicho, por lo que su residuo es excesivamente grande. En una representación bidimensional se corresponde con desviaciones en el eje Y.
- **Observación influyente:** Observación que influye sustancialmente en el modelo, su exclusión afecta al ajuste. No todos los *outliers* tienen por qué ser influyentes.
- **Observación con alto leverage:** Observación con un valor extremo para alguno de los predictores. En una representación bidimensional se corresponde con desviaciones en el eje X. Son potencialmente puntos influyentes.

Independientemente de que el modelo se haya podido aceptar, siempre es conveniente identificar si hay algún posible *outlier*, *observación con alto leverage* u observación altamente influyente, puesto que podría estar condicionando en gran medida el modelo. La eliminación de este tipo de observaciones debe de analizarse con detalle y dependiendo de la finalidad del modelo. Si el fin es predictivo, un modelo sin estas observaciones puede lograr mayor precisión en la mayoría de casos. Sin embargo, es muy importante prestar atención a estos valores ya que, de no ser errores de medida, pueden ser los casos más interesantes. El modo adecuado a proceder cuando se sospecha de algún posible valor atípico o influyente es calcular el modelo de regresión excluyendo dicho valor.

Test de Benferroni para detectar Outliers.

```
# Prueba de Benferroni para detectar outliers
```

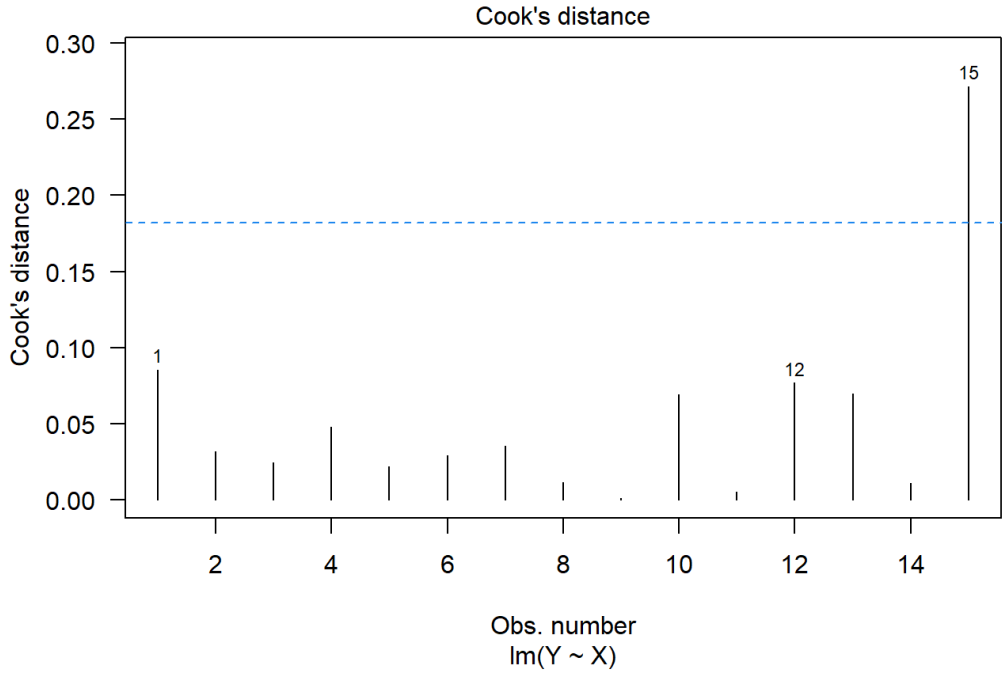
```
library(car)
outlierTest(modelo_lineal, cutoff=Inf, n.max=4)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 15  3.3004752          0.0063352     0.095028
## 12 -1.3674927          0.1965400           NA
## 10  1.3173787          0.2123100           NA
## 4  -0.9728475          0.3498300           NA
```

El posible valor atípico ubicado en la posición 15, no es significativo como outlier y mucho menos los otros valores (12,10,4).

Otros procedimientos.

```
# Distancia de Cook, detección de valores influyentes
cutoff <- 4 / (26-2-2)  # Cota
plot(modelo_lineal, which=4, cook.levels=cutoff, las=1)
abline(h=cutoff, lty="dashed", col="dodgerblue2")
```




```
# Identificación de valores atípicos: outliers, leverage y
observaciones influyentes
library(ggrepel)
library(dplyr)
```

```
datos$studentized_residual <- rstudent(modelo_lineal)
ggplot(data = datos,aes(x = prediccion, y =
abs(studentized_residual))) +
  geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +
  # Se identifican en rojo residuos studentized absolutos > 3
  geom_point(aes(color = ifelse(abs(studentized_residual) > 3, "red",
"black"))) +
  scale_color_identity() +
  #se muestra el equipo al que pertenece la observacion atipica,
  geom_text_repel(data = filter(datos, abs(studentized_residual) > 3),
                  aes(label ="" )) +
  labs(title = "Distribucion de los residuos studentized",x =
"prediccion modelo") +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5), legend.position =
"none")
```

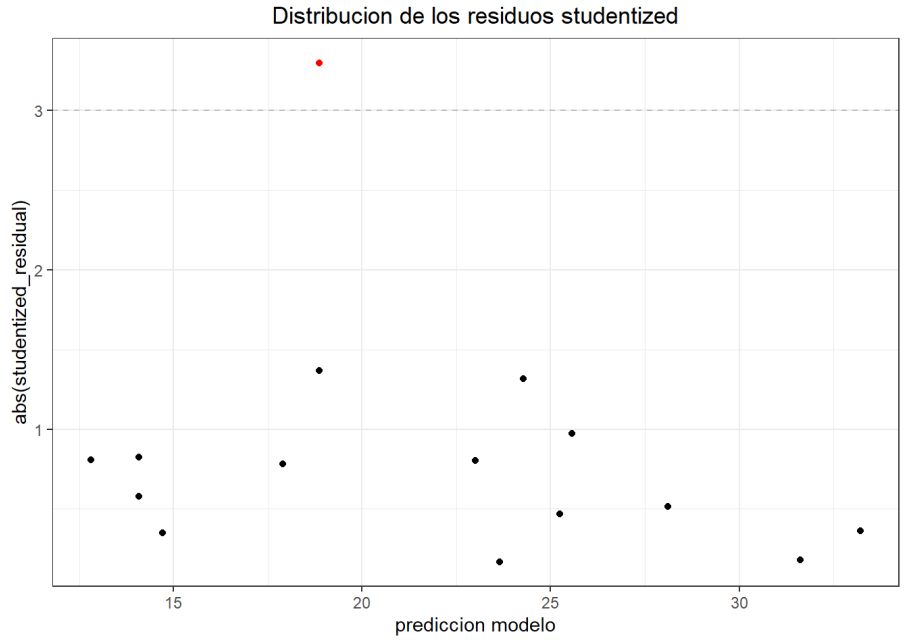


Figura 18. *Distribución de los residuos estudiantizados para observar valores outliers*

Se observa un punto por encima de 3. ¿Cuál es el valor reconocido como outliers?

```
datos %>% filter(abs(studentized_residual) > 3)
```

```
##    X  Y prediccion residuos studentized_residual
## 1 35 28   18.86345 9.136549             3.300475
```

```
which(abs(datos$studentized_residual) > 3)
```

```
## [1] 15
```

El estudio de los residuos *studentized* identifica al dato 15 con valores (35 y 28), como una posible observación atípica.

El hecho de que un valor sea atípico o con alto grado de *leverage* no implica que sea influyente en el conjunto del modelo. Sin embargo, si un valor es influyente, suele ser o atípico o de alto *leverage*. En R se dispone de la función outlierTest() del paquete car y de las funciones influence.measures(), influencePlot() y hatvalues() para identificar las observaciones más influyentes en el modelo.

```
library(car)
summary(influence.measures(model = modelo_lineal))
```

```
## Potentially influential observations of
##   lm(formula = Y ~ X, data = datos) :
##
##    dfb.1_ dfb.X dffit cov.r   cook.d hat
## 6   0.14  -0.20 -0.23  1.62_*  0.03   0.29
## 11  0.05  -0.08 -0.10  1.52_*  0.01   0.23
## 15  0.73  -0.41  0.98  0.35_*  0.27   0.08
```

No se detectan valores influyentes ni atípicos significativos que puedan modificar el modelo, pero podemos identificar gráficamente cuales eran esos posibles valores.

```
influencePlot(model = modelo_lineal)
```

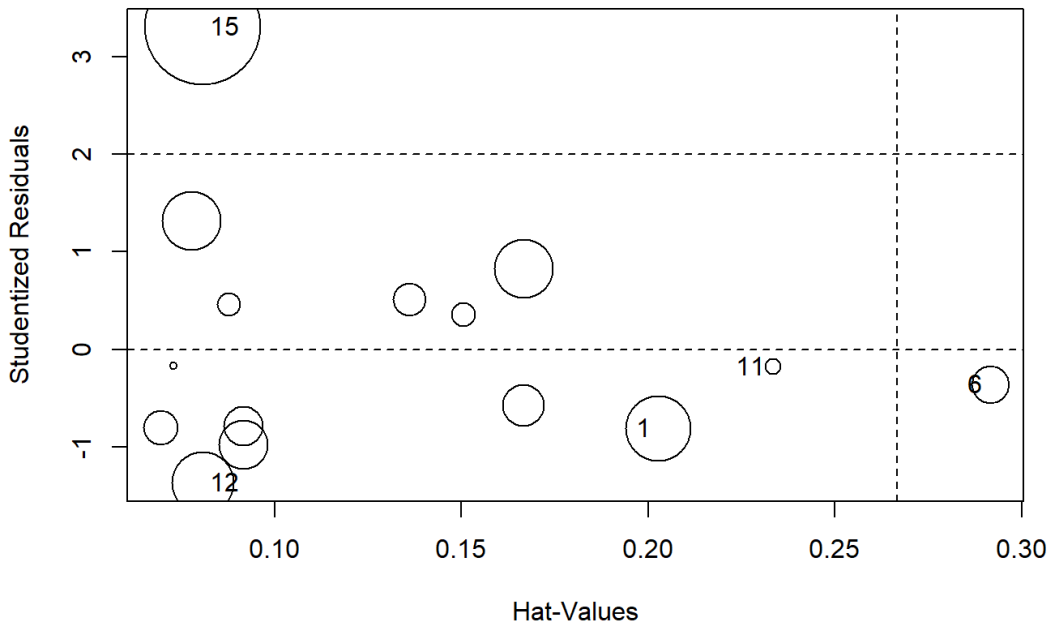


Figura 19. *Posibles valores influyentes*

```
##        StudRes        Hat       CookD
## 1   -0.8088578 0.20268332 0.085429608
## 6   -0.3633091 0.29151053 0.029097425
## 11 -0.1773640 0.23339116 0.005174119
## 12 -1.3674927 0.08071941 0.076951250
## 15  3.3004752 0.08071941 0.271575682
```

i) Prediciendo nuevos valores.

El modelo obtenido cumple con los supuestos y no presenta valores atípicos o influyentes:

$$Y = 7.70280 + 0.31888(X)$$

Tanto el intercepto como la variable X (gastos en publicidad) son estadísticamente significativas al 5%, en ambos casos $[p(0.0074) < \alpha(0.05) y\ p(0.0000266) < \alpha(0.05)]$, con tendencia positiva. El modelo tiene un coeficiente de determinación de 75.44% (el modelo calculado explica el 75.44% de la variabilidad presente en la variable respuesta

(volumen de ventas) mediante la variable independiente (gastos en publicidad) y un ANVA significativo.

¿Cuánto será el volumen de ventas, si se invierte 63 mil pesetas en publicidad?

```
# prediciendo nuevos valores, cuando X = 63
predict_value <- predict(modelo_lineal, data.frame(X= c(63)))
predict_value
```

```
##        1
## 27.79197
```

Si se invierten 63 mil pesetas en publicidad, el volumen de ventas será de 27.79 millones de pesetas.

Por cada unidad (mil pesetas) que se invierte en gastos de publicidad, el volumen de ventas se incrementa en promedio 0.3188 millones de pesetas.

11. Ejemplo manual para mostrar los cálculos de las Bandas de Confianza:

En la producción de herramientas la deformación del acero a cierta temperatura puede afectar la dureza del acero, para investigar esta relación se ha tomado la siguiente muestra.

Deformación: mm²(X)	6	9	11	13	22	26	28	33	35
Dureza en Kg/mm²(Y)	68	67	65	53	44	40	37	34	32

Las sumas necesarias son:

$\sum X_i = 183 \quad \sum Y_i = 440 \quad \sum X_i^2 = 4665 \quad \sum Y_i^2 = 23232 \quad \sum X_i Y_i = 7701$

$\overline{X} = 20.33 \quad \overline{Y} = 48.89$

1) Diagrama de dispersión

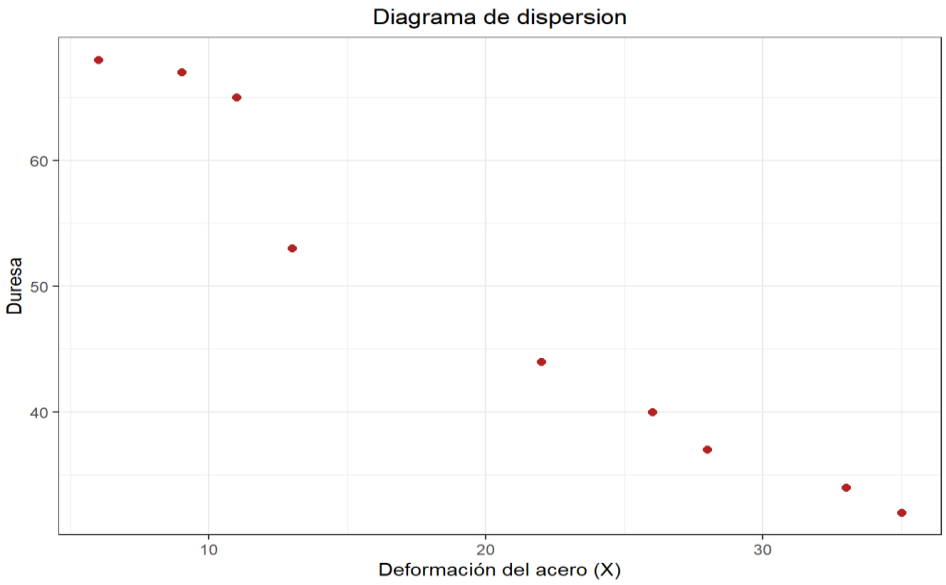


Figura 20. *Diagrama de dispersión*

2) Obteniendo parámetros

Transformado los datos en desviaciones:

$$\sum x_i^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = 4665 - \frac{183^2}{9} = 944$$

$$\sum x_i y_i = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n} = 7701 - \frac{183(440)}{9} = -1245.67$$

Calculando los parámetros

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{-1245.67}{944} = -1.32$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 48.89 - (-1.32)(20.33) = 75.73$$

La recta de regresión mínimo cuadrática es:

$$\hat{Y} = 75.73 - 1.32X_i$$

3) Recta de regresión lineal simple

Usando puntos extremos para ajustar la recta:

$$\hat{Y}_1 = 75.73 - 1.32(6) = 67.81$$

$$\hat{Y}_9 = 75.73 - 1.32(35) = 29.53$$

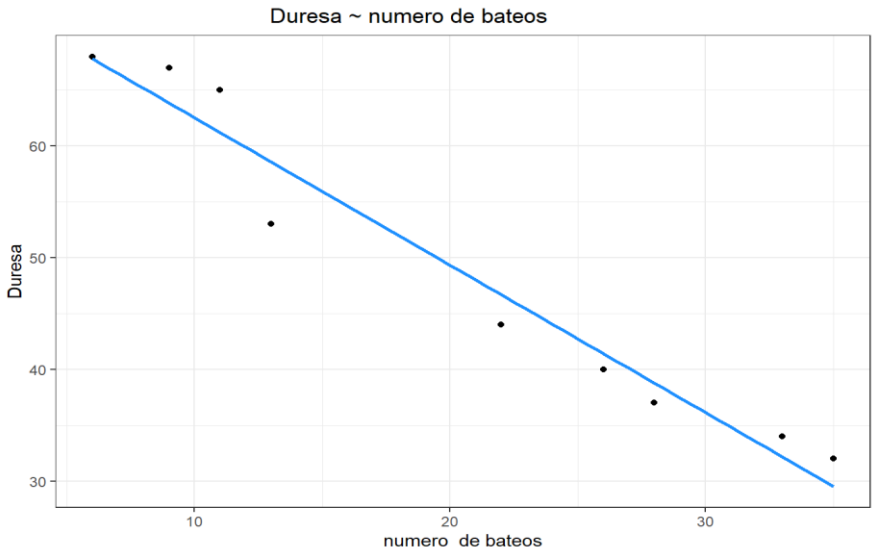


Figura 21. *Modelo de rls*

4) Intervalo de confianza

Intervalo de confianza para el parámetro β

$$p(\hat{\beta}_1 - t_{(n-2),\alpha/2}S_b < \beta_1 < \hat{\beta}_1 + t_{(n-2),\alpha/2}S_b) = 1 - \alpha$$

$$p(-1.32 - 2.365(0.1073) < \beta_1 < -1.32 + 2.365(0.1073)) = 1 - 0.05$$

$$p(-1.57 < \beta_1 < -1.07) = 0.95$$

Donde:

$$S_e^2 = \frac{\sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum X_i Y_i}{n-2} = \frac{23232 - (75.73)(440) - (-1.32)(7701)}{9-2}$$

$$= 10.87$$

$$S_e = \sqrt{10.87} = 3.297$$

$$Var(\hat{\beta}_1) = S_{\hat{\beta}_1}^2 = \frac{S_e^2}{\sum x_i^2} = \frac{10.87}{944} = 0.0115$$

entonces: $S_b^2 = \sqrt{0.0115} = 0.1073$

Por otra fórmula:

$$S_b = \frac{S_e}{\sqrt{\sum x_i^2}} = \frac{3.297}{\sqrt{944}} = 0.1073$$

$$t_t = t_{(n-2),\alpha/2} = t_{7,0.025} = 2.365$$

Intervalo de confianza para el parámetro *$Y = a + bX$ pronosticación (E(Y/X))*

$$p\left[\hat{Y}_p - t_{\alpha/2(n-2)} S_{\hat{Y}_p} < E(Y/X) < \hat{Y}_p + t_{\alpha/2(n-2)} S_{\hat{Y}_p}\right] = 1 - \alpha$$

$$t_t = t_{(n-2),\alpha/2} = t_{7,0.025} = 2.365$$

$$S_{\hat{Y}_p} = S_e \sqrt{\frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum x_i^2}} = 3.297 \sqrt{\frac{1}{9} + \frac{(X_p - 20.33)^2}{944}} = A_o$$

$$p\left[\hat{Y}_p - 2.365 A_o < E(Y/X) < \hat{Y}_p + 2.365 A_o\right] = 1 - \alpha$$

Para X_p=6:

$$p\left[67.81 \pm 2.365((3.297)\sqrt{\frac{1}{9} + \frac{(X_p - 20.33)^2}{944}}) < E(Y/X) <)\right] = 0.95$$

$$p[63.34 < E(Y/X) < 72.28] = 0.95$$

Zona de confianza para E(Y/X)

Cuanto más se desvía X del promedio, los valores tabulares son mayores y por lo tanto los intervalos son más amplios.

Intervalo de confianza para el parámetro *Y predicción*

$$p\left[\hat{Y}_p \pm t_{\alpha/2(n-2)} S_{\hat{Y}_p - Y_p} < E(Y/X) <\right] = 1 - \alpha$$

$$t_t = t_{(n-2),\alpha/2} = t_{7,0.025} = 2.365$$

$$S_{\hat{Y}_p - Y_p} = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum x_i^2}} = 3.297 \sqrt{1 + \frac{1}{9} + \frac{(X_p - 20.33)^2}{944}} = A_1$$

Reemplazando los valores de X, en la tabla

Se forma la siguiente tabla

Tabla 5.

Xi	$\hat{Y}_p$	$2.365A_o$	$\hat{Y}_p - 2.365A_o$	$\hat{Y}_p + 2.365A_o$	$\hat{Y}_p - 2.365A_1$	$\hat{Y}_p + 2.365A_1$
6	67.81	4.47	63.34	72.28	58.82	76.80
9	63.85	3.88	59.97	67.73	55.14	72.56
11	61.21	3.52	57.69	64.73	52.66	69.76
13	58.57	3.20	55.37	61.77	50.14	67.00
20.33	48.89	2.60	46.29	51.49	40.67	57.11
22	46.69	2.63	44.06	49.32	38.46	54.92
26	41.41	2.97	38.44	44.38	33.07	49.75
28	38.77	3.25	35.52	42.02	30.32	47.22
33	32.17	4.13	28.04	36.30	23.34	41.00
35	29.53	4.54	24.99	37.07	20.51	38.55

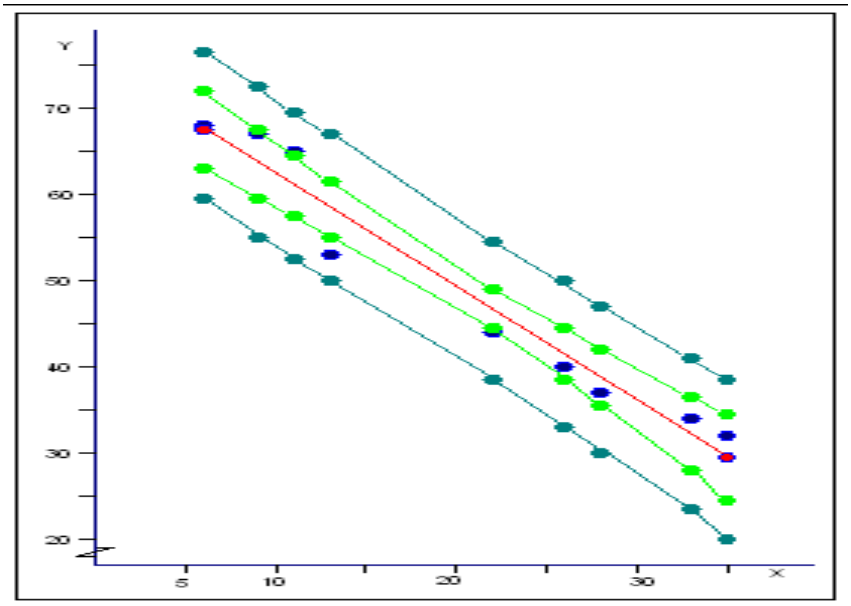


Figura 22. *Zona de confianza para Y_o*

Los limites pueden usarse para estimar el intervalo correspondiente a un X dado; ejemplo: estimar el valor de Y para X=18

$$\hat{Y} = 75.73 - 1.32(18) = 51.97$$

Hallando el intervalo de predicción:

$$p[51.97 \pm 2.365(3.48) < Y_o) <] = 0.95$$

$$p[43.74 \ , \ 60.20] = 0.05.$$

12. Ejemplo 2 en R.

Análisis de Regresión lineal simple con valor atípico.

Un analista de deportes quiere saber si existe una relación lineal entre el número de bateos que realiza un equipo de béisbol y el número de corridas que consigue. En caso de existir y de establecer un modelo, podría predecir el resultado del partido (Amat, 2016).

Ingreso de datos (también pueden ser importados de una base de datos)

```
equipos <- c("Texas","Boston","Detroit","Kansas","St.", "New_S.",
             "Ne w_Y.","Milwaukee","Colorado","Houston","Baltimore",
             "Los_An.","Chica go"
"Cincinnati","Los_P.","Philadelphia",
             "Chicago","Cleveland","Ari zona","Toronto","Minnesota",
             "Florida","Pittsburgh","Oakland","Tampa","Atlanta",
             "Washington”,"San.F","San.I","Seattle")
numero_bateos <- c(5659,5710,5563,5672,5532,5600,5518,5447,5544,5598,
                   5585,5436,5549,5612,5513,5579,5502,5509,5421,5559,
                   5487,5508,5421,5452,5436,5528,5441,5486,5417,5421)
corridas <- c(855,875,787,730,762,718,967,721,735,615,708,644,654,735,
          667,713,654,704,731,743,619,625,610,645,707,641,624,570,
          593,556)
```

Creando un dada.frame y observando 6 datos

```
datos <- data.frame(equipos, numero_bateos, corridas)
head(datos)
```

```
##   equipos numero_bateos corridas
## 1   Texas          5659      855
## 2  Boston          5710      875
## 3 Detroit          5563      787
## 4  Kansas          5672      730
## 5     St.          5532      762
## 6  New_S.          5600      718
```

La estructura (str) sirve para observar la estructura de ingreso de la data (numéricas, o cualitativas)

```
str(datos)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ equipos      : chr  "Texas" "Boston" "Detroit" "Kansas" ...
##  $ numero_bateos: num  5659 5710 5563 5672 5532 ...
##  $ corridas     : num  855 875 787 730 762 718 967 721 735 615 ...
```

Podemos ver que la data set está conformada por 30 observaciones y 3 variables:

- equipos está reconocido como variable cualitativa(chr)
- numero _bateos, como variable numérica
- runs, como variable numérica.

El número de corridas (Y) estará en función al número de bateos (X), dicho de otra manera, en número de bateos influye en el número de corridas.

a) Observar las variables descriptivamente.

```
# histograma para las variables
require(ggplot2)
par(mfrow = c(1, 2))
hist(datos$numero_bateos, breaks = 10, main = "", xlab = "número de ba
teos",ylab="Frecuencia", border = "darkred")
hist(datos$corridas, breaks = 10, main = "", xlab = "corridas",ylab="F
recuencia", border = "blue")
```

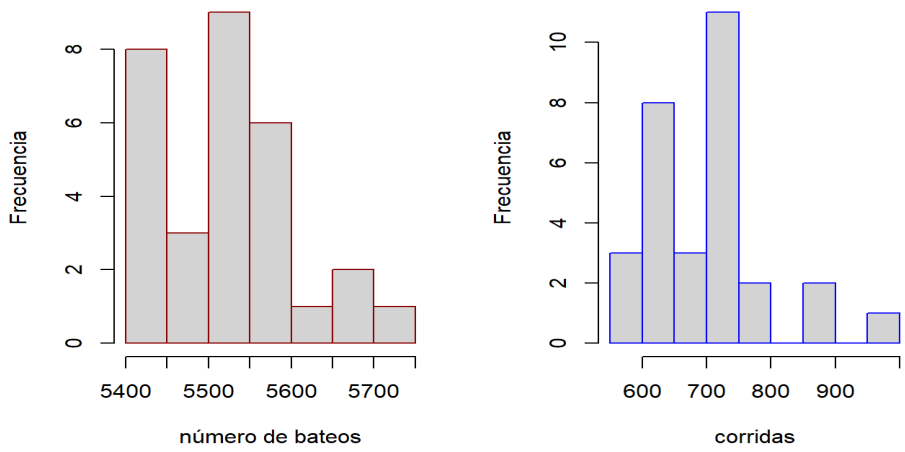


Figura 23. *Histograma de frecuencias para el número de bateos y numero de corridas.*

```
summary(datos)
```

```
##    equipos           numero_bateos     corridas
##  Length:30          Min.   :5417   Min.   :556.0
##  Class :character   1st Qu.:5448   1st Qu.:629.0
##  Mode  :character   Median :5516   Median :705.5
##                     Mean   :5524   Mean   :696.9
##                     3rd Qu.:5575   3rd Qu.:734.0
##                     Max.   :5710   Max.   :967.0
```

El promedio del número de bateos es 5524 con un mínimo de 5417 y máximo de 5710. El promedio de corridas es 697 aproximadamente con un mínimo de corridas de 556 y máximo 967. En promedio, para realizar 697 corridas se necesita 5524 bateos.

b) Representación gráfica de las observaciones

El primer paso antes de generar un modelo de regresión es representar los datos gráficamente para poder intuir si existe una relación lineal y cuantificar mediante un coeficiente de correlación. Si en este paso no se detecta la posible relación lineal, no tiene sentido seguir adelante

generando un modelo lineal (se tendrían que probar otros modelos no lineales).

Usando la librería ggplot2

```
require(ggplot2)
ggplot(data = datos, mapping = aes(x = numero_bateos, y= corridas))
  +  geom_point(color = "firebrick", size = 3) +
  (labs(title = "Diagrama de dispersión", x = "número  de bateos")) +
  theme_bw() +
  theme(plot.title = element_text(hjust =0.5))
```

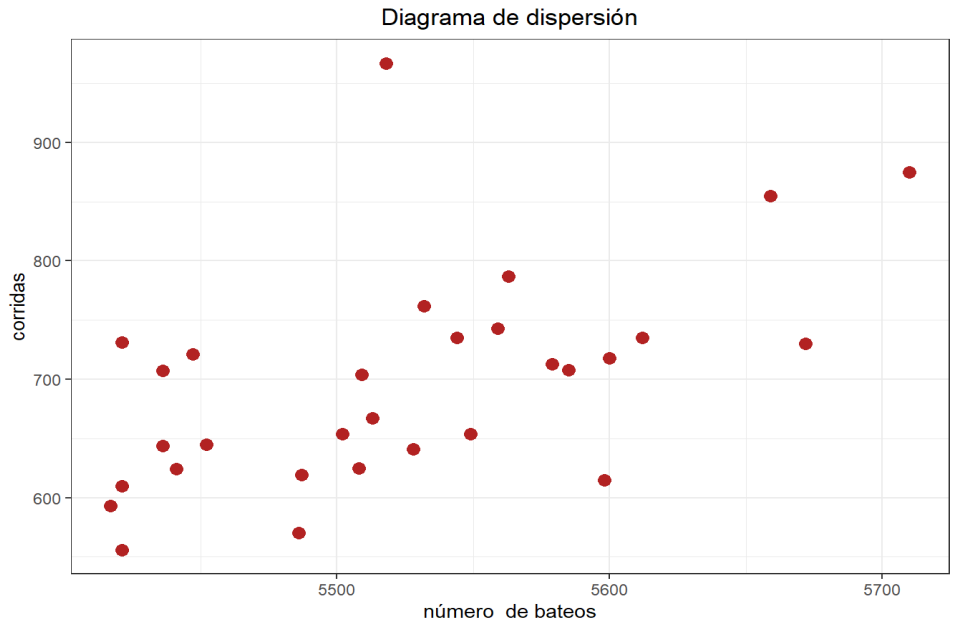


Figura 24. *Diagrama de dispersión*

La figura muestra un comportamiento (tendencia) positivo, a más bateos más corridas, así mismo, muestra posible outlier y puntos influyentes.

c) comportamiento del coeficiente de correlación

Estadístico de correlación de Pearson (para datos cuantitativos)

```
cor.test(x = datos$numero_bateos, y = datos$corridas, method =
"pearson")
```

```
##  Pearson's product-moment correlation
##
## data:  datos$numero_bateos and datos$corridas
## t = 3.477, df = 28, p-value = 0.001673
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2354692 0.7592163
## sample estimates:
##       cor
## 0.5491526
```

Resumen:

Coeficiente de correlación	Test de significancia	Intervalo de confianza de r	p-value
r = 0.549	t = 3.477	0.2354692 - 0.7592163	0.001673

El test de correlación muestra una relación lineal significativa $p(0.001673) < \alpha(0.05)$, de intensidad considerable (r = 0.549). Tiene sentido intentar generar un modelo de regresión lineal que permita predecir el número de corridas en función del número de bateos.

d) Cálculo del modelo de regresión lineal simple

```
# Cálculo del modelo de regresión lineal simple
modelo_lineal <- lm(corridas ~ numero_bateos, data=datos)
summary(modelo_lineal)
```

```
## Call:
## lm(formula = corridas ~ numero_bateos, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -128.69  -50.54  -19.96   51.10  273.52
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2769.4894   997.0462  -2.778  0.00966 **
## numero_bateos     0.6276     0.1805   3.477  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 77.63 on 28 degrees of freedom
## Multiple R-squared:  0.3016, Adjusted R-squared:  0.2766
## F-statistic: 12.09 on 1 and 28 DF,  p-value: 0.001673
```

El modelo de regresión lineal simple es:

$$Y = -2769.4894 + 0.6276(numero\_bateos)$$

Por cada unidad que se incrementa el número de bateos, el número de corridas aumenta en promedio 0.6276 unidades.

Realizando la inferencia de los parámetros con la prueba t, ambos son significativos. $p < \alpha$, es decir, que tienen importancia en el modelo individualmente.

El coeficiente de determinación R^2 indica que el modelo calculado explica el 30.16 de la variabilidad presente en la variable respuesta (corridas) mediante la variable independiente (*número de bateos*). Podría indicarse un ajuste no muy bueno.

La prueba de Análisis de varianza observa al modelo en su conjunto. El *p-value* obtenido en el test $p(0.001673) < \alpha(0.05)$ determina que es significativamente superior la varianza explicada por el modelo en comparación a la varianza total. Es el parámetro que determina que el modelo en su conjunto es significativo.

Observando en detalle en ANVA

```
anova(modelo_lineal)
```

```
## Analysis of Variance Table
##
## Response: corridas
##                Df Sum Sq Mean Sq F value    Pr(>F)
## numero_bateos   1  72867   72867   12.09 0.001673 **
## Residuals      28 168761    6027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

Hasta aquí se puede concluir que el modelo es bueno y existe relación entre las variables (excepto que el ajuste R^2 es bajo).

e) Intervalos de confianza para los parámetros del modelo

```
# Intervalos de confianza para los parámetros del modelo
confint(modelo_lineal, level=0.95)
```

```
##                        2.5 %       97.5 %
## (Intercept)    -4811.8458919 -727.1328213
## numero_bateos      0.2578569    0.9972975
```

Los intervalos de confianza contienen a los coeficientes obtenidos.

Por cada unidad que se incrementa el número de bateos, el número de corridas aumenta en promedio entre 0.2579 y 0.9973 unidades.

f) Representación gráfica del modelo

```
# Representación gráfica del modelo
ggplot(data = datos, mapping = aes(x = numero_bateos, y= corridas))
+
  geom_point(size=3) +
  labs(title = "corridas~número de bateos", x="número  de bateos")
+
```

```
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.4))
```

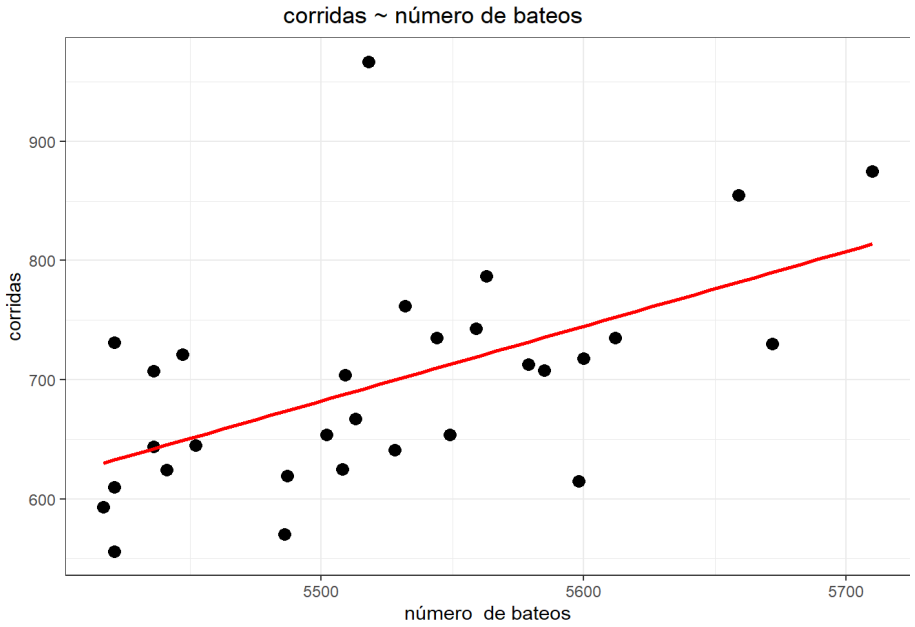


Figura 25. *Representación de la línea de regresión*

```
names(modelo_lineal)
```

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

Además de la línea de mínimos cuadrados es recomendable incluir los límites superior e inferior del intervalo de confianza. Esto permite identificar la región en la que, según el modelo generado y para un determinado nivel de confianza, se encuentra el valor promedio de la variable dependiente.

Para poder representar el intervalo de confianza a lo largo de todo el modelo se recurre a la función predict() para predecir valores que

abarquen todo el eje *X*. Se añaden al gráfico líneas formadas por los límites superiores e inferiores calculados para cada predicción (Amat, 2016).

```
# predecir valores para Y con valores de X originales
predict_value <- predict(modelo_lineal, datos)
head(predict_value)
```

```
##        1        2        3        4        5        6
## 781.9700 813.9765 721.7226 790.1285 702.2677 744.9430
```

Calculando el error medio cuadrático (RMSE)

```
# MSE: ERROR MEDIO CUADRATICO
# RMSE: raiz del error cuadratico medio
par(mfrow = c(1, 1))
error = predict_value - datos$corridas
head(error)
```

```
##         1         2         3         4         5         6
## -73.02996 -61.02352 -65.27737  60.12855 -59.73226  26.94299
```

Estos resultados también se encuentran en residuales del modelo.

```
MSE <- mean(modelo_lineal$residuals^2)
MSE
```

```
## [1] 5625.35
```

```
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 75.00233
```

```
 Cálculo del error cuadrático medio RMSE con los residuales del
modelo.
sqrt(mean(modelo_lineal$residuals^2))
```

```
## [1] 75.00233
```

Bandas de confianza para valores medios.

```
# gráfico de bandas de confianza (una banda). valores medios
par(mfrow = c(1, 1))
puntos <- seq(from = min(datos$numero_bateos),
              to = max(datos$numero_bateos), length.out = 100)
limites_intervalo <- predict(object = modelo_lineal,
                             newdata = data.frame( numero_bateos
= puntos),
                             interval = "confidence", level =
0.95)
head(limites_intervalo, 3)
```

```
##         fit      lwr      upr
## 1 630.0964 581.1740 679.0188
## 2 631.9537 583.9076 679.9998
## 3 633.8111 586.6322 680.9900
```

```
plot(datos$numero_bateos, datos$corridas, col = "firebrick", pch = 19,
     ylab = "corridas", xlab = "número de bateos",
     main = "Banda de confianza para valores medios")
abline(modelo_lineal, col = 1)
lines(x = puntos, y = limites_intervalo[, 2], type = "l", col = 2, lty =
3)
lines(x = puntos, y = limites_intervalo[, 3], type = "l", col = 3, lty =
3)
```

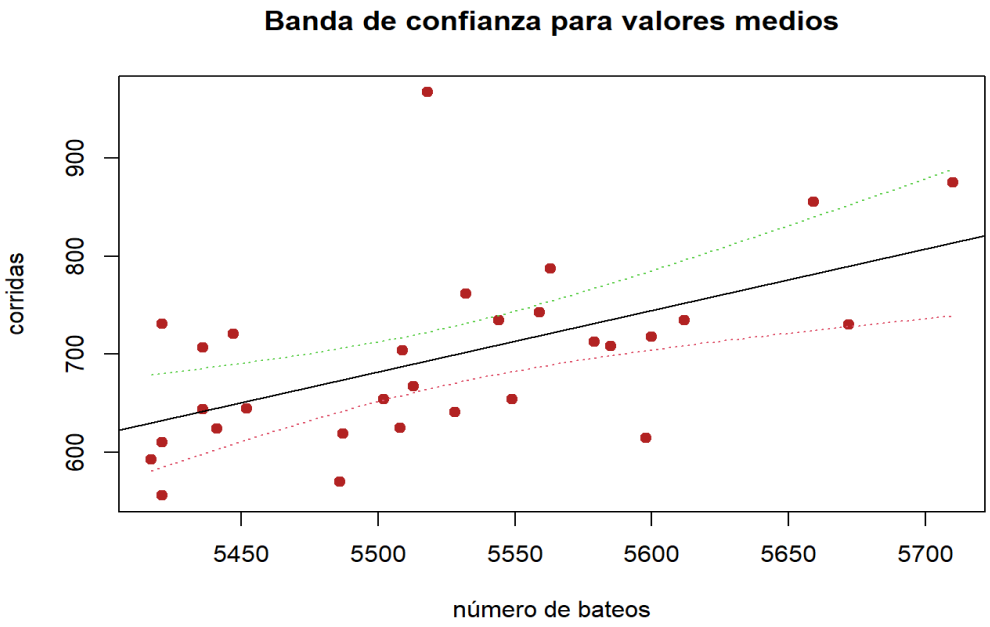


Figura 26. *Bandas de confianza para el ajuste de la regresión.*

La función geom_smooth() del paquete ggplot2 genera la regresión y su intervalo de forma directa.

```
# con ggplot2
ggplot(data = datos, mapping = aes(x = numero_bateos, y = corridas))
+
  geom_point(color = "firebrick", size = 2) +
  labs(title = "Diagrama de dispersion", x = "numero  de bateos") +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

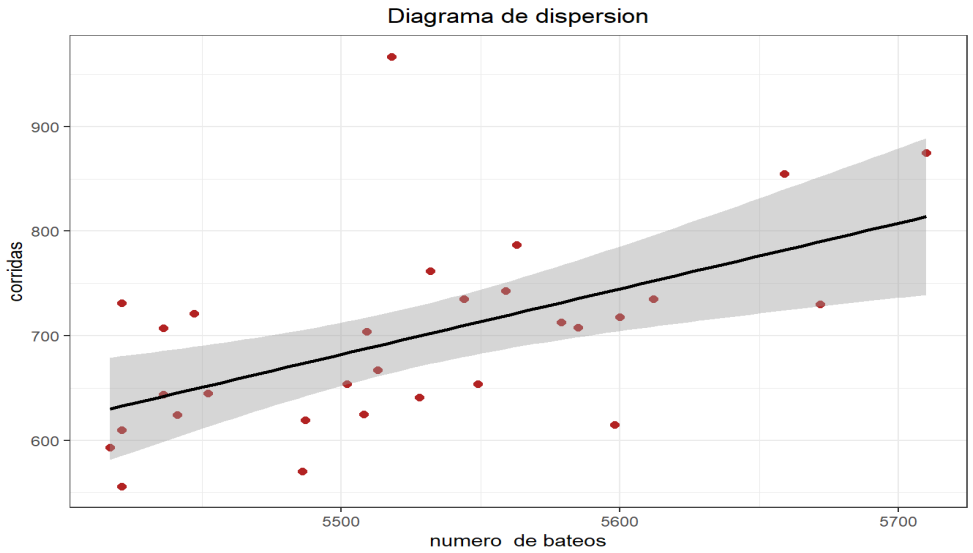


Figura 27. *Bandas de confianza para la media y la predicción.*

```
# Por defecto incluye la región de 95% de confianza
# dos bandas de confianza
# Grafico dispersión y recta
plot(datos$numero_bateos, datos$corridas, col = "firebrick", pch
= 19,
     ylab = "corridas", xlab = "numero de bateos",
     main = "Bandas de confianza para valores medios y predicció
n")
abline(modelo_lineal, col = 1)

# Intervalos de confianza de la respuesta media:
# valores medios
nuevos_datos <- data.frame(numero_bateos= seq(min(numero_bateos)
,
                                              max(numero_bateos)
))
ic <- predict(modelo_lineal, nuevos_datos, interval = 'confidenc
e')
lines(nuevos_datos$numero_bateos, ic[, 2], lty = 2)
lines(nuevos_datos$numero_bateos, ic[, 3], lty = 2)

# Intervalos de predicción
# para cualquier valor
ic <- predict(modelo_lineal, nuevos_datos, interval = 'predictio
n')
```

```
lines(nuevos_datos$numero_bateos, ic[, 2], lty = 2, col = 'red')
lines(nuevos_datos$numero_bateos, ic[, 3], lty = 2, col = 'red')
```

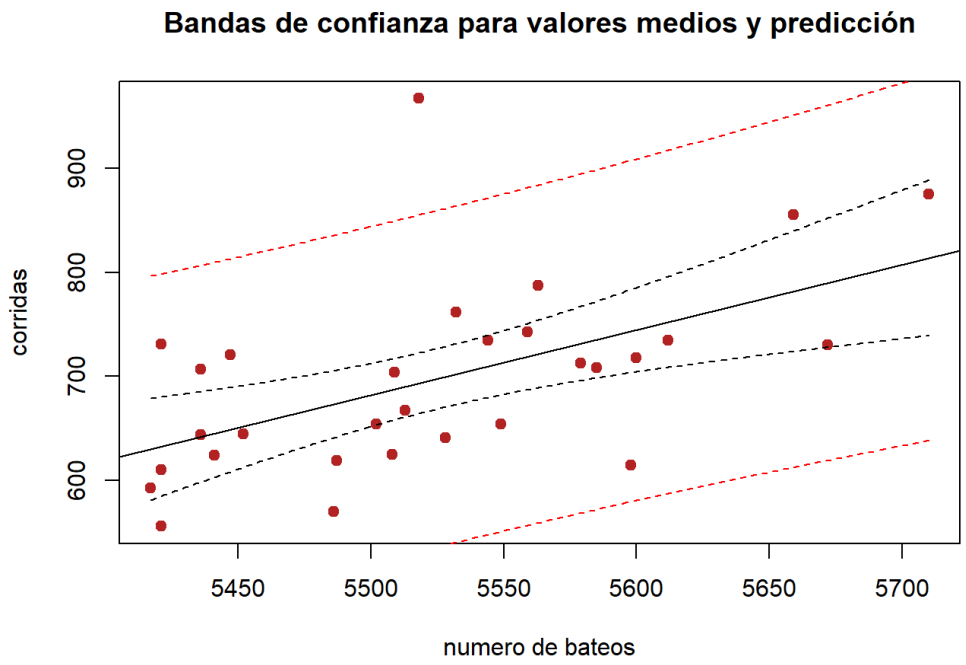


Figura 28. *El punto fuera de la banda de confianza de la predicción implica un valor atípico.*

Observando los valores reales, predichos y los residuales

```
# observando datos, valores predichos y residuales
datos$prediccion <- modelo_lineal$fitted.values
datos$residuos <- modelo_lineal$residuals
head(datos)
```

```
##   equipos numero_bateos corridas prediccion  residuos
## 1   Texas          5659      855   781.9700  73.02996
## 2  Boston          5710      875   813.9765  61.02352
## 3 Detroit          5563      787   721.7226  65.27737
## 4  Kansas          5672      730   790.1285 -60.12855
## 5     St.          5532      762   702.2677  59.73226
## 6  New_S.          5600      718   744.9430 -26.94299
```

g) Verificar condiciones para poder aceptar un modelo lineal

Relación lineal entre variable dependiente e independiente:

Se calculan los residuos para cada observación y se representan (scatterplot). Si las observaciones siguen la línea del modelo, los residuos se deben distribuir aleatoriamente entorno al valor 0.

```
# gráfico de residuales
ggplot(data = datos, aes(x = prediccion, y = residuos)) +
  geom_point(aes( color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_hline(yintercept = 0) +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  labs(title = "Distribucion de los residuos", x = "prediccion modelo"
,
       y = "residuo") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "non
e")
```

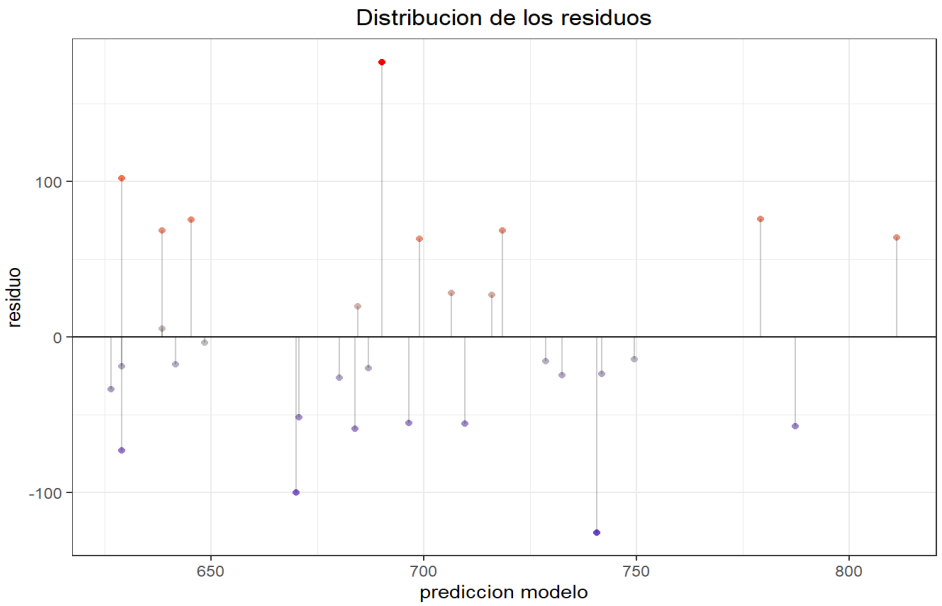


Figura 29. *Distribución de residuos.*

Los residuos se distribuyen de forma aleatoria entorno al 0 por lo que se acepta la linealidad.

Distribución normal de los residuos:

Los residuos se deben distribuir de forma normal con media 0. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a un test de contraste de normalidad.

Histograma de residuos.

```
par(mfrow = c(1, 1))
# Distribución normal de los residuos:
ggplot(data = datos, aes(x = residuos)) +
  geom_histogram(aes(y = ..density..)) +
  labs(title = "Histograma de los residuos") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

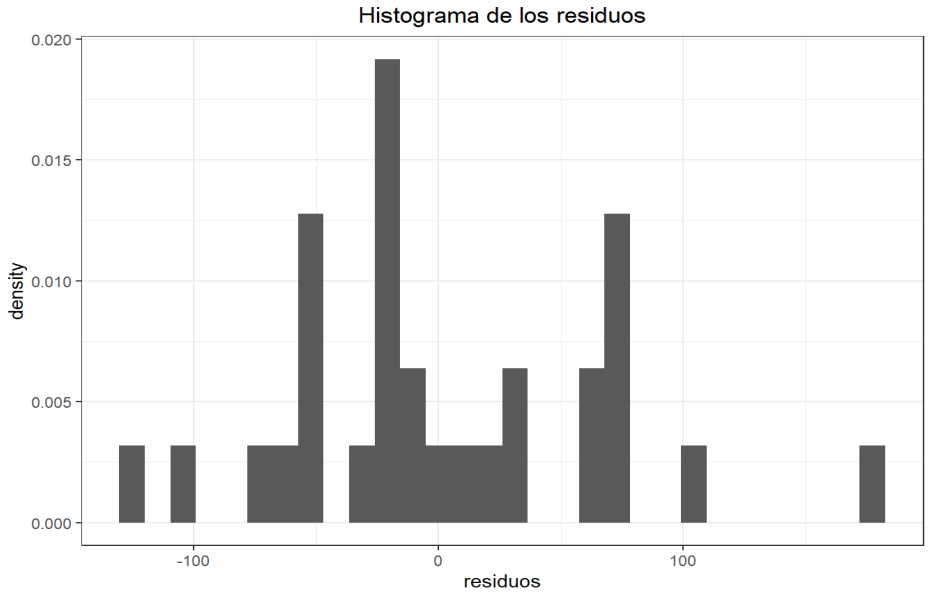


Figura 30. *Histograma de residuos del modelo.*

No se evidencia una clara distribución normal de los residuos.

Gráfico de cuantiles.

```
# grafico de cuantiles
```

```
qqnorm(modelo_lineal$residuals, main = "Gráfico de cuantiles QQ plot",
       col = "darkred")
qqline(modelo_lineal$residuals)
```

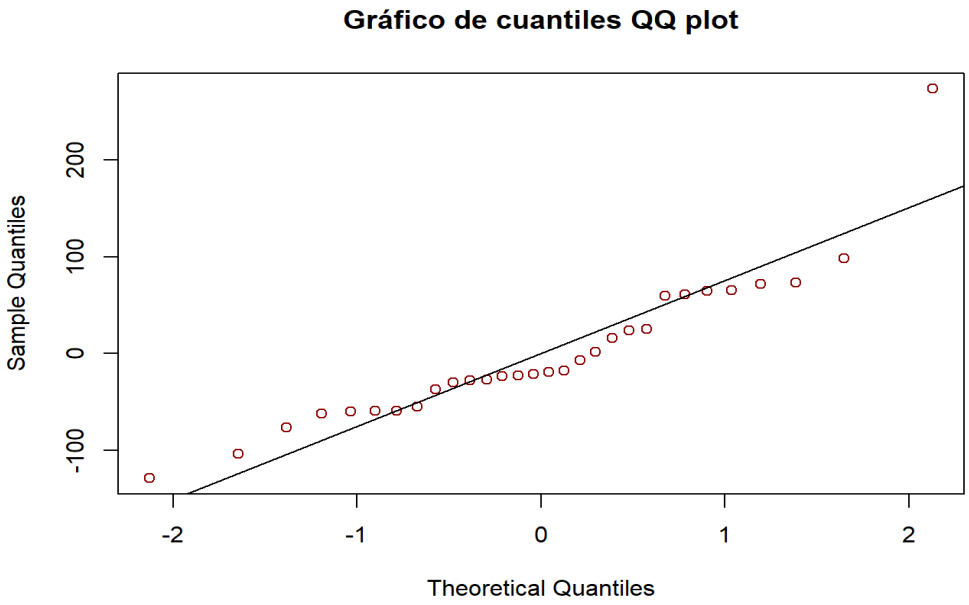


Figura 31. *Gráfico de cuantiles de los residuos.*

Un punto se aleja de la línea, posible no normalidad en los errores, finalmente realizamos una prueba de hipótesis de normalidad, por la cantidad de datos el estadístico más adecuado sería el test de Shapiro.

Prueba de hipótesis

```
# Test de normalidad
#shapiro.test
shapiro.test(modelo_lineal$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  modelo_lineal$residuals
## W = 0.88535, p-value = 0.003751
```

1) Prueba de hipótesis

Ho: existe normalidad en los residuos

Ha: no existe normalidad en los residuos

2) nivel de significancia 0.05

3) Prueba estadística. Shapiro y Wilks

W = 0.88535 con p=0.003751

4) decisión: $p(0.003751) < \alpha(0.05)$, se rechaza la Ho, los residuos no siguen una distribución normal, a diferencia de test de Kolmogorov.

Test de Kolmogorov- Smirnov

```
# Kolmogorov test
ks.test(modelo_lineal$residuals, "pnorm",
        mean = mean(modelo_lineal$residuals),
        sd = sd(modelo_lineal$residuals))
```

```
##  One-sample Kolmogorov-Smirnov test
##
## data:  modelo_lineal$residuals
## D = 0.15726, p-value = 0.4062
## alternative hypothesis: two-sided
```

$p(0.4062) > \alpha(0.05)$, No se rechaza la Ho. Los datos tienden a una distribución normal (esta distribución es para datos grandes)

Varianza constante de los residuos (Homocedasticidad):

La variabilidad de los residuos debe de ser constante a lo largo del eje X.

```
# Varianza constante de los residuos (Homocedasticidad):

ggplot(data = datos, aes(x = prediccion, y = residuos)) + geom_point(a
es( color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  geom_smooth(se = FALSE, color = "firebrick") +
  labs(title = "Distribucion de los residuos", x = "prediccion modelo"
, y = "residuo") +
```

```
  geom_hline(yintercept = 0) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "non
e")
```

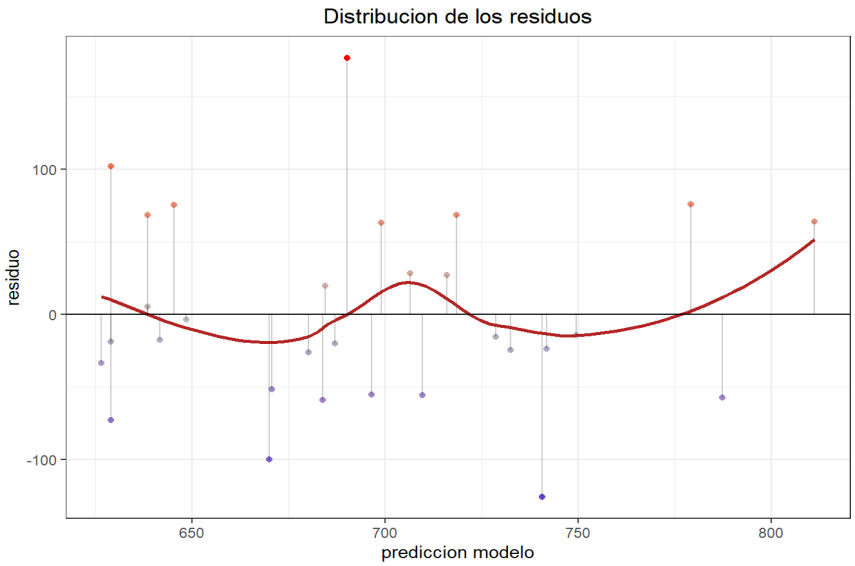


Figura 32. *Distribución de los residuos con valores predichos.*

Los residuos se comportan de manera aleatoria, no existe un patrón, existe homocedasticidad. Verifiquemos con un test de hipótesis.

Prueba de hipótesis de Breush - Pagan

```
# Test de Breush-Pagan
library(lmtest)
bptest(modelo_lineal)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo_lineal
## BP = 0.00011221, df = 1, p-value = 0.9915
```

Ni la representación gráfica ni el contraste de hipótesis muestran evidencias que haga sospechar falta de homocedasticidad.

1) Prueba de hipótesis

   Ho: existe homocedasticidad.

   Ha: falta de homocedasticidad

2) nivel de significancia 0.05
3) Prueba estadística. Pagan

   W = 0.0001122 con p=0. 0.9915

5) decisión: $p(0.9915) > \alpha(0.05)$, se acepta la Ho de existencia de homocedasticidad.

Autocorrelación de residuos:

Cuando se trabaja con intervalos de tiempo, es muy importante comprobar que no existe autocorrelación de los residuos, es decir que son independientes. Esto puede hacerse detectando visualmente patrones en la distribución de los residuos cuando se ordenan según han registrado o con el test de Durbin-Watson dwt() del paquete Car (Amat, 2016).

```
# Autocorrelación de residuos:
ggplot(data = datos, aes(x = seq_along(residuos), y = residuos)) +
 geom_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
geom_line(size = 0.3) +
  labs(title = "Distribucion de los residuos", x = "index", y =
"residuo")+
  geom_hline(yintercept = 0) +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```

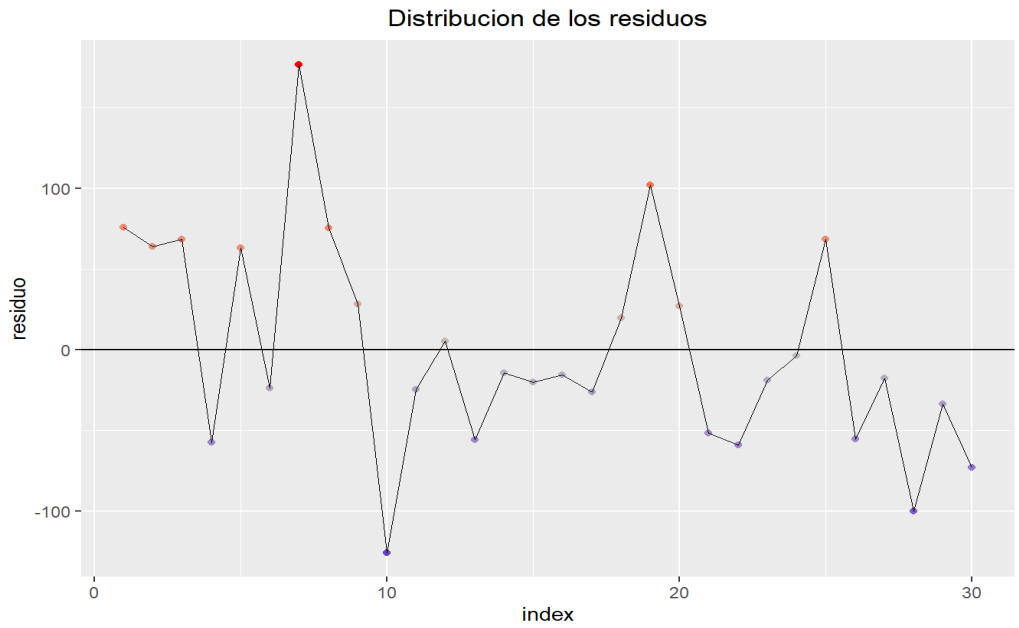


Figura 33. *Distribución de los residuos*

En este caso, la representación de los residuos no muestra ninguna tendencia.

Test de Durbin Watson

```
library(lmtest)
dwtest(modelo_lineal)
```

```
##  Durbin-Watson test
##
## data:  modelo_lineal
## DW = 1.59, p-value = 0.1045
## alternative hypothesis: true autocorrelation is greater than 0
```

1) Prueba de hipótesis

   Ho: no existe autocorrelación.

   Ha: existe autocorrelación

2) nivel de significancia 0.05
3) Prueba estadística. Durwin watson

DW = 1.59 con p=0.1045

4) decisión: $p(0.1045) > \alpha(0.05)$, se acepta la Ho de no existencia de autocorrelación.

h) Identificación de valores atípicos: *outliers*, *leverage* y observaciones influyentes (Amat, 2016).

Outlier u observación atípica: Observaciones que no se ajustan bien al modelo. El valor real se aleja mucho del valor predicho, por lo que su residuo es excesivamente grande. En una representación bidimensional se corresponde con desviaciones en el eje Y. es una observación que es numéricamente distante del resto de los datos. **Observación influyente:** Observación que influye sustancialmente en el modelo, su exclusión afecta al ajuste. No todos los *outliers* tienen por qué ser influyentes. Punto que tiene impacto en las estimativas del modelo.

- **Observación con alto leverage:** Observación con un valor extremo para alguno de los predictores. En una representación bidimensional se corresponde con desviaciones en el eje X. Son potencialmente puntos influyentes.

Independientemente de que el modelo se haya podido aceptar, siempre es conveniente identificar si hay algún posible *outlier*, *observación con alto leverage* u observación altamente influyente, puesto que podría estar condicionando en gran medida el modelo. La eliminación de este tipo de observaciones debe de analizarse con detalle y dependiendo de la finalidad del modelo. Si el fin es predictivo, un modelo sin estas observaciones puede lograr mayor precisión en la mayoría de casos. Sin embargo, es muy importante prestar atención a estos valores ya que, de no ser errores de medida, pueden ser los casos más interesantes. El modo adecuado a proceder cuando se sospecha de algún posible valor atípico o influyente es calcular el modelo de regresión incluyendo y excluyendo dicho valor.

Prueba de Bonferroni para detectar outliers

```r
# Prueba de Bonferroni para detectar outliers
library(car)
outlierTest(modelo_lineal, cutoff=Inf, n.max=4)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 7   4.782991         5.4575e-05    0.0016372
## 10 -1.777555         8.6744e-02           NA
## 28 -1.381908         1.7833e-01           NA
## 19  1.347912         1.8889e-01           NA
```

En la salida vemos cuatro observaciones (n.max=4) que tienen los mayores valores de residual estudentizado. La observación ubicada en la línea 7 es la única con un valor-p muy pequeño y por lo tanto hay evidencias para considerar esa observación como observación outlier (Hernandez, 2023). $p < \alpha$

Distancia de Cook

```r
# Distancia de Cook, detección de valores influyentes
cutoff <- 4 / (26-2-2)  # Cota
plot(modelo_lineal, which=4, cook.levels=cutoff, las=1)
abline(h=cutoff, lty="dashed", col="dodgerblue2")
```

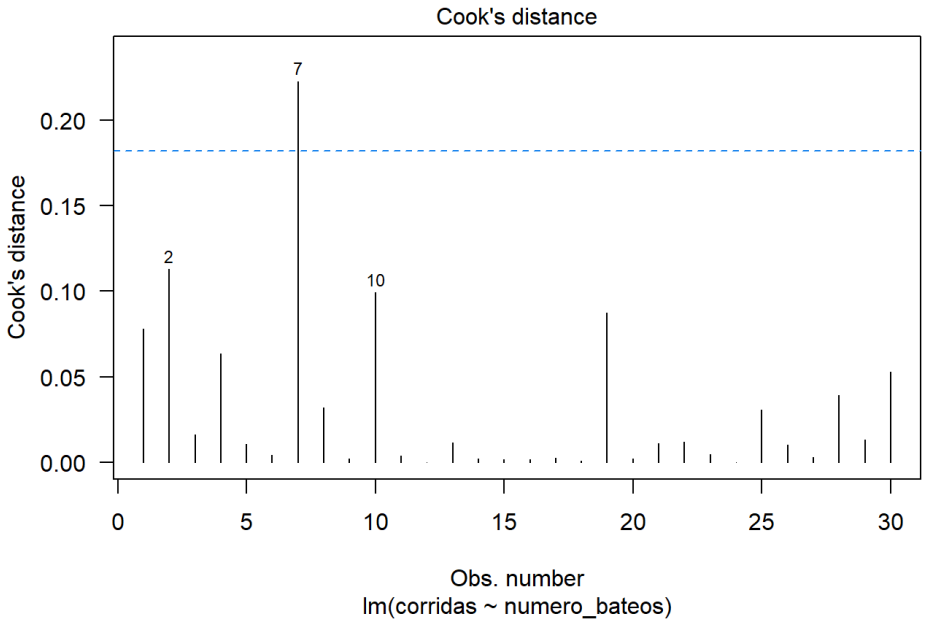


Figura 34. *Valores influyentes con la distancia de Cook*

En esta figura es claro que la observación 7 tiene Di por encima de la cota y se considera observación outlier. Podemos observar además valores influyentes como el 2 y 10, pero no significativas.

Utilizando otros indicadores.

```
# Identificación de valores atípicos: outliers, leverage y
observaciones influyentes
library(ggrepel)
library(dplyr)
datos$studentized_residual <- rstudent(modelo_lineal)
ggplot(data = datos,aes(x = prediccion, y =
abs(studentized_residual))) +
  geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +
# Se identifican en rojo residuos studentized absolutos > 3
geom_point(aes(color = ifelse(abs(studentized_residual) > 3, "red",
"black"))) +
scale_color_identity() +
#se muestra el equipo al que pertenece la observacion atipica,
geom_text_repel(data = filter(datos, abs(studentized_residual) > 3),
 aes(label = equipos)) +
labs(title = "Distribucion de los residuos studentized",x =
"prediccion modelo") +
theme_bw()+
```

```
theme(plot.title = element_text(hjust = 0.5), legend.position =
"none")
```

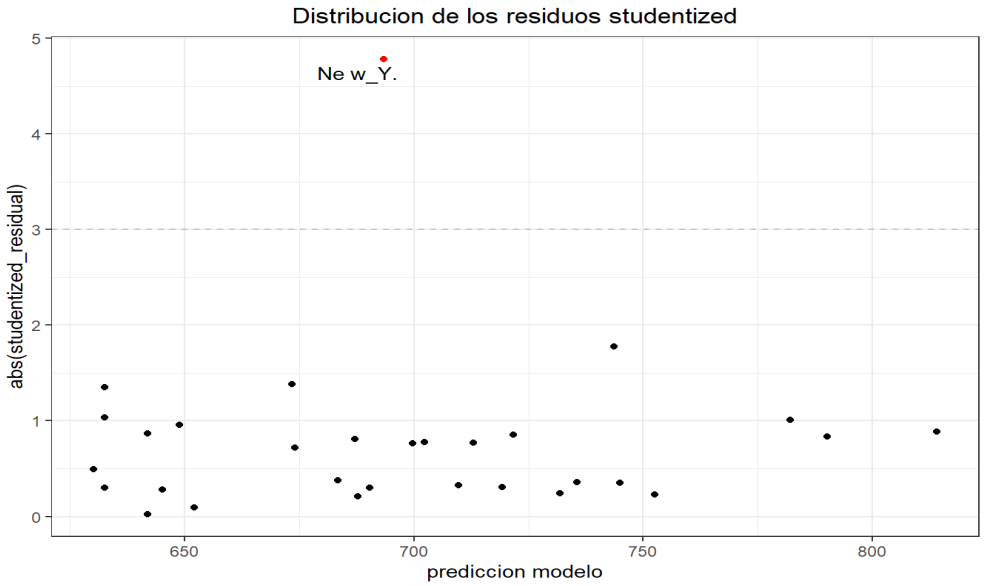


Figura 35. *Distribución de los residuos estudentizados para observar valores outliers.*

Cuales son o cual es el o los valores reconocidos como outliers

```
datos %>% filter(abs(studentized_residual) > 3)
```

```
##   equipos numero_bateos corridas prediccion residuos
studentized_residual
## 1 Ne w_Y.          5518      967   693.4817 273.5183
4.782991
```

```
which(abs(datos$studentized_residual) > 3)
```

```
## [1] 7
```

El estudio de los residuos *studentized* identifica al equipo de New_Y. como una posible observación atípica. Esta observación ocupa la posición 7 en la tabla de datos.

El hecho de que un valor sea atípico o con alto grado de *leverage* no implica que sea influyente en el conjunto del modelo. Sin embargo, si un valor es influyente, suele ser o atípico o de alto *leverage*. En R se dispone de la función outlierTest() del paquete car y de las funciones influence.measures(), influencePlot() y hatvalues() para identificar las observaciones más influyentes en el modelo (Amat, 2016).

```
library(car)
summary(influence.measures(model = modelo_lineal))
```

```
## Potentially influential observations of
##   lm(formula = corridas ~ numero_bateos, data = datos) :
##
##   dfb.1_ dfb.nmr_ dffit   cov.r   cook.d  hat
## 2 -0.43   0.44     0.47    1.30_*  0.11   0.22_*
## 7  0.07  -0.06     0.89_*  0.33_*  0.22   0.03
```

Se detectan como influyente la observación 2 y 7.

```
influencePlot(model = modelo_lineal)
```

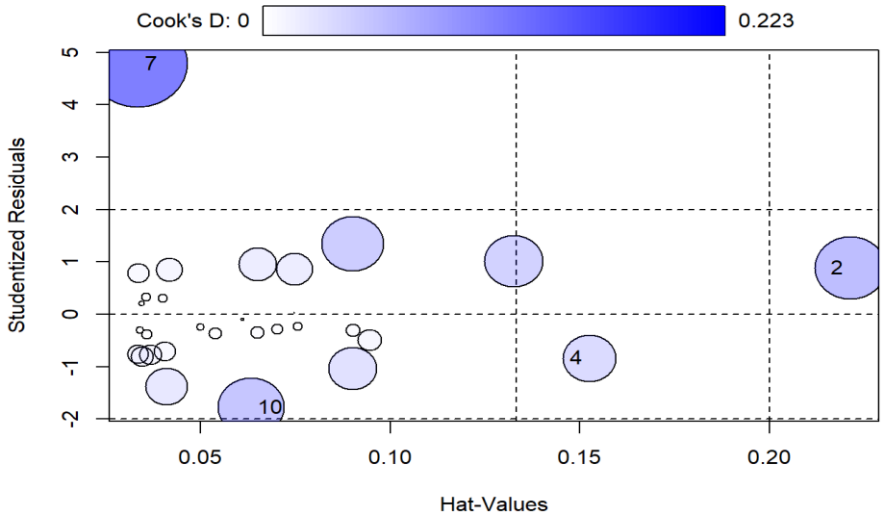


Figura 36. Observaciones *influyentes*

```
##        StudRes         Hat      CookD
## 2    0.8873817 0.22133381 0.11277084
## 4   -0.8368049 0.15252728 0.06369637
## 7    4.7829913 0.03349684 0.22254988
## 10 -1.7775546 0.06333282 0.09917231
```

Las funciones influence.measures() e influencePlot() detectan la observación 7 como atípica pero no significativamente influyente. Sí detectan como influyente la observación que ocupa la segunda posición. Para evaluar hasta qué punto condiciona el modelo, se recalcula la recta de mínimos cuadrados excluyendo esta observación.

Librería performance para evaluar supuestos y valores atípicos

```
# uso de libreria performance para evaluar
# supuestos y atipicos

library(tidyverse)          # data manipulation
model_performance(modelo_lineal)
```

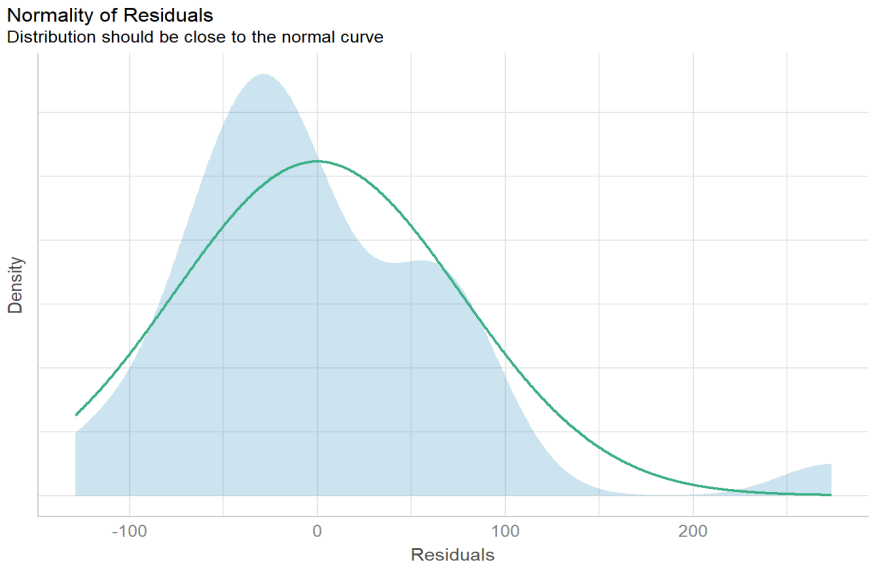
```
## # Indices of model performance
```

```
##
## AIC     |     AICc |      BIC |    R2 | R2 (adj.) |    RMSE |   Sigma
## ------------------------------------------------------------------------
## 350.187 | 351.111 | 354.391 | 0.302 |     0.277 | 75.002 | 77.635
```

```
check_normality(modelo_lineal)
```

```
## Warning: Non-normality of residuals detected (p = 0.005).
```

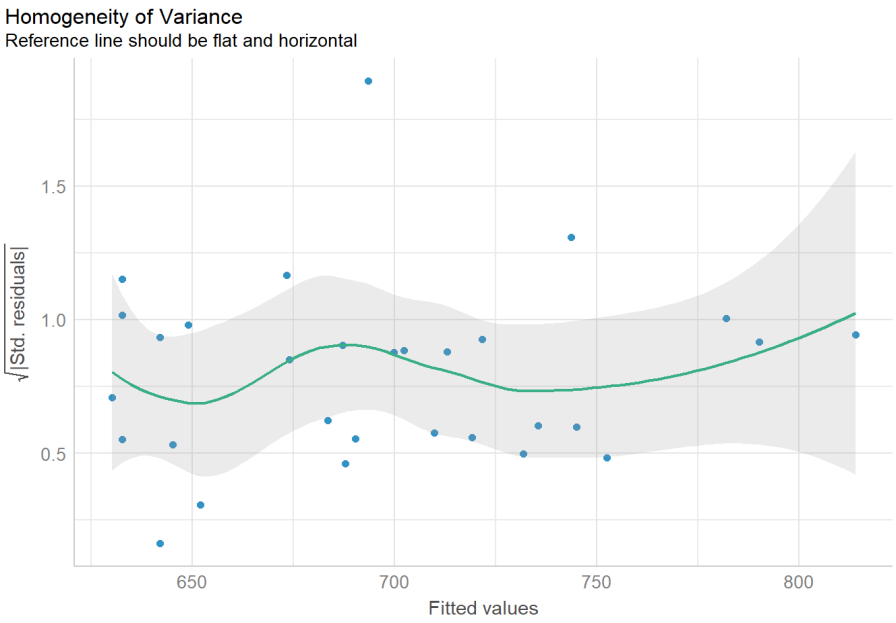
```
plot( check_normality(modelo_lineal) )
```




```
check_heteroscedasticity(modelo_lineal)
```

```
## OK: Error variance appears to be homoscedastic (p = 0.986).
```

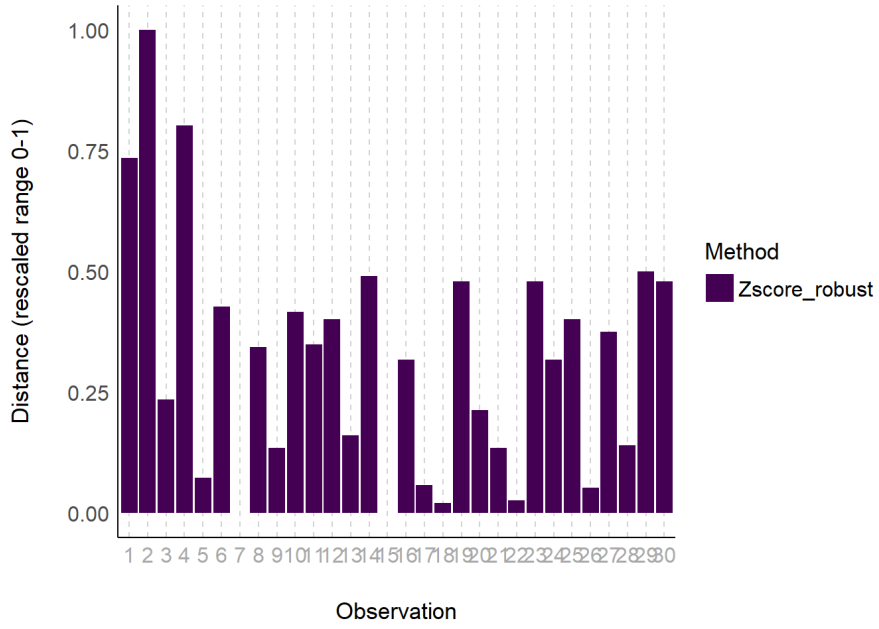
```
plot( check_heteroscedasticity(modelo_lineal) )
```




```
#outliers

check_outliers(numero_bateos)
```

```
## OK: No outliers detected.

## - Based on the following method and threshold: zscore_robust (3.291
).

## - For variable: numero_bateos
```

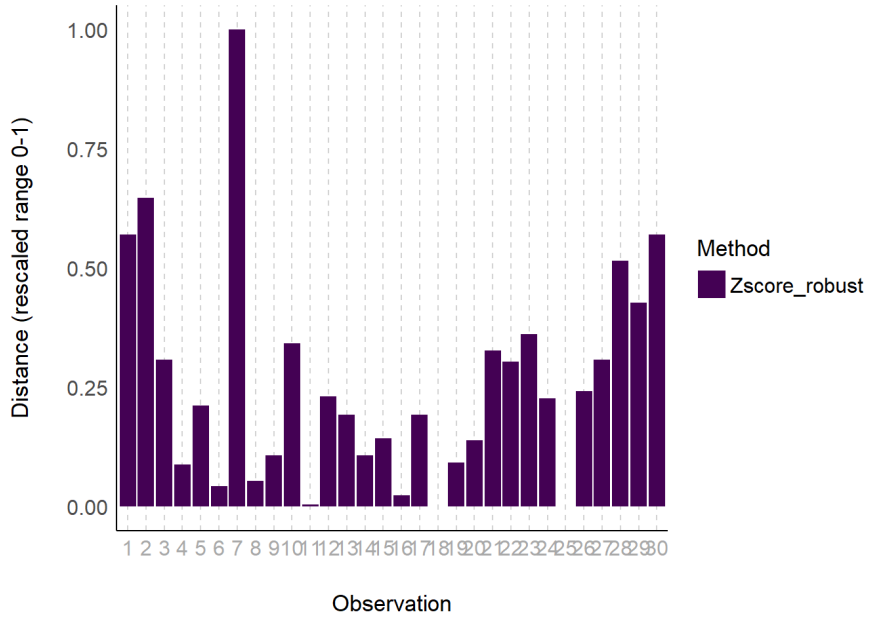
```
plot(check_outliers(numero_bateos))
```




```
check_outliers(corridas)
```

```
## OK: No outliers detected.

## - Based on the following method and threshold: zscore_robust (3.291
).

## - For variable: corridas
```

```
plot(check_outliers(corridas))
```



Vista del modelo excluyendo la observación atípica. (7)

```
# excluyendo el valor atípico
ggplot(data = datos, mapping = aes(x = numero_bateos, y = corridas)) +
  geom_point(color = "grey50", size = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  # se resalta el valor excluido
  geom_point(data = datos[7, ], color = "red", size = 2) +
  # se anade la nueva recta de minimos cuadrados
  geom_smooth(data = datos[-7, ], method="lm", se =FALSE,color =  "blue")
+
  labs(title = "Diagrama de dispersion", x = "numero  de bateos") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```

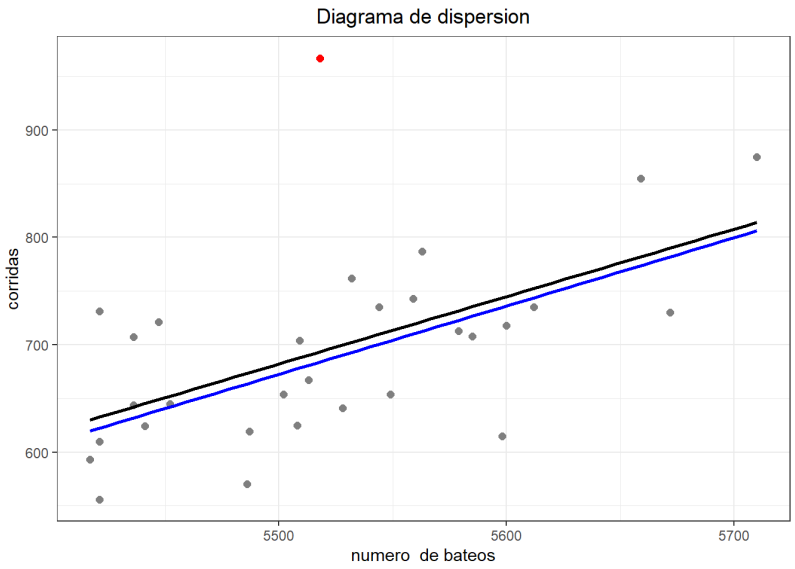


Figura 37. *Modelo inicial y modelo final*

Observamos cambio en la línea de regresión, veamos más detalladamente los parámetros. Obteniendo el modelo con todos los datos (modelo_inicial) y el modelo excluyendo el dato 7 (modelo_final).

```
modelo_inicial <- lm(formula = corridas ~ numero_bateos, data = datos)
summary(modelo_inicial)
```

```
## Call:
## lm(formula = corridas ~ numero_bateos, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -128.69  -50.54  -19.96   51.10  273.52
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2769.4894   997.0462  -2.778  0.00966 **
## numero_bateos     0.6276     0.1805   3.477  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.63 on 28 degrees of freedom
## Multiple R-squared:  0.3016, Adjusted R-squared:  0.2766
```

```
## F-statistic: 12.09 on 1 and 28 DF,  p-value: 0.001673
```

```
modelo_final <- lm(formula = corridas ~ numero_bateos, data = datos[-
7, ])
summary(modelo_final)
```

```
## Call:
## lm(formula = corridas ~ numero_bateos, data = datos[-7, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.88  -45.29  -11.03   34.46  108.69
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2825.3914   747.1319  -3.782 0.000786 ***
## numero_bateos     0.6360     0.1352   4.702 6.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.17 on 27 degrees of freedom
## Multiple R-squared:  0.4503, Adjusted R-squared:  0.4299
## F-statistic: 22.11 on 1 and 27 DF,  p-value: 6.776e-05
```

En ambos, los estimadores $\hat{\beta}_o$ y $\hat{\beta}_1$ son significativos, la diferencia esta en el coeficiente de determinación. En el modelo inicial se tiene un $R^2 =$ 30.16% y en el modelo final $R^2 = 45.03\%$, mejorando el ajuste de la varianza.

Hasta aquí el mejor modelo es el que excluye el valor atípico.

Vista del modelo excluyendo la observación influyente

En líneas arriba se conoce que el punto influyente significativo es el punto **2.**

Graficando el modelo inicial y el modelo excluyendo la observación influyente

```
ggplot(data = datos, mapping = aes(x = numero_bateos, y = corridas)) +
  geom_point(color = "grey50", size = 2) +
geom_smooth(method = "lm", se = FALSE, color = "black") +
  # se resalta el valor excluido
geom_point(data = datos[2, ], color = "red", size = 2) +
  # se anade la nueva recta de minimos cuadrados
geom_smooth(data = datos[-2, ], method="lm", se =FALSE,color =  "blue"
) +
labs(title = "Diagrama de dispersion", x = "numero  de bateos") +
theme_bw() +
theme(plot.title = element_text(hjust = 0.5), legend.position = "none"
)
```

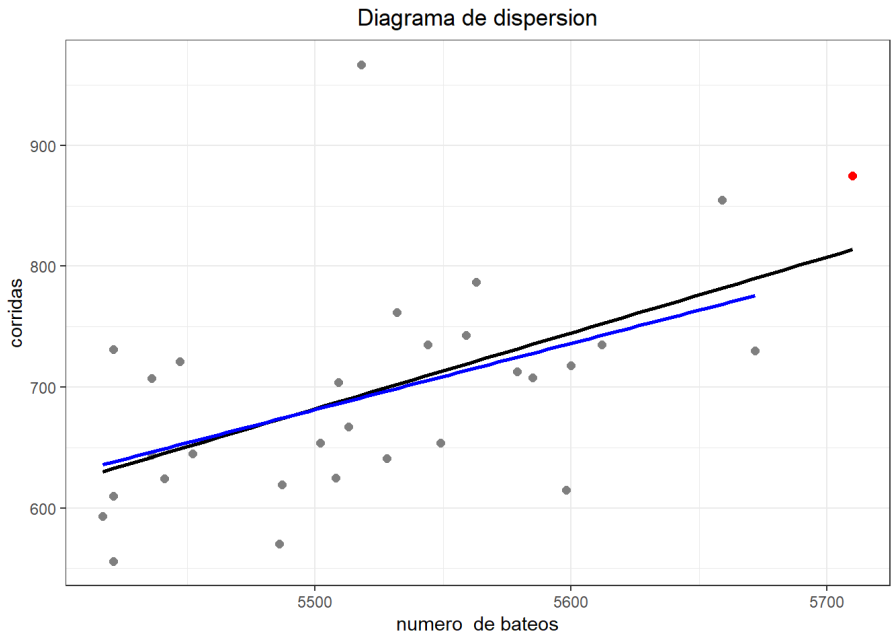


Figura 38. *Modelo sin observación influyente.*

La eliminación del valor identificado como influyente apenas cambia la recta de mínimos cuadrados. Para conocer con exactitud el resultado de excluir la observación se comparan los coeficientes del modelo inicial.

```
modelo_sin_influ <- lm(formula = corridas ~ numero_bateos, data =
datos[-2, ])
```

```
summary(modelo_sin_influ)
```

```
## Call:
## lm(formula = corridas ~ numero_bateos, data = datos[-2, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.19  -45.78  -18.29   29.43  275.70
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2335.7462  1113.8315  -2.097   0.0455 *
## numero_bateos     0.5486     0.2019   2.717   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.93 on 27 degrees of freedom
## Multiple R-squared:  0.2148, Adjusted R-squared:  0.1857
## F-statistic: 7.385 on 1 and 27 DF,  p-value: 0.01134
```

En ambos, los estimadores $\hat{\beta}_o$ *y* $\hat{\beta}_1$ son significativos, la diferencia está en el coeficiente de determinación. En el modelo inicial se tiene un $R^2 = 30.16\%$ y en el modelo sin observación influyente, $R^2 = 21.48\%$, bajando el ajuste de la varianza. Hasta aquí el mejor modelo es el que excluye el valor atípico.

i) Modelo final.

Análisis del modelo final excluyendo el dato atípico 7.

Datos completos.

```
#### MODELO FINAL ############
# Datos originales.
equipos <- c("Texas", "Boston", "Detroit", "Kansas", "St.", "New_S.",
             "Ne w_Y.", "Milwaukee", "Colorado", "Houston", "Baltimore",
             "Los_An.", "Chica go", "Cincinnati", "Los_P.", "Philadelphia",
             "Chicago", "Cleveland", "Ari zona", "Toronto", "Minnesota",
             "Florida", "Pittsburgh", "Oakland", "Tampa ","Atlanta","Washingto
n",
```

```
             "San.F", "San.I", "Seattle")
numero_bateos <- c(5659,5710,5563,5672,5532,5600,5518,5447,5544,5598,
                   5585,5436,5549,5612,5513,5579,5502,5509,5421,5559,
                   5487,5508,5421,5452,5436,5528,5441,5486,5417,5421)
corridas <- c(855,875,787,730,762,718,967,721,735,615,708,644,654,735,
              667,713,654,704,731,743,619,625,610,645,707,641,624,570,
              593,556)
datos <- data.frame(equipos, numero_bateos, corridas)
head(datos)
```

```
#### MODELO FINAL
# eliminando el dato 7 y guardando los datos en un nuevo archivo
datos1 <- datos[-c(7), ]
head(datos1)
```

```
##   equipos numero_bateos corridas
## 1   Texas          5659      855
## 2  Boston          5710      875
## 3 Detroit          5563      787
## 4  Kansas          5672      730
## 5     St.          5532      762
## 6  New_S.          5600      718
```

```
str(datos1)
```

```
## 'data.frame':    29 obs. of  3 variables:
##  $ equipos      : chr  "Texas" "Boston" "Detroit" "Kansas" ...
##  $ numero_bateos: num  5659 5710 5563 5672 5532 ...
##  $ corridas     : num  855 875 787 730 762 718 721 735 615 708 ...
```

Quedan 29 observaciones y 3 variables.

```
# histograma para las variables
library(ggplot2)
par(mfrow = c(1, 2))
hist(datos1$numero_bateos, breaks = 10, main = "", xlab = "número de
bateos",
```

```
     ylab="Frecuencia", border = "darkred")
hist(datos1$corridas, breaks = 10, main = "", xlab =
"corridas",ylab="Frecuencia",
     border = "blue")
```

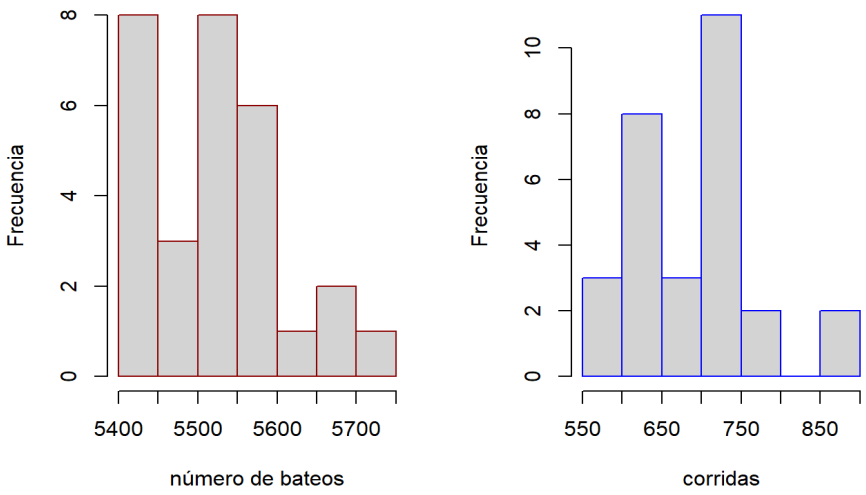


Figura 39. *Histogramas*

Resumen de estadísticos descriptivos.

```
summary(datos1)
```

```
##    equipos           numero_bateos     corridas
##  Length:29          Min.   :5417   Min.   :556.0
##  Class :character   1st Qu.:5447   1st Qu.:625.0
##  Mode  :character   Median :5513   Median :704.0
##                     Mean   :5524   Mean   :687.6
##                     3rd Qu.:5579   3rd Qu.:731.0
##                     Max.   :5710   Max.   :875.0
```

```
# representación gráfica
require(ggplot2)
ggplot(data = datos1, mapping = aes(x = numero_bateos, y = corridas))
+
  geom_point(color = "firebrick", size = 3) +
  (labs(title = "Diagrama de dispersión", x = "número  de bateos")) +
  theme_bw() +
```

```
  theme(plot.title = element_text(hjust =0.5))
```

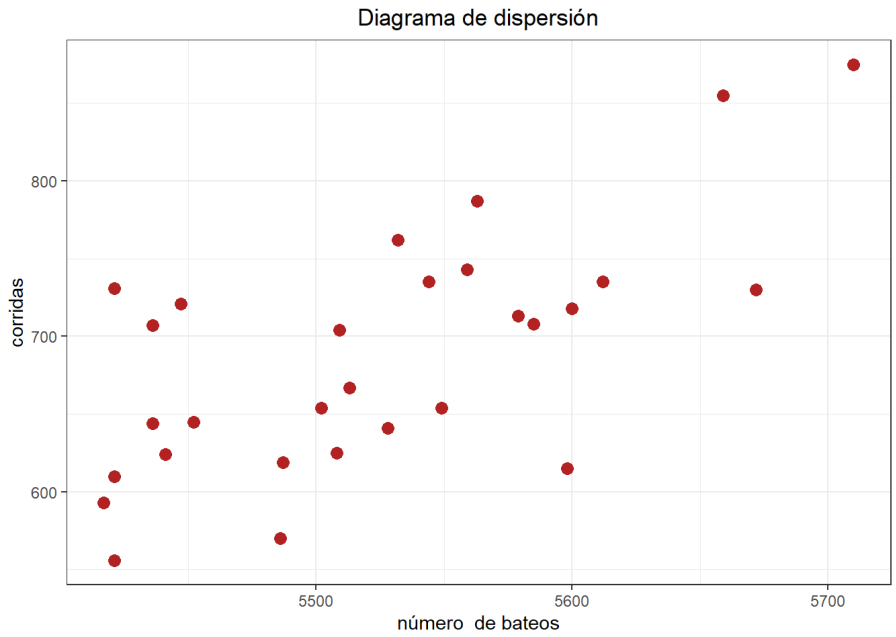


Figura 40. *Diagrama de dispersión*

```
cor.test(x = datos1$numero_bateos, y = datos1$corridas, method =
"pearson")
```

```
##  Pearson's product-moment correlation
##
## data:  datos1$numero_bateos and datos1$corridas
## t = 4.7025, df = 27, p-value = 6.776e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4038110 0.8327235
## sample estimates:
##       cor
## 0.6710081
```

El coeficiente de correlación se ha incrementado a 0.67 y es significativa con $p(0.0000677) < \alpha(0.05)$, los datos pueden ser ajustados a una regresión lineal.

```
# Cálculo del modelo de regresión lineal simple
```

```
modelo_final<- lm(corridas ~ numero_bateos, data=datos1)
summary(modelo_final)
```

```
## Call:
## lm(formula = corridas ~ numero_bateos, data = datos1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.88  -45.29  -11.03   34.46  108.69
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2825.3914   747.1319  -3.782 0.000786 ***
## numero_bateos     0.6360     0.1352   4.702 6.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.17 on 27 degrees of freedom
## Multiple R-squared:  0.4503, Adjusted R-squared:  0.4299
## F-statistic: 22.11 on 1 and 27 DF,  p-value: 6.776e-05
```

El intercepto y la variable X (número de bateos) resultan ser significativos al 5% con un coeficiente de determinación de 45.03% y un ANVA significativo $p(0.0000677) < \alpha(0.05)$.

```
# Intervalos de confianza para los parámetros del modelo
confint(modelo_final, level=0.95)
```

```
##                      2.5 %        97.5 %
## (Intercept)   -4358.3793538 -1292.4034930
## numero_bateos     0.3584894     0.9134908
```

```
# Representación gráfica del modelo
ggplot(data = datos1, mapping = aes(x = numero_bateos, y = corridas))
+
  geom_point(size=3) +
```

```
  labs(title = "corridas ~ numero de bateos", x = "numero  de bateos")
+
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.4))
```

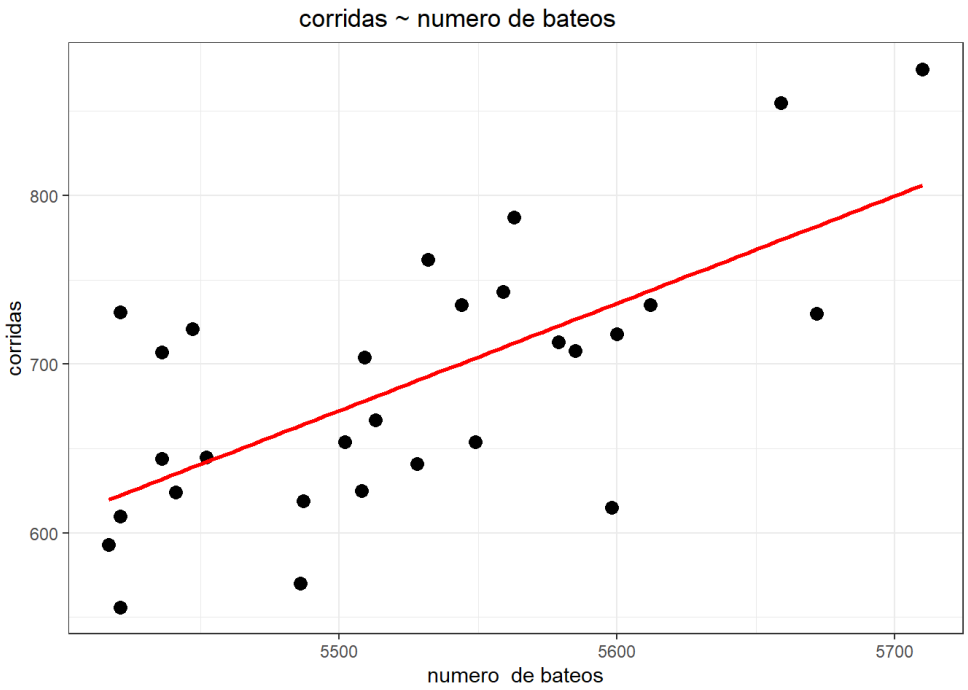


Figura 41. *Modelo de regresión*

```
# predecir

nuevos_datos <- data.frame(numero_bateos= seq(min(numero_bateos),max(n
umero_bateos)))

predict_value <- predict(modelo_final)

head(predict_value)
```

```
##        1        2        3        4        5        6
## 773.6767 806.1122 712.6217 781.9446 692.9060 736.1533
```

Cálculo del error medio cuadrático

```
y_predict <- predict(modelo_final)
```

```
str(corridas)
```

```
# Cálculo del error
error = y_predict - datos1$corridas
head(error)
```

```
#         1           2           3           4           5           6
## -81.323266  -68.887770  -74.378319   51.944605  -69.094013   18.153
316
```

```
# Cálculo del error cuadrático medio RMSE:
sqrt(mean(error^2))
```

```
## [1] 56.12652
```

O puede usar.

```
# Cálculo del error cuadrático medio RMSE:
sqrt(mean(modelo_final$residuals^2))
```

```
## [1] 56.12652
```

Bandas de confianza

```
# una banda
par(mfrow = c(1, 1))
puntos <- seq(from = min(datos1$numero_bateos),
              to = max(datos1$numero_bateos), length.out = 100)
limites_intervalo <- predict(object = modelo_final,
                             newdata = data.frame( numero_bateos = puntos)
,
                             interval = "confidence", level = 0.95)
head(limites_intervalo, 3)
```

```
##        fit      lwr      upr
## 1 619.7671 582.7842 656.7501
## 2 621.6494 585.3206 657.9782
## 3 623.5317 587.8501 659.2132
```

```
plot(datos1$numero_bateos, datos1$corridas, col = "firebrick", pch = 19,
     ylab = "corridas", xlab = "numero de bateos",
     main = "corridas ~ numero de bateos")
abline(modelo_final, col = 1)
lines(x = puntos, y = limites_intervalo[, 2], type = "l", col = 2, lty = 3
)
lines(x = puntos, y = limites_intervalo[, 3], type = "l", col = 3, lty = 3
)
```

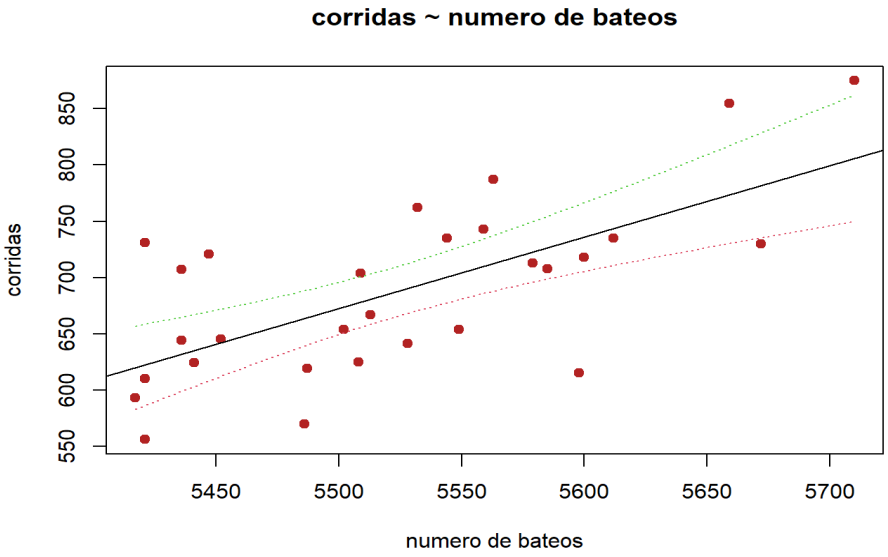


Figura 42. Bandas de confianza *del modelo de regresión*

```
# Por defecto incluye la región de 95% de confianza
# dos bandas
# Grafico dispersión y recta
plot(datos1$numero_bateos, datos1$corridas, col = "firebrick", pch = 1
9,
     ylab = "corridas", xlab = "numero de bateos",
     main = "corridas ~ numero de bateos")
abline(modelo_final, col = 1)
```

```
# Intervalos de confianza de la respuesta media:
# valores medios

ic <- predict(modelo_final, nuevos_datos, interval = 'confidence')
lines(nuevos_datos$numero_bateos, ic[, 2], lty = 2)
lines(nuevos_datos$numero_bateos, ic[, 3], lty = 2)

# Intervalos de prediccion
# para cualquier valor
ic <- predict(modelo_final, nuevos_datos, interval = 'prediction')
lines(nuevos_datos$numero_bateos, ic[, 2], lty = 2, col = 'red')
lines(nuevos_datos$numero_bateos, ic[, 3], lty = 2, col = 'red')
```

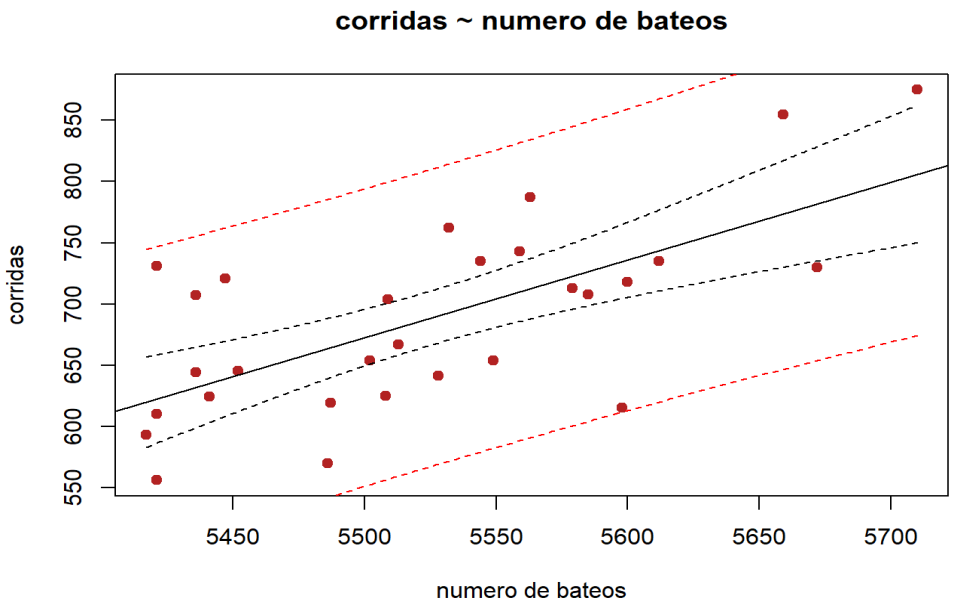


Figura 43. *Modelo con intervalos de confianza*

Usando el modelo para predecir el valor de Y con datos nuevos.

```
# prediciendo nuevos valores, para X = 5459 y 5455

prediciendo<- predict(modelo_final, data.frame(numero_bateos= c(5459,5
455)))

prediciendo
```

```
##        1        2
## 646.4787 643.9347
```

```
# observando datos, valores predichos y residuales
datos1$prediccion <- modelo_final$fitted.values
datos1$residuos <- modelo_final$residuals
head(datos1)
```

```
##   equipos numero_bateos corridas prediccion  residuos
## 1   Texas          5659      855   773.6767  81.32327
## 2  Boston          5710      875   806.1122  68.88777
## 3 Detroit          5563      787   712.6217  74.37832
## 4  Kansas          5672      730   781.9446 -51.94461
## 5     St.          5532      762   692.9060  69.09401
## 6  New_S.          5600      718   736.1533 -18.15332
```

```
# gráfico de residuales
ggplot(data = datos1, aes(x = prediccion, y = residuos)) +
  geom_point(aes( color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_hline(yintercept = 0) +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  labs(title = "Distribucion de los residuos", x = "prediccion modelo"
,
       y = "residuo") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "non
e")
```

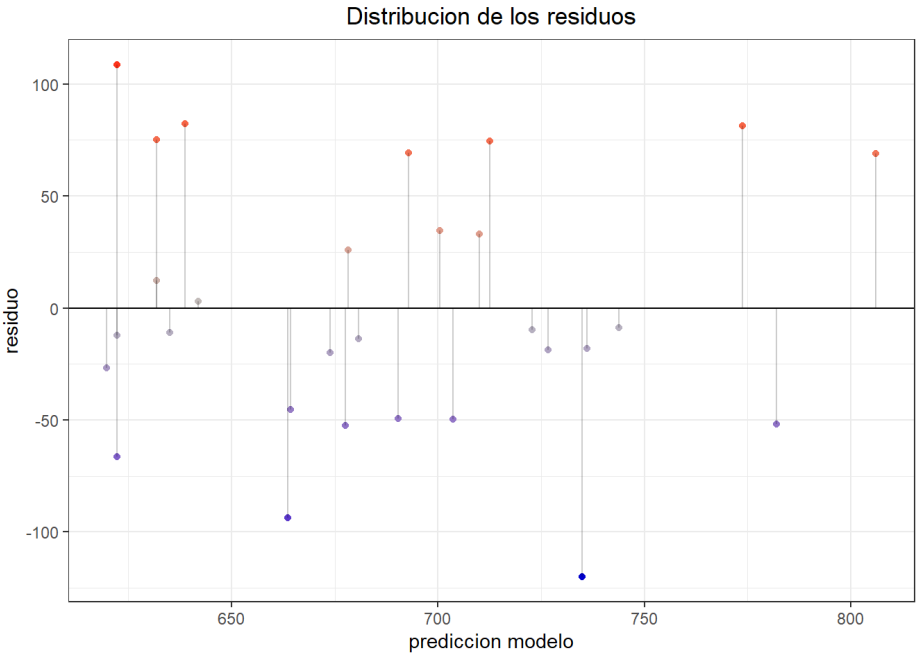


Figura 44. *Gráfico de residuales*

```
par(mfrow = c(1, 1))
# Distribución normal de los residuos:
ggplot(data = datos1, aes(x = residuos)) +
  geom_histogram(aes(y = ..density..)) +
  labs(title = "Histograma de los residuos") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

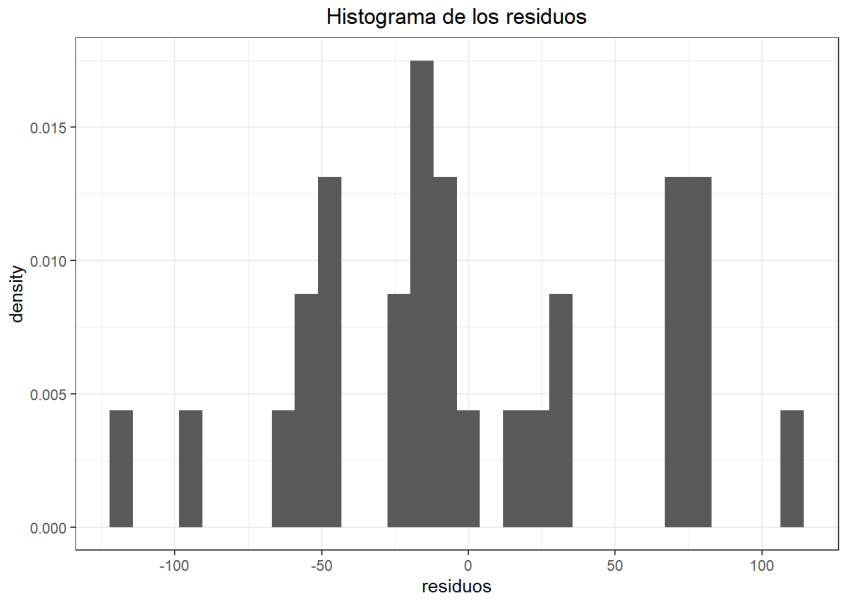


Figura 45. *Histograma de los residuales*

```
# gráfico de cuantiles
qqnorm(modelo_final$residuals, main = "Gráfico de cuantiles QQ plot",
       col = "darkred")
qqline(modelo_final$residuals)
```

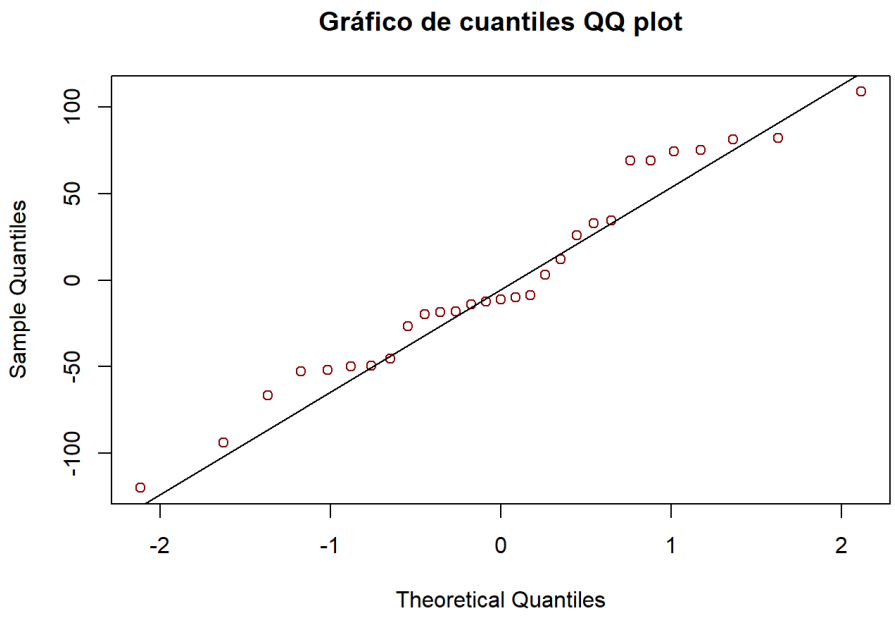


Gráfico 46. *Gráfico de cuantiles QQ.*

```
# Test de normalidad
#shapiro.test
shapiro.test(modelo_final$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  modelo_final$residuals
## W = 0.96269, p-value = 0.3822
```

Al ser $p(0.3822) > \alpha(0.05)$, no se rechaza la Ho, los datos siguen una distribución normal.

```r
# Kolmogorov test
ks.test(modelo_final$residuals, "pnorm",
        mean = mean(modelo_final$residuals),
        sd = sd(modelo_final$residuals))
```

```
##  One-sample Kolmogorov-Smirnov test
##
## data:  modelo_final$residuals
## D = 0.14732, p-value = 0.5083
## alternative hypothesis: two-sided
```

```r
# Varianza constante de los residuos (Homocedasticidad):
ggplot(data = datos1, aes(x = prediccion, y = residuos)) + geom_point(
aes( color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +
  geom_smooth(se = FALSE, color = "firebrick") +
  labs(title = "Distribucion de los residuos", x = "prediccion modelo"
, y = "residuo") +
  geom_hline(yintercept = 0) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "non
e")
```

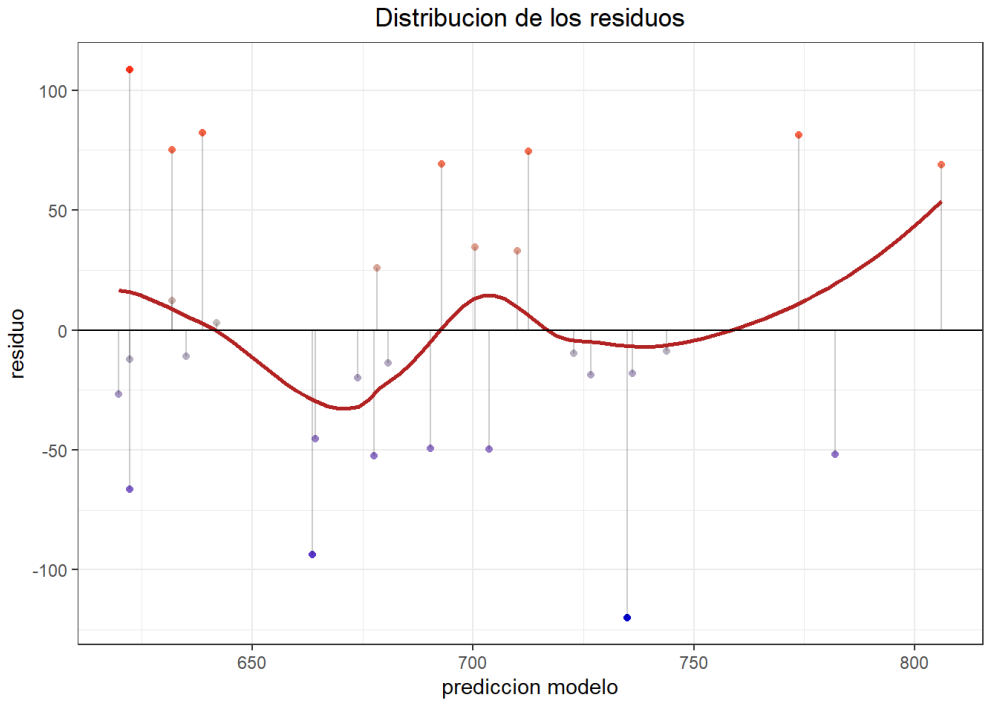


Figura 47. *Gráfico de residuos*

```
# Test de Breush-Pagan
library(lmtest)
bptest(modelo_final)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo_final
## BP = 0.076886, df = 1, p-value = 0.7816
```

Al ser $p(0.7816) > \alpha(0.05)$, no se rechaza Ho, los datos tienen varianza constante, es decir son homocedasticos.

```
# Autocorrelación de residuos:
ggplot(data = datos1, aes(x = seq_along(residuos), y = residuos)) + ge
om_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") + g
eom_line(size = 0.3) +
  labs(title = "Distribucion de los residuos", x = "index", y = "resid
uo")+
```

```
  geom_hline(yintercept = 0) +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "non
e")
```

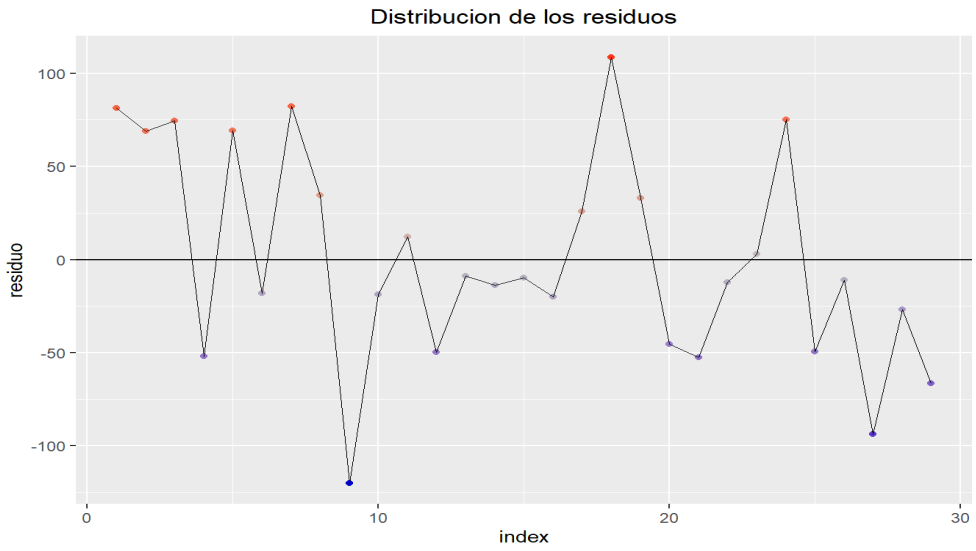


Figura 48. *Gráfico de distribución de residuos.*

```
#test de Durwin Watson
library(lmtest)
dwtest(modelo_final)
```

```
##  Durbin-Watson test
##
## data:  modelo_final
## DW = 1.633, p-value = 0.1327
## alternative hypothesis: true autocorrelation is greater than 0
```

Al ser $p(0.1327) > \alpha(0.05)$, no se rechaza la Ho, no existe autocorrelacion

```
# Prueba de Bonferroni para detectar outliers
library(car)
outlierTest(modelo_final, cutoff=Inf, n.max=4)
```

```
##     rstudent unadjusted p-value Bonferroni p
```

```
## 10 -2.292364           0.030226      0.87657
## 19  2.077211           0.047800           NA
## 28 -1.701835           0.100720           NA
## 1   1.539515           0.135760           NA
```

El punto 10 que es el nuevo outlier no es significativo.

```
# Distancia de Cook, detección de valores influyentes
cutoff <- 4 / (26-2-2)  # Cota
plot(modelo_final, which=4, cook.levels=cutoff, las=1)
abline(h=cutoff, lty="dashed", col="dodgerblue2")
```

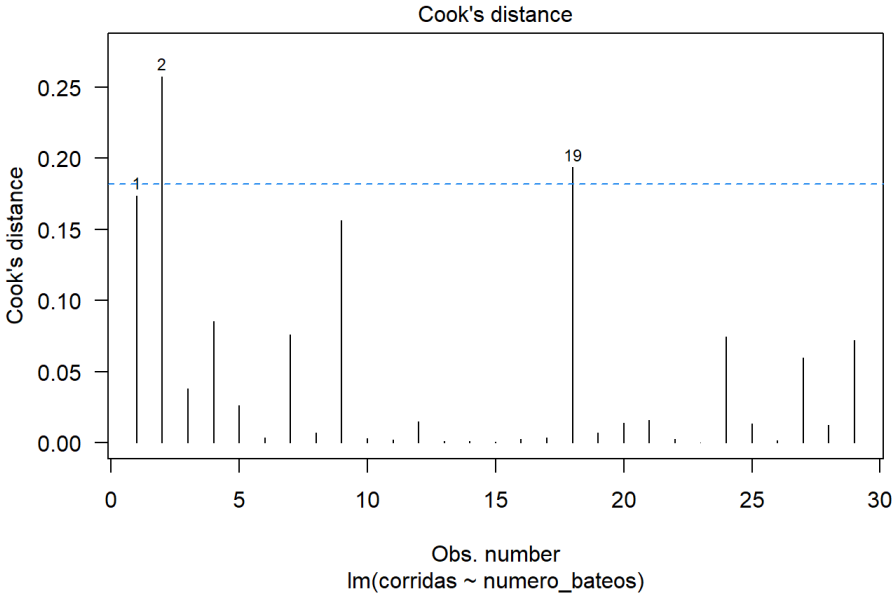


Figura 49. *Gráfico de distancia de Cook.*

La figura, muestra valores por encima de la línea como valores influyentes, veamos su significancia:

```
# Identificacion de valores atipicos: outliers, leverage y observacion
es influyentes

library(ggrepel)
library(dplyr)
datos1$studentized_residual <- rstudent(modelo_final)
```

```
ggplot(data = datos1,aes(x = prediccion, y = abs(studentized_residual)
)) +
  geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +
  # Se identifican en rojo residuos studentized absolutos > 3
  geom_point(aes(color = ifelse(abs(studentized_residual) > 3, "red",
"black"))) +
  scale_color_identity() +
  #se muestra el equipo al que pertenece la observacion atipica,
  geom_text_repel(data = filter(datos1, abs(studentized_residual) > 3)
,
                  aes(label = equipos)) +
  labs(title = "Distribucion de los residuos studentized",x = "predicc
ion modelo") +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5), legend.position = "non
e")
```

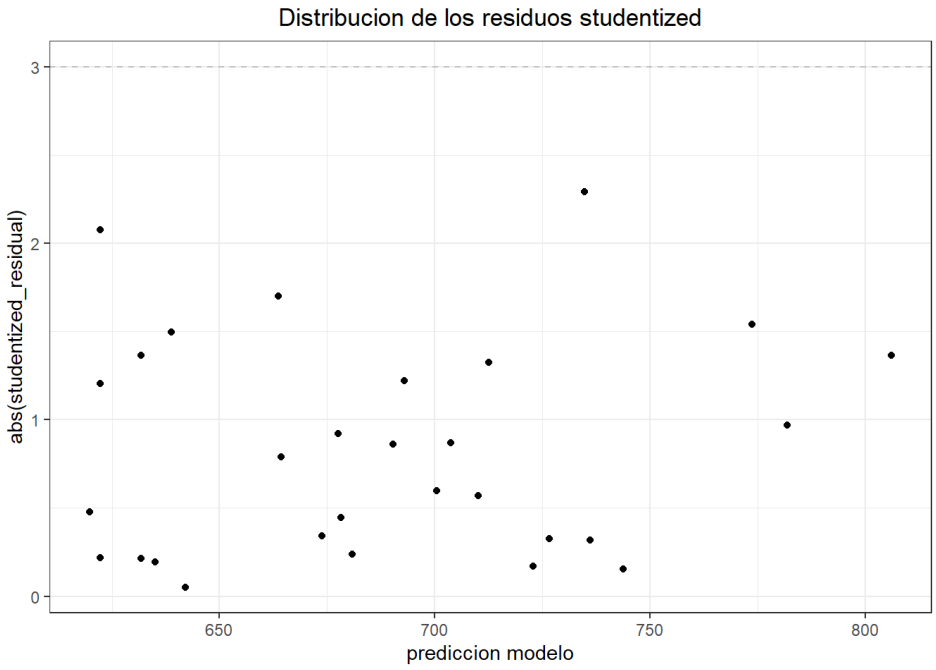


Figura 50. *Gráfico de residuos estudizado.*

La figura, no muestra valores mayores a 3, no hay valores atípicos significativos

```
datos1 %>% filter(abs(studentized_residual) > 3)
```

```
## [1] equipos              numero_bateos        corridas
## [4] prediccion           residuos             studentized_residual
## <0 rows> (or 0-length row.names)
```

```
which(abs(datos1$studentized_residual) > 3)
```

```
## integer(0)
```

No se observa valores atípicos no influyentes

```
# prediciendo nuevos valores, cuando X = 5459 y 5455
prediciendo<- predict(modelo_final, data.frame(numero_bateos= c(5459,5455)))
prediciendo
```

```
##        1      2
## 646.4787   643.9347
```

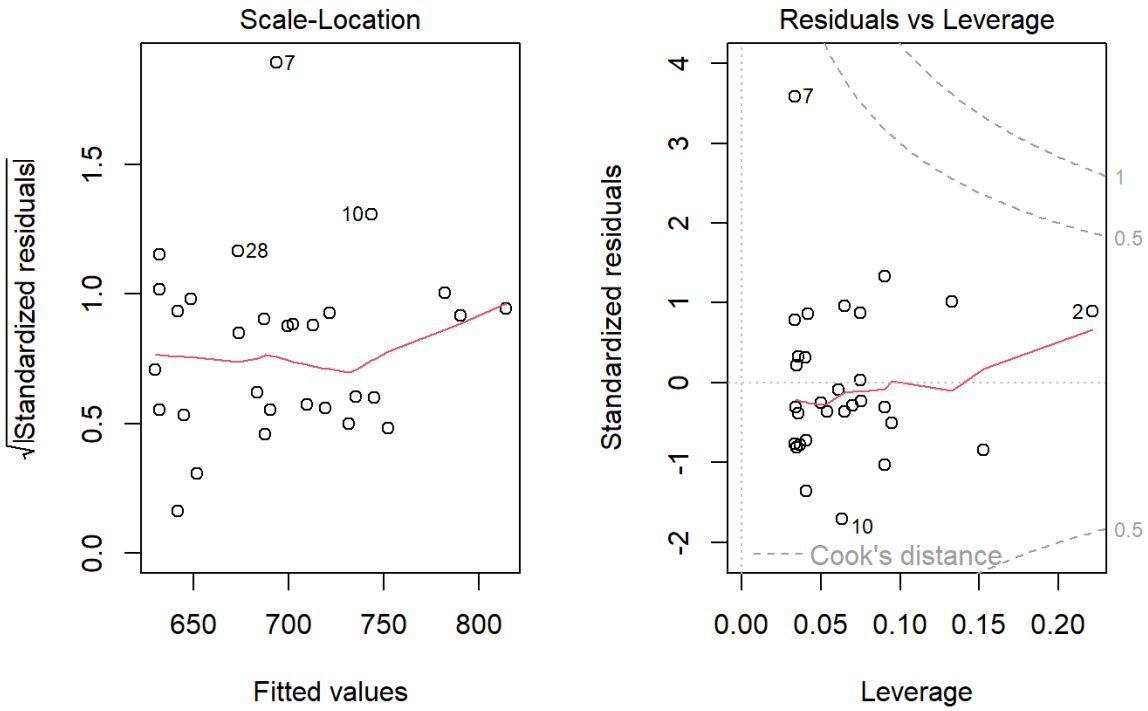
Conclusión

Dado que se satisfacen todas las condiciones para considerar válido un modelo de regresión lineal por mínimos cuadrados y que el *p-value* indica que el ajuste es significativo, se puede aceptar el modelo lineal.

```
# COMPARACION DE MODELOS
library(performance)
compare_performance(modelo_lineal, modelo_final, rank = TRUE)
```

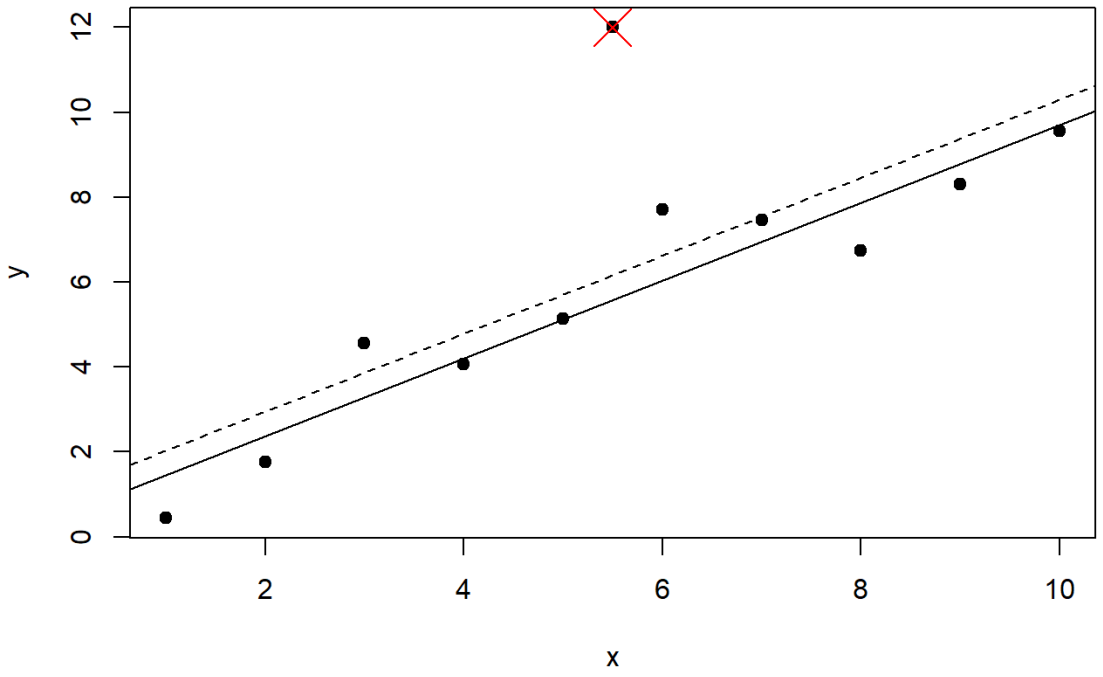
```
## # Comparison of Model Performance Indices
##
# Name          | Model |  R2  | R2 (adj.) |  RMSE |  Sigma | AIC weights | AICc weights | BIC weights | Performance-Score
## ---------------------------------------------------------------------------------------------------------------
## modelo_final |    lm | 0.450 |    0.430 | 56.127 | 58.168 |       1.000 |        1.000 |        1.000 |          100.00%
## modelo_lineal |    lm | 0.302 |    0.277 | 75.002 | 77.635 |    7.20e-07 |     7.34e-07 |     6.84e-07 |            0.00%
```

```r
# Evaluacion de los residuos de un modelo lineal simple mediante grafi
cos R
par(mfrow = c(1, 2))
plot(modelo_lineal)
```

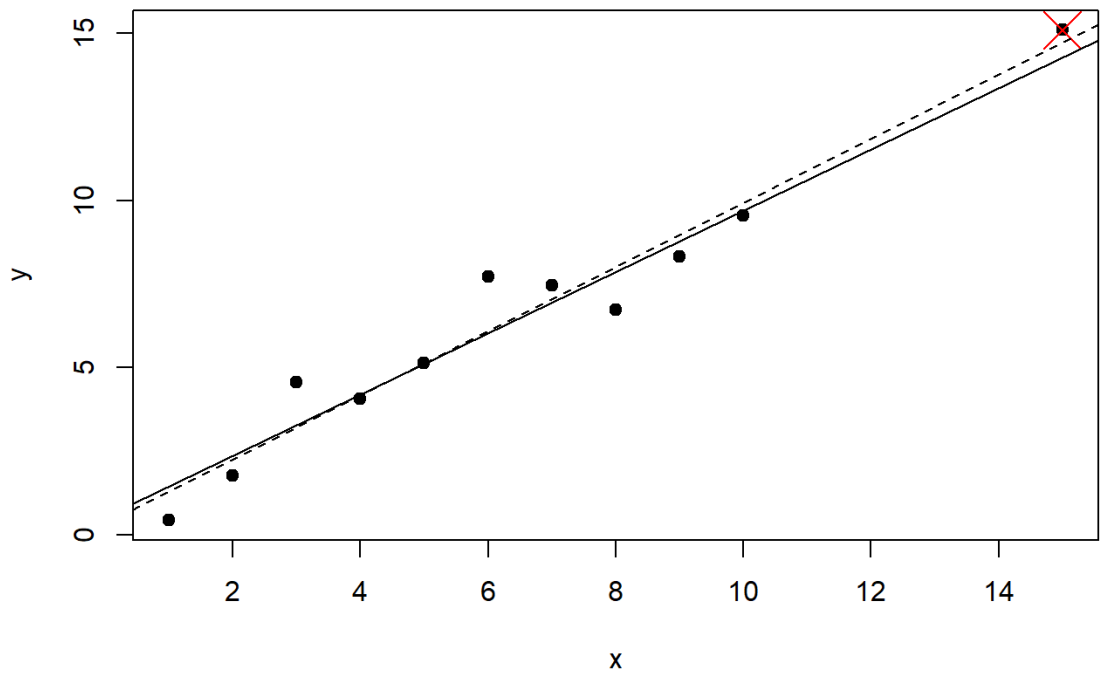



```r
# Identificacion de outliers, observaciones con alto leverage
# y observaciones influyentes
```

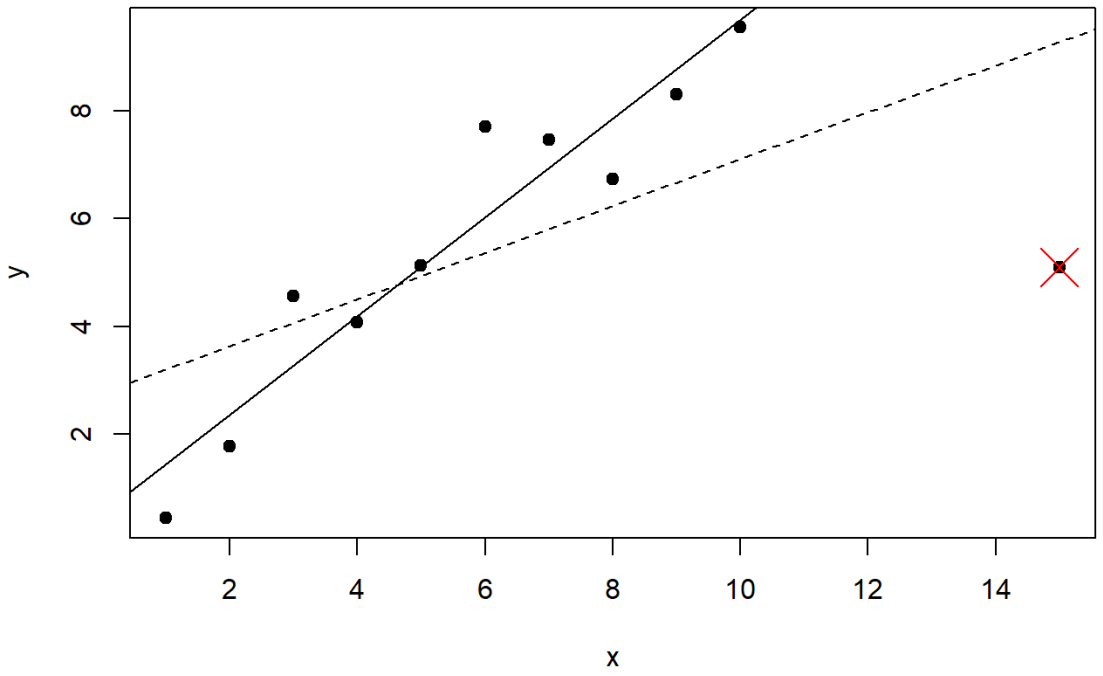
```r
par(mfrow = c(1, 1))
set.seed(123)
datos <- data.frame(x = 1:10, y = 1:10 + rnorm(10))
modelo_1 <- lm(y ~ x, data = datos)
observacion <- c(5.5, 12)
modelo_2 <- lm(y ~ x, data = rbind(datos, observacion))
plot(y ~ x, data = rbind(datos, observacion), pch = 19)
points(5.5, 12, pch = 4, cex = 3, col = "red")
abline(modelo_1)
abline(modelo_2, lty = 2)
```




```r
# Observacion con alto leverage
observacion <- c(15, 15.1)
modelo_2 <- lm(y ~ x, data = rbind(datos, observacion))
plot(y ~ x, data = rbind(datos, observacion), pch = 19)
points(15, 15.1, pch = 4, cex = 3, col = "red")
abline(modelo_1)
abline(modelo_2, lty = 2)
```




```
# Observacion influyente
observacion <- c(15, 5.1)
modelo_2 <- lm(y ~ x, data = rbind(datos, observacion))
plot(y ~ x, data = rbind(datos, observacion), pch = 19)
points(15, 5.1, pch = 4, cex = 3, col = "red")
abline(modelo_1)
abline(modelo_2, lty = 2)
```



13. Regresión lineal con un predictor categórico de dos niveles.

El set de datos sexab del paquete faraway contiene los resultados de un estudio en el que se investigó las secuelas que padecen mujeres adultas debido a abusos sufridos durante la infancia. En una clínica médica se midió el nivel de estrés post-traumático (*ptsd*) y nivel de abuso físico sufrido (*cpa*), ambos en escalas estandarizadas, en 45 mujeres que fueron víctimas en su infancia (*csa*). Las mismas mediciones se registraron para 31 mujeres que no sufrieron ningún tipo de abuso (Amat, 2016).

cpa: Abuso físico infantil en escala estándar

ptsd: Trastorno de estrés postraumático en escala estándar

csa: Abuso sexual infantil: abusado o no abusado

```
library(faraway)
library(ggplot2)
data(sexab)
head(sexab)
```

```
##        cpa          ptsd        csa
## 1  2.04786    9.71365 Abused
## 2  0.83895    6.16933 Abused
## 3 -0.24139  15.15926 Abused
## 4 -1.11461  11.31277 Abused
## 5  2.01468    9.95384 Abused
## 6  6.71131    9.83884 Abused
```

```
str(sexab)
```

```
## 'data.frame':    76 obs. of  3 variables:
##  $ cpa : num  2.048 0.839 -0.241 -1.115 2.015 ...
##  $ ptsd: num  9.71 6.17 15.16 11.31 9.95 ...
##  $ csa : Factor w/ 2 levels "Abused","NotAbused": 1 1 1 1 1 1 1 1 1 1 ...
```

La data contiene 76 observaciones y 3 variables.

Resumen de estadísticos descriptivos

```
summary(sexab)
```

```
##                  cpa              ptsd                  csa
##  Min.    : -3.1204   Min.    : -3.349      Abused   :45
##  1st Qu. :  0.8321   1st Qu. : 6.173      NotAbused:31
##  Median : 2.0707   Median : 8.909
##  Mean    : 2.3547   Mean    : 8.986
##  3rd Qu. : 3.7387   3rd Qu. :12.240
##  Max.    : 8.6469   Max.    :18.993
```

Para el análisis, tomaremos las variables ptsd y csa

a) Diagrama de dispersión

```
# representación grafica
require(ggplot2)
ggplot(data = sexab, mapping = aes(x = csa, y =ptsd )) +
  geom_point(color = "firebrick", size = 2) +
  (labs(title = "Diagrama de dispersion", x = "Deformación del acero
(X)")) +
  theme_bw() +
  theme(plot.title = element_text(hjust =0.5))
```

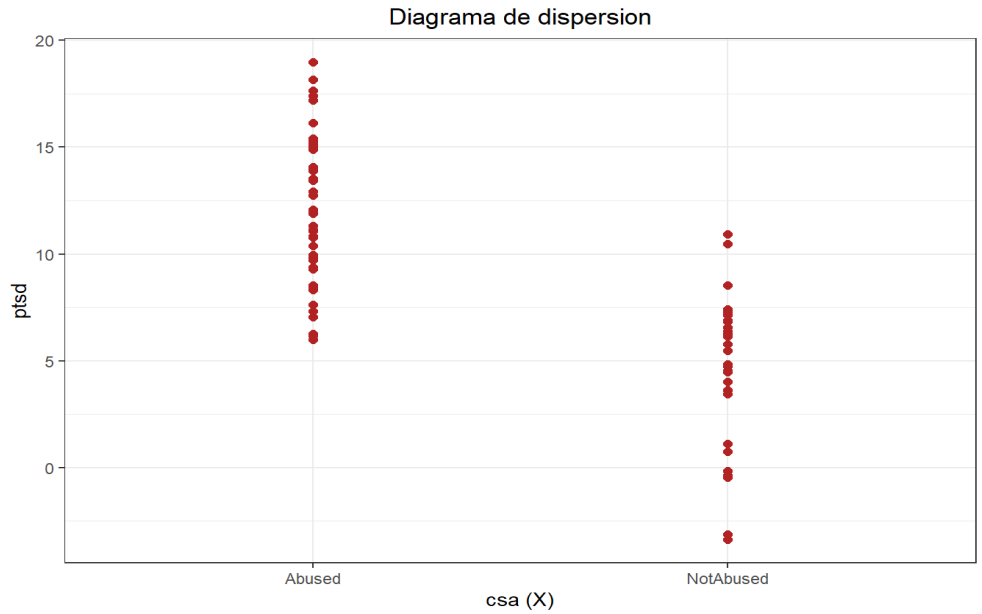


Figura 51. *Diagrama de dispersión.*

El diagrama de dispersión está en función de CSA (variable cualitativa). Observamos puntajes más altos de ptsd en el grupo Abused.

b) Estadísticos descriptivos para cada variable.

```
by(data = sexab, INDICES = sexab$csa, summary)
```

```
## sexab$csa: Abused
##                cpa             ptsd                csa
```

```
##  Min.    :-1.115    Min.    : 5.985    Abused    :45
##  1st Qu. : 1.415    1st Qu.  : 9.374    NotAbused: 0
##  Median : 2.627     Median :11.313
##  Mean    : 3.075    Mean    :11.941
##  3rd Qu. : 4.317    3rd Qu. :14.901
##  Max.    : 8.647    Max.    :18.993
## ------------------------------------------------------------
## sexab$csa: NotAbused
##                cpa              ptsd                csa
##  Min.    :-3.1204    Min.    :-3.349     Abused   : 0
##  1st Qu. :-0.2299    1st Qu. : 3.544    NotAbused:31
##  Median : 1.3216     Median : 5.794
##  Mean    : 1.3088    Mean    : 4.696
##  3rd Qu.: 2.8309     3rd Qu. : 6.838
##  Max.    : 5.0497    Max.    :10.914
```

Se observa que las víctimas de abusos tienen niveles promedio (11.941) más altos de estrés postraumático en comparación al promedio (4.696) de mujeres que no han sufrido abusos.

c) Diagrama de cajas

```
ggplot(data = sexab, aes(x = csa, y = ptsd, colour = csa)) + geom_boxplot () +
  theme_bw() + theme(legend.position = "none")
```

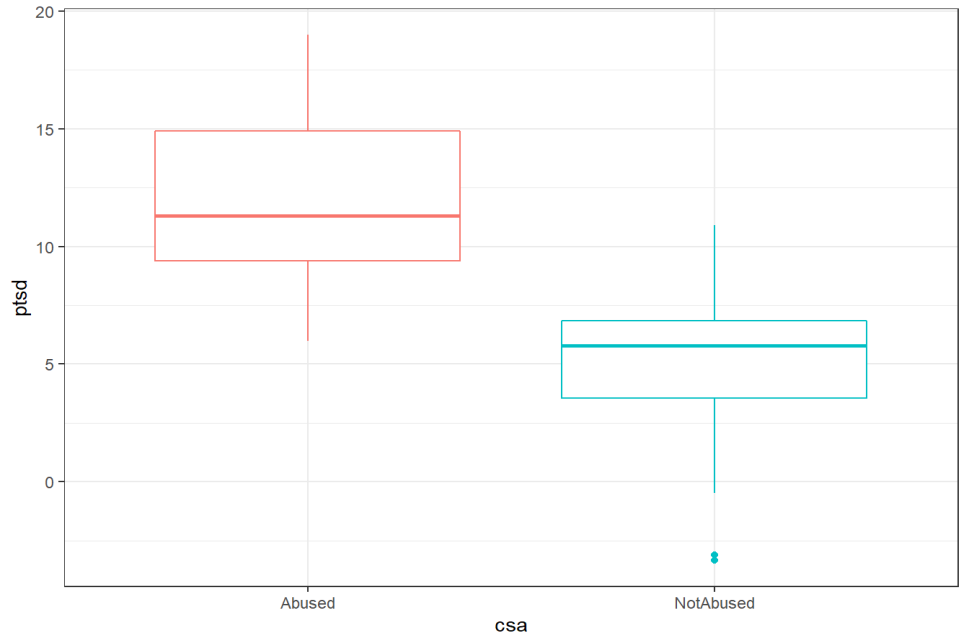


Figura 52. *Gráfico BoxPlot*

Se observa que las víctimas de abusos tienen niveles más altos de estrés postraumático en comparación a las mujeres que no han sufrido abusos(Amat, 2016).

Una forma de comparar si está diferencia es significativa, es mediante un t-test.

```
# Calculo de la varianza de cada grupo para determinar si son similares
aggregate(ptsd ~ csa, data = sexab, FUN = var)
```

```
##          csa     ptsd
## 1     Abused 11.83464
## 2 NotAbused 12.38859
```

```
t.test(formula = ptsd ~ csa, data = sexab, var.equal = TRUE)
```

```
##  Two Sample t-test
##
```

```
## data:  ptsd by csa
## t = 8.9387, df = 74, p-value = 2.172e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.630165 8.860273
## sample estimates:
##    mean in group Abused mean in group NotAbused
##               11.941093                4.695874
```

El contraste de hipótesis muestra una clara significancia estadística en la diferencia de medias $p(\ (2.172e-13) < \alpha(0.05)$. se conforma que el promedio de las mujeres abusadas en mayor al promedio de las mujeres no abusadas.

El modelo lineal puede obtenerse realizarse desde la perspectiva de un modelo lineal incluyendo la variable cualitativa como predictor. Para hacerlo, cada uno de los niveles del predictor se tiene que convertir en una variable dummy cuyo valor puede ser 0 o 1.

$$ptsd = \beta_0 + \beta_1 dummy_{abused} + \beta_2 dummy_{NotAbused}$$

$$dummy_i = \begin{cases} 0\ la\ observacion\ no\ pertenece\ al\ nivel\ i \\ 1\ la\ observacion\ pertenece\ al\ nivel\ i \end{cases}$$

Para cada observación, únicamente una de las variables dummy toma el valor 1.

En R la función lm() identifica automáticamente si un predictor es de tipo cualitativo y escoge como nivel de referencia el primero en base al orden alfabético.

d) Modelo lineal

```
# modelo lineal
modelo <- lm(ptsd ~ csa, data = sexab)
summary(modelo)
```

```
##
## Call:
## lm(formula = ptsd ~ csa, data = sexab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0451 -2.3123  0.0951  2.1645  7.0514
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.9411     0.5177  23.067  < 2e-16 ***
## csaNotAbused  -7.2452     0.8105  -8.939 2.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.473 on 74 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5127
## F-statistic:  79.9 on 1 and 74 DF,  p-value: 2.172e-13
```

$$y = 11.9411 - 7.2452(csaNotAbuded) = 11.9411$$

Esto indica que se categorizo la dummy como: Abused: 0, not abused: 1

Prediciendo:

notAbused:

$$y = 11.9411 - 7.2452(1) = 4.6959$$

Abused:

$$y = 11.9411 - 7.2452(0) = 11.9411$$

Si se desea especificar cuál debe ser el nivel de referencia empleado por lm(), se puede recurrir a la función modelo$xlevels y modelo$assign.

```
modelo$xlevels
```

```
## $csa
```

```
## [1] "Abused"    "NotAbused"
```

```
modelo$assign
```

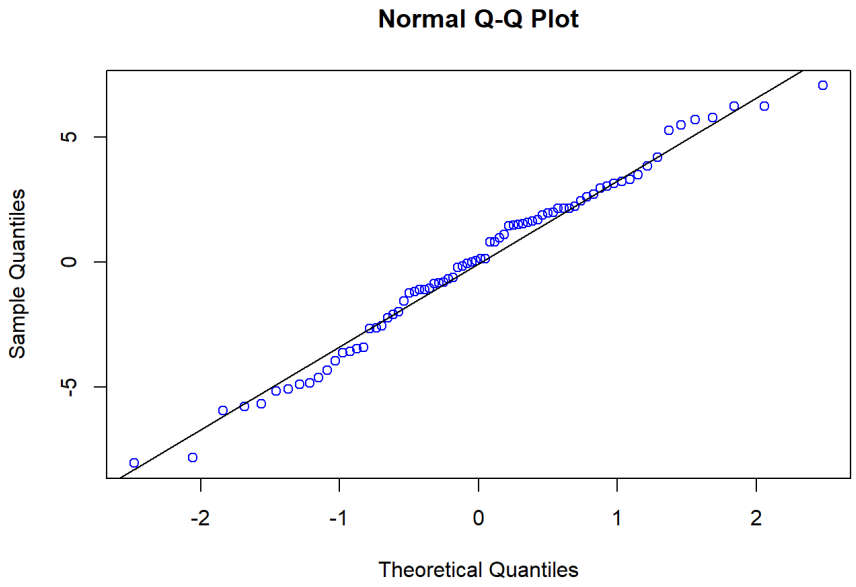
```
## [1] 0 1
```

```
# predecir valores de Y.
nuevos_datos <- data.frame(csa= seq(min(0),max(1)))
predict_value <- predict(modelo)
head(predict_value)
```

```
##        1        2        3        4        5        6
## 11.94109 11.94109 11.94109 11.94109 11.94109 11.94109
```

e) Observando la normalidad

```
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```




```
shapiro.test(modelo$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.98418, p-value = 0.465
```

Como $p(0.465) > \alpha(0.05)$, se observa normalidad en los residuos

f) Varianza constante

```
# Varianza constante de los residuos (Homocedasticidad):
# Test de Breush-Pagan
library(lmtest)
bptest(modelo)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo
```

```
## BP = 0.015186, df = 1, p-value = 0.9019
```

Como $p(0.9019) > \alpha(0.05)$, se observa homocedasticidad

g) Autocorrelación

```
# Autocorrelacion de residuos:
#test de Durwin Watson
library(lmtest)
dwtest(modelo)
```

```
##  Durbin-Watson test
##
## data:  modelo
## DW = 1.8303, p-value = 0.1948
## alternative hypothesis: true autocorrelation is greater than 0
```

Como $p(0.1948) > \alpha(0.05)$, se observa la no presencia de autocorrelación

Podríamos crear dos columnas para la variable dummy y obtener modelos, veamos.

Creando modelos utilizando transformaciones dummy

```
sexab$abused <- ifelse(test = sexab$csa == "Abused", yes = 1, no = 0)
sexab$not_abused <- ifelse(test = sexab$csa == "NotAbused", yes = 1, no = 0)
rbind(head(sexab, 3), tail(sexab, 3))
```

```
##         cpa    ptsd        csa    abused    not_abused
## 1   2.04786   9.71365     Abused       1          0
## 2   0.83895   6.16933     Abused       1          0
```

```
## 3  -0.24139  15.15926       Abused      1         0
## 74 -1.85753  -0.46996   NotAbused      0         1
## 75  2.85253   6.84304   NotAbused      0         1
## 76  0.81138   7.12918   NotAbused      0         1
```

Se puede observar que la información de las dos variables dummy es redundante, al haber solo dos niveles, y dado que solo uno de ellos puede tomar el valor 1, conociendo uno se conoce el otro. Para evitar que aparezca el problema de la singularidad al ajustar el modelo, una de las dos variables se excluye del modelo y se considera como el nivel de referencia (Amat, 2016).

h) Modelo lineal utilizando la variable dummy abused

```
modelo <- lm(ptsd ~ abused, data = sexab)
summary(modelo)
```

```
## Call:
## lm(formula = ptsd ~ abused, data = sexab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0451 -2.3123  0.0951  2.1645  7.0514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6959     0.6237   7.529 1.00e-10 ***
## abused        7.2452     0.8105   8.939 2.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.473 on 74 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5127
```

```
## F-statistic:  79.9 on 1 and 74 DF,  p-value: 2.172e-13
```

El modelo presenta el intercepto y la variable independiente abused significativas. El modelo se escribe como:

Para predecir abused:

$$y = 4.6959 + 7.2452(1) = 11.9411$$

```
# media de ptsd en mujeres víctimas de abusos
mean(sexab[sexab$csa == "Abused", "ptsd"])
```

```
## [1] 11.94109
```

Independientemente del nivel que se escoja como referencia, el resultado es el mismo. Lo único que cambia es el valor de la intersección, ya que cambia el nivel de referencia, y el signo de la pendiente.

i) Modelo lineal utilizando NotAbused

```
# modelo para not_abused
modelo <- lm(ptsd ~ not_abused, data = sexab)
summary(modelo)
```

```
## Call:
## lm(formula = ptsd ~ not_abused, data = sexab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0451 -2.3123  0.0951  2.1645  7.0514
##
## Coefficients:
##                 Estimate    Std. Error   t value   Pr(>|t|)
## (Intercept)     11.9411     0.5177    23.067    < 2e-16 ***
```

```
## not_abused     -7.2452     0.8105    -8.939    2.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.473 on 74 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5127
## F-statistic:  79.9 on 1 and 74 DF,  p-value: 2.172e-13
```

$$y = 11.9411 - 7.2452(1) = 4.6959$$

Es importante tener en cuenta que los *p-values* obtenidos por un *t-test* y por un modelo lineal que contenga un predictor cualitativo con dos niveles serán los mismos siempre y cuando el *t-test* no incluya una corrección de varianzas desiguales.

```r
# media de ptsd en mujeres no victimas de abusos
mean(sexab[sexab$csa == "NotAbused", "ptsd"])
```

```
## [1] 4.695874
```

```r
# predecir
nuevos_datos <- data.frame(csa= seq(min(0),max(1)))
predict_value <- predict(modelo)
head(predict_value)
```

```
##        1        2        3        4        5        6
## 11.94109 11.94109 11.94109 11.94109 11.94109 11.94109
```

```r
error = predict_value - sexab$ptsd
head(error)
```

```
##          1          2          3          4          5          6
##  2.2274429  5.7717629 -3.2181671  0.6283229  1.9872529  2.1022529
```

```
# error del modelo
sqrt(mean(error^2))
```

```
## [1] 3.426641
```

14. Ejemplo con varias categorías de respuesta en la variable cualitativa:

Caso tres niveles.

El siguiente ejemplo lo extraemos de Camacho, C. (2020), esta prueba es equivalente a la prueba de análisis de la varianza donde se estudia el efecto de una variable cualitativa de varias categorías con otra cuantitativa. Como se sabe, para aplicar el modelo de regresión lineal han de respetarse los supuestos del modelo: linealidad, normalidad y homocedasticidad. Los dos últimos son los mismos que los supuestos del análisis de la varianza.

La solución consiste en generar tantas variables independientes como categorías haya en el factor, y a continuación codificar cada una de estas variables con “ceros” y “unos” según la categoría a la que pertenezca los distintos sujetos.

La variable cualitativa FUMA fue codificada como:

0: no fuma

1: fumador

2: ex-fumador

Una solución podría ser crear tantas variables como categorías. No sirve porque serían combinación lineal y el modelo es irresoluble.

La solución es crear tantas variables como categorías menos 1, denominadas variables indicadoras con el siguiente esquema

	X1	X2	
No-fumador	0	0	categoría de referencia
Fumador1	0		
Ex-fumador	0	1	

Las variables X1 y X2 ya no son combinación lineal y, por tanto, el modelo es resoluble. El modelo quedaría como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

β_0: es el valor de Y, cuando X1 y X2 son ambas cero

El modelo cuando X1 es =0, queda como:

$$Y = \beta_0 + \beta_0 X_2$$

El modelo cuando X1 es =1, queda como:

$$Y = \beta_0 + \beta_1 X_2$$

Ejemplo:

Los siguientes datos proceden de una muestra hipotética, sobre presión arterial en cm de Hg y "status" de fumador, codificado como 0: no-fumador, 1: fumador y 2: ex-fumador. Discutir el modelo de regresión entre presión arterial y "status" de fumador y estimar por intervalos la presión arterial media según el "status" de fumador, a partir de los resultados del modelo más adecuado.

```
# Predictores numericos y 3 categoricos
# presion arterial y fumadores
datos<-data.frame(presion=c(15, 19, 16.3, 22, 18, 19.8, 23.2, 14.4, 20
.3,
                            22, 20.5, 19, 12.7, 14, 11.8, 11.2, 14
,
                            19.5, 22.3, 15, 12.6, 16.4, 13.5, 13.7
),
```

```
                    tabaco = c(0,2,1,1,2,0,1,0,2,1,2,2,0,0,0,2,0,1,1,0
,2,0,2,1))

head(datos, 6)
```

```
##   presion tabaco
## 1    15.0      0
## 2    19.0      2
## 3    16.3      1
## 4    22.0      1
## 5    18.0      2
## 6    19.8      0
```

```
str(datos)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ presion: num  15 19 16.3 22 18 19.8 23.2 14.4 20.3 22 ...
##  $ tabaco : num  0 2 1 1 2 0 1 0 2 1 ...
```

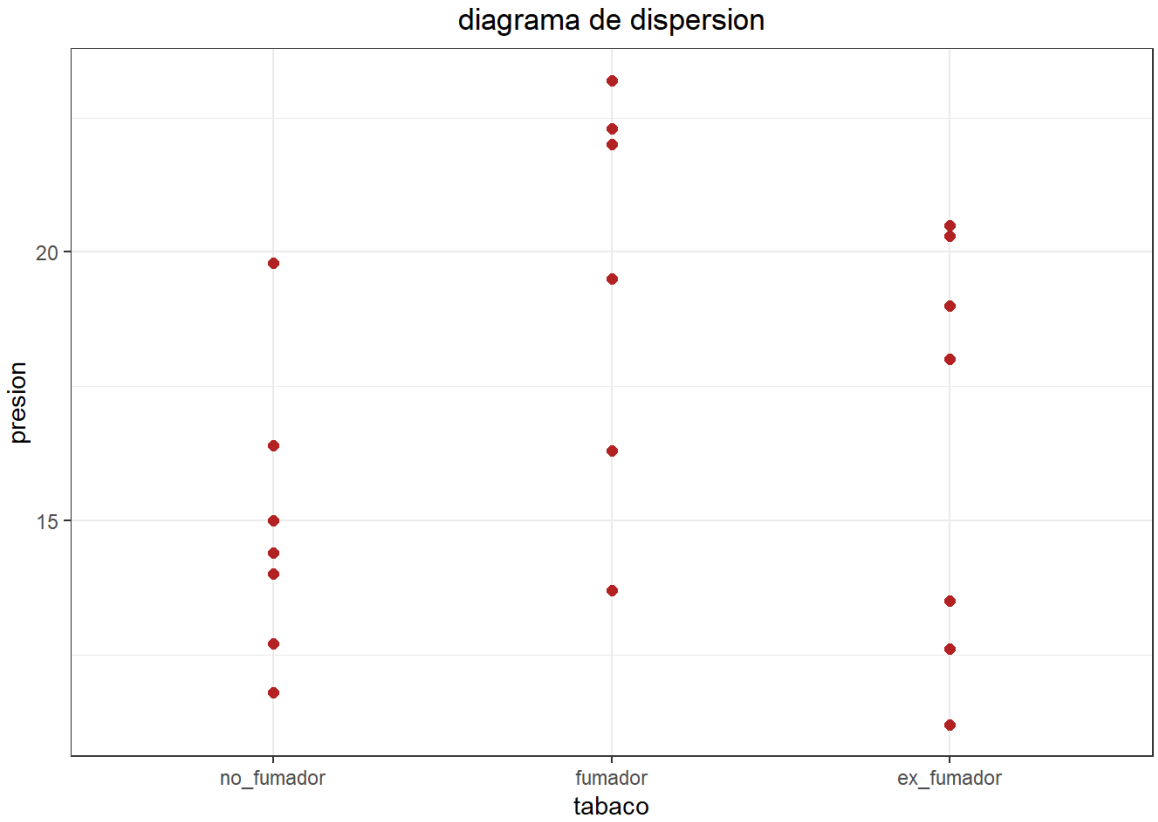
```
datos$tabaco <- factor(datos$tabaco, levels=c(0,1,2), labels=c("no_fum
ador", "fumador", "ex_fumador"))

str(datos)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ presion: num  15 19 16.3 22 18 19.8 23.2 14.4 20.3 22 ...
##  $ tabaco : Factor w/ 3 levels "no_fumador","fumador",..: 1 3 2 2 3
1 2 1 3 2 ...
```

a) Diagrama de dispersión

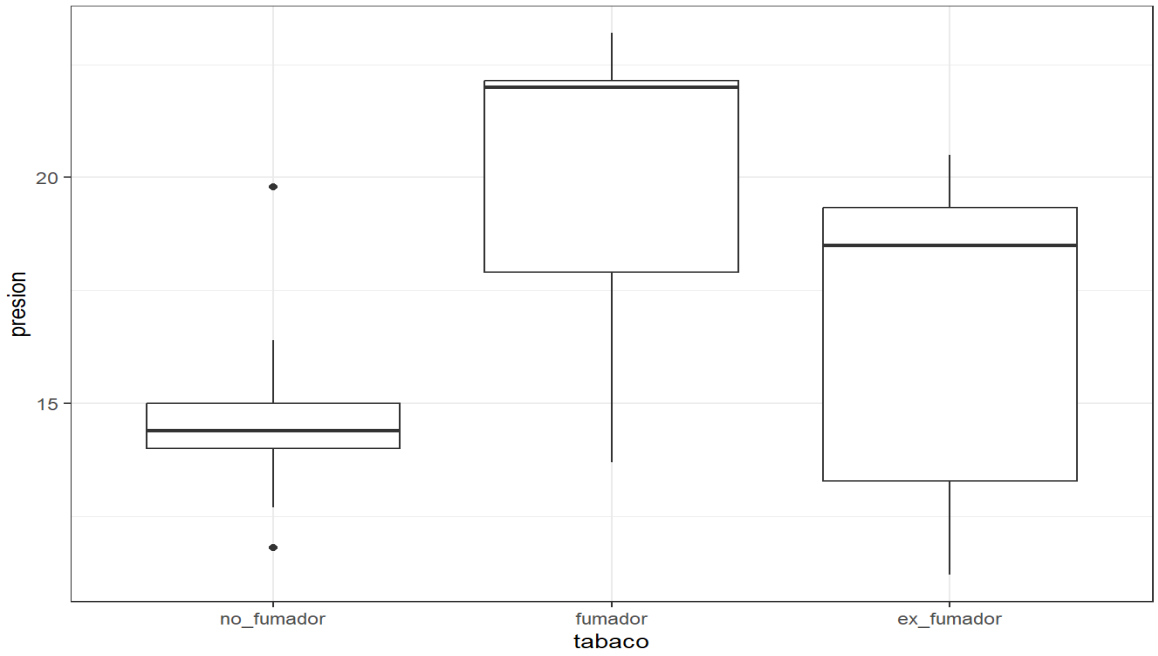
```
# Diagrama de dispersión

library(ggplot2)

ggplot(data = datos, mapping = aes(x = tabaco, y = presion)) +
```

```
  geom_point( color="firebrick", size = 2) +
  (labs(title = "diagrama de dispersion", x = "tabaco")) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



b) Diagrama de cajas

```
# diagrama de cajas
ggplot(data = datos, aes(x = tabaco, y = presion)) +
  geom_boxplot () +
  theme_bw() +
  theme(legend.position = "none")
```



c) Descripción de datos

```
# descriptiva segun la variable cualitativa
by(data = datos, INDICES = datos$tabaco, summary)
```

```
## datos$tabaco: no_fumador
##     presion             tabaco
##  Min.   :11.80   no_fumador:9
##  1st Qu.:14.00   fumador   :0
##  Median :14.40   ex_fumador:0
##  Mean   :14.79
##  3rd Qu.:15.00
##  Max.   :19.80
## ------------------------------------------------------------
## datos$tabaco: fumador
##     presion             tabaco
##  Min.   :13.70   no_fumador:0
##  1st Qu.:17.90   fumador   :7
##  Median :22.00   ex_fumador:0
##  Mean   :19.86
##  3rd Qu.:22.15
##  Max.   :23.20
```

```
## ------------------------------------------------------------
## datos$tabaco: ex_fumador
##     presion             tabaco
##  Min.   :11.20   no_fumador:0
##  1st Qu.:13.28   fumador   :0
##  Median :18.50   ex_fumador:8
##  Mean   :16.76
##  3rd Qu.:19.32
##  Max.   :20.50
```

d) Modelo lineal

```
# modelo lineal
modelo <- lm(presion ~ tabaco, data = datos)
summary(modelo)
```

```
##
## Call:
## lm(formula = presion ~ tabaco, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1571 -2.3139  0.2111  2.2375  5.0111
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        14.789      1.070  13.820 5.17e-12 ***
## tabacofumador       5.068      1.618   3.133  0.00503 **
## tabacoex_fumador    1.974      1.560   1.265  0.21968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.21 on 21 degrees of freedom
## Multiple R-squared:  0.3192, Adjusted R-squared:  0.2543
## F-statistic: 4.922 on 2 and 21 DF,  p-value: 0.01766
```

Según el modelo, la categoría de respuesta no fumador, es la de referencia

	X1	X2	
No-fumador	0	0	categoría de referencia
Fumador	1	0	
Ex-fumador	0	1	

Modelos:

no-fumador,

$$y = 14.789 + 5.068(tabacoFumador) + 1.974(tabacoExfumador)$$
$$= 14.789 + 5.068(0) + 1.974(0)$$

fumador

$$y = 14.789 + 5.068(tabacoFumador) + 1.974(tabacoEx\_fumador$$
$$= 14.789 + 5.068(1) + 1.974(0)$$

Ex_fumador.

$$y = 14.789 + 5.068(tabacoFumador) + 1.974(tabacoEx\_fumador$$
$$= 14.789 + 5.068\ (0) + 1.974(1)$$

```
# media de presion en no fumadores
mean(datos[datos$tabaco == "no_fumador", "presion"])
```

```
## [1] 14.78889
```

```
# media de presion en fumadores
mean(datos[datos$tabaco == "fumador", "presion"])
```

```
## [1] 19.85714
```

```
# media de presion en fumadores
mean(datos[datos$tabaco == "ex_fumador", "presion"])
```

```
## [1] 16.7625
```

```
confint(modelo)
```

```
##                       2.5 %    97.5 %
## (Intercept)       12.563400 17.014377
## tabacofumador      1.703632  8.432876
## tabacoex_fumador -1.270568  5.217790
```

e) Normalidad de errores

```
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```

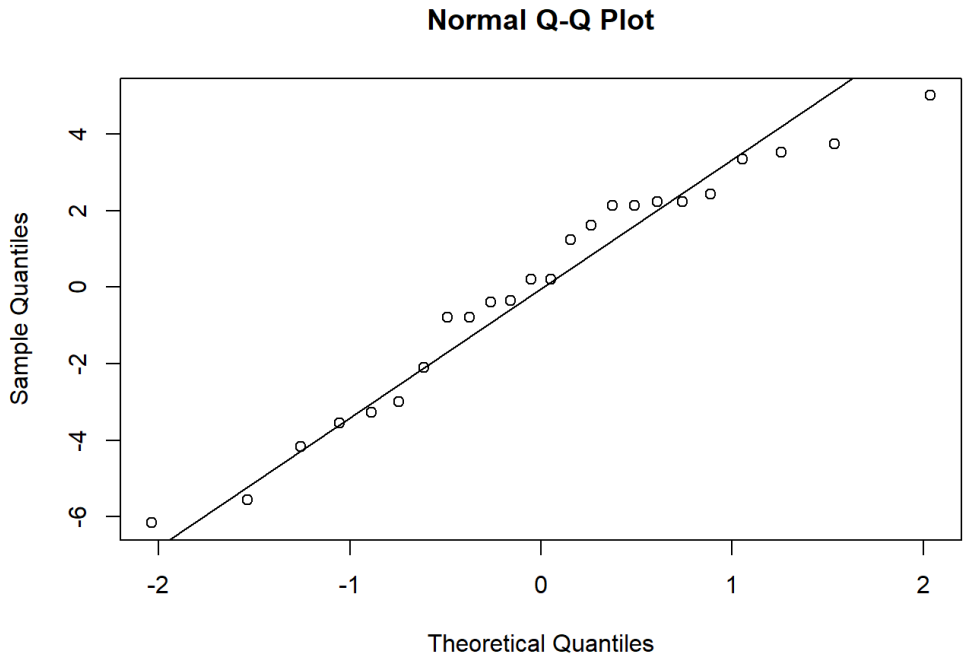


Figura 55. Diagrama QQpplot de normalidad

```
shapiro.test(modelo$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.95611, p-value = 0.3654
```

f) Varianza constante

```
# Varianza constante de los residuos (Homocedasticidad):
# Test de Breush-Pagan
library(lmtest)
```

```
bptest(modelo)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 2.7518, df = 2, p-value = 0.2526
```

g) Autocorrelación

```
# Autocorrelacion de residuos:
#test de Durwin Watson
library(lmtest)
dwtest(modelo)
```

```
#
##  Durbin-Watson test
```

```
##
## data:  modelo
## DW = 1.2768, p-value = 0.03784
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# predecir
nuevos_datos <- data.frame(tabaco= seq(min(0),max(2)))
predict_value <- predict(modelo)
head(predict_value)
```

```
##        1        2        3        4        5        6
## 14.78889 16.76250 19.85714 19.85714 16.76250 14.78889
```

```
error = predict_value - datos$presion
head(error)
```

```
##          1          2          3          4          5          6
## -0.2111111 -2.2375000  3.5571429 -2.1428571 -1.2375000 -5.0111111
```

```
##          1          2          3          4          5          6
## -0.2111111 -2.2375000  3.5571429 -2.1428571 -1.2375000 -5.0111111
```

h) Error (MSE)

```
# error del modelo (MSE)
sqrt(mean(error^2))
```

```
## [1] 3.003087
```