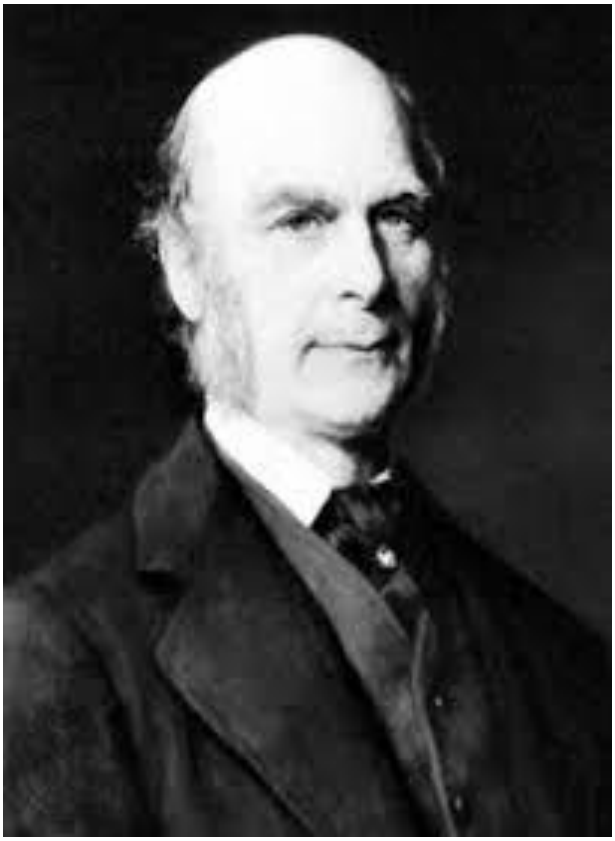
Unidad I

ANÁLISIS DE CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

"Es verdaderamente absurdo ver como un número limitado de observaciones pueden convertirse, en manos de hombres, en ideas preconcebidas"
Francis Galton

Introducción

El término regresión fue utilizado por primera vez como un concepto estadístico en 1877 por sir Francis Galton, quien llevó a cabo un estudio que mostró que la estatura de los niños nacidos de padres altos tiende a retroceder o "regresar" hacia la estatura media de la población. Galton usó esta ecuación para explicar el fenómeno que los hijos de padres altos tienden a ser altos pero no tan altos como sus padres mientras los hijos de padres bajos tienden a ser bajos pero no tan bajos como sus padres. Este efecto se llama el efecto de la regresión. Designó la palabra regresión como el nombre del proceso general de predecir una variable (la estatura de los niños) a partir de otra (la estatura del padre o de la madre).


Francis Galton (1822-1911)

El análisis de regresión es una técnica estadística utilizada para investigar, modelar o explicar a una variable respuesta (Y) mediante una o muchas variables predictoras (Xs)
La variable respuesta (Y) también es llamada variable dependiente, variable endógena o variable target.
Las variables predictoras (X) también son llamadas variables independientes, variables exógenas, variables regresoras o variables explicativas.
Por ejemplo, nos preguntamos ¿el peso (en kg) de una persona puede ser explicado por su estatura (en cm)?
¿Qué otras variables podrían explicar el peso de una persona?

La presente unidad tiene como objetivo discutir los conceptos de correlación y regresión lineal simple

1. Coeficiente de correlación

La correlación es una medida de asociación entre dos variables.
El estimador del parámetro ρ es el coeficiente de correlación muestral

$$r = \frac{\sum_{i=1}^{n} y_i(\mathrm{x}_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Si bien la regresión y la correlación tienen una relación estrecha, la regresión es más poderosa en muchos casos.
La correlación de Pearson sólo es una medida de la asociación lineal, y tiene poco uso en predicción, sin embargo, los métodos de regresión se pueden aplicar para desarrollar relaciones cuantitativas entre las variables, y esas relaciones se pueden usar en predicciones. Es decir, la existencia de correlación no implica causalidad de una variable con respecto a la otra.

El coeficiente de correlación toma valores entre -1 y 1.
Si r es cercano a 1 se dice que existe una fuerte relación lineal directa entre X e Y.
Si r es cercano a -1 se dice que existe una fuerte relación lineal indirecta entre X e Y.
Si r es cercano a 0 se dice que no existe una relación lineal entre X e Y

Con frecuencia es útil probar la hipótesis que el coeficiente de correlación es cero, esto es:
H_0: ρ = 0
H_1: ρ ≠ 0

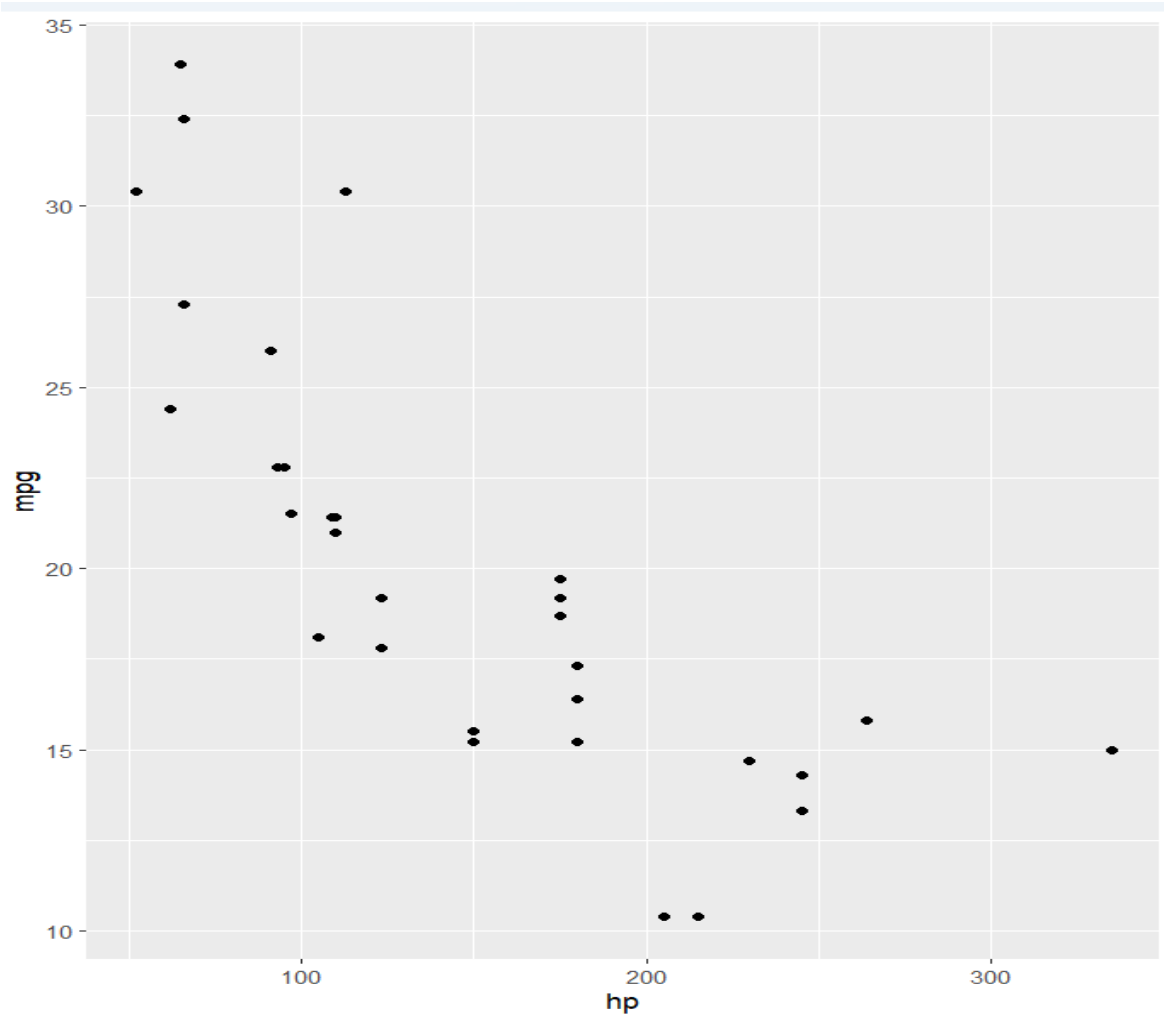
Para la hipótesis anterior, el estadístico de prueba adecuado es:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

Se rechaza si $|t_0| > t_{(1-\alpha/2,n-2)}$

Ejemplo
Se utilizará el conjunto de datos mtcars de R, que está compuesta por características de diferentes modelos y marcas de vehículos que son utilizadas para explicar el rendimiento en millas por galón (mpg) de dichos vehículos.

```
data("mtcars")
mtcars
library(ggplot2)
data(mtcars)
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point()
```




```
#Tamaño de muestra
n<-nrow(mtcars)
n
[1] 32

#Cálculo de la correlación
r<-cor(mtcars$mpg,mtcars$hp,method = "p")
r
[1] -0.7761684

#Comprobación del cálculo del estadístico
tcal<-r*sqrt(n-2)/sqrt(1-r^2)
tcal
[1] -6.742389

#Valores críticos
qt(c(0.05/2,1-0.05/2),n-2)
#Cálculo de pvalor
2*pt(tcal,n-2)
[1] 1.787835e-07
```

```
#Evaluación de la significancia
cor.test(mtcars$mpg,mtcars$hp,alternative="t",method = "p")
```

```
     Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$hp
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8852686 -0.5860994
sample estimates:
       cor
-0.7761684
```

```
##Probar si es significativamente inverso
#Valor crítico
qt(0.05,n-2)
#Cálculo de pvalor
pt(-6.7424,n-2)

cor.test(mtcars$mpg,mtcars$hp,alternative="l",method = "p")
```

```
Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$hp
t = -6.7424, df = 30, p-value = 8.939e-08
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.0000000 -0.6231988
sample estimates:
       cor
-0.7761684
```

En general, el procedimiento de prueba de hipótesis para un valor hipotético ρ_0 es:
H_0: $\rho = \rho_0$
H_1: $\rho \neq \rho_0$

Para muestras moderadamente grandes (n>25):

$$Z = arctanh\, r = \frac{1}{2} ln \frac{1+r}{1-r}$$

Tiene distribución aproximadamente normal, con media

$$\mu_Z = arctanh\,\rho = \frac{1}{2} ln \frac{1+\rho}{1-\rho}$$

Y varianza

$$\sigma_Z^2 = (n-3)^{-1}$$

Así, para probar la hipótesis H_0: $\rho = \rho_0$ se calcula el estadístico

$$Z_0 = (arctanh\, r - arctanh\, \rho_0)\sqrt{(n-3)}$$

Se rechaza si $|Z_0| > Z_{(1-\frac{\alpha}{2})}$

Por ejemplo, si se quiere probar:
H_0: $\rho = -0.8$
H_1: $\rho \neq -0.8$

```
z<-(atanh(r)-atanh(-0.8))*sqrt(n-3)
[1] 0.3390095

qnorm(c(0.05/2,1-0.05/2))
[1] -1.959964  1.959964

2*(1-pnorm(abs(z)))
[1] 0.7346026
```

Pero la función cor.test no permite realizar esta hipótesis, por lo que se elaboró la siguiente función:
Función para evaluar la correlación con cualquier valor hipotético

```
cor_test_rho <- function(x, y, rho_h0, alternative =
c("two.sided", "less", "greater")) {
  alternative <- match.arg(alternative)
  n <- length(x)
  r <- cor(x, y)
  z_obs <- atanh(r)
  z_h0 <- atanh(rho_h0)
  se <- 1 / sqrt(n - 3)
  z <- (z_obs - z_h0) / se
  if (alternative == "two.sided") {
    p_value<- felse(z>0,2*pnorm(z,lower.tail=F),2*pnorm(z))
  } else if (alternative == "less") {
    p_value <- pnorm(z)
  } else if (alternative == "greater") {
    p_value <- 1 - pnorm(z)
  }
    return(list(Zcal=z,pvalor=p_value))
}
```

```
cor_test_rho(mtcars$mpg,mtcars$hp,rho_h0=-
0.8,alternative="two.sided)
```

También se puede establecer un intervalo de confianza 100(1-α)% para ρ siendo este:

$$tanh\left(arctanh\, r\ - \frac{Z_{1-\alpha/2}}{\sqrt{(n-3)}}\right) \leq \rho \leq tanh\left(arctanh\, r\ + \frac{Z_{1-\alpha/2}}{\sqrt{(n-3)}}\right)$$

```
cor.test(mtcars$mpg,mtcars$hp,conf.level=0.95,method="p",al
ternative="t")
```

```
Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$hp
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8852686 -0.5860994
sample estimates:
       cor
-0.7761684
```

```
r<-cor(mtcars$mpg,mtcars$hp)
n<-nrow(mtcars)
#hallamos el limite inferior del IC
LI<-tanh((atanh(r)-(qnorm((1-0.05/2)))/sqrt(n-3)))
LI
[1] -0.8852686

#Hallamos el limite superior del IC
LS<-tanh((atanh(r)+(qnorm((1-0.05/2)))/sqrt(n-3)))
[1] -0.5860994

c(LI,LS)

[1] -0.8852686 -0.5860994
```

El intervalo que va de -0.885 a-0.586 brinda un 95% de confianza de contener a la correlación entre el rendimiento en millas por galón y los caballos de fuerza.

2. Análisis de Regresión Lineal Simple

Supongamos que recolectamos datos de 32 elementos elegidos al azar y los representamos en un gráfico denominado Diagrama de Dispersión como se puede apreciar a continuación:

Gráfico de Dispersión

```
plot(mtcars$hp,mtcars$mpg,xlab="hp",ylab="mpg",
     main="Gráfico de dispersión",col="red")
```

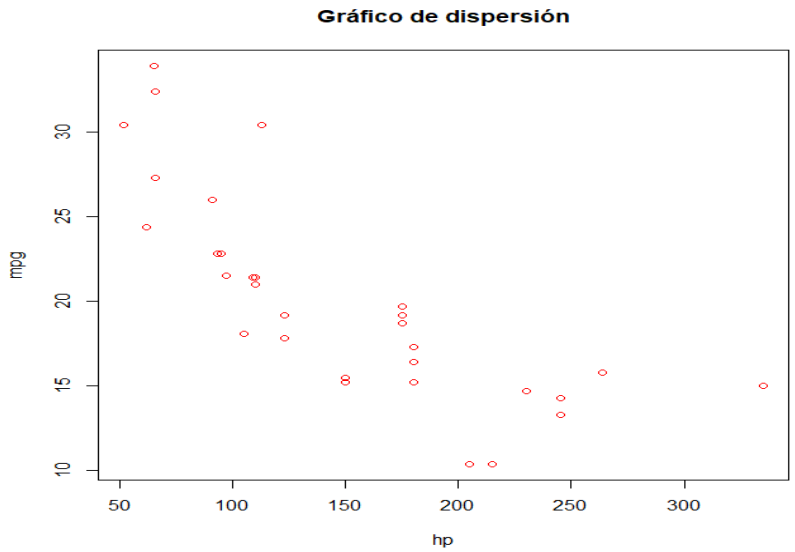
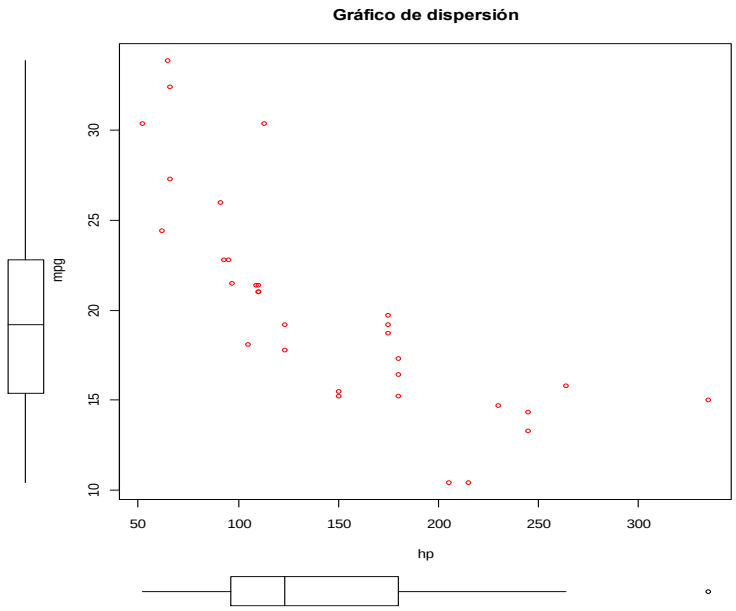


Gráfico Marginal

```
car::scatterplot(mpg ~ hp, data = mtcars, regLine = F, grid
= F, smooth = F, main = "Gráfico de dispersión", col = "red
")
```



Otras formas de obtener el gráfico marginal

Primera forma

```
data(mtcars)
library(ggplot2)
par(fig = c(0, 0.8, 0, 0.8), new = TRUE)
plot(mtcars$hp, mtcars$mpg, xlab = "Caballos de fuerza", yl
ab = "Millas por galon")
par(fig = c(0, 0.8, 0.55, 1), new = TRUE)
boxplot(mtcars$hp, horizontal = TRUE, axes = FALSE)
par(fig = c(0.65, 1, 0, 0.8), new = TRUE)
boxplot(mtcars$mpg, axes = FALSE)
mtext("Gráfico marginal", side = 3, outer = TRUE, line = -3
)
```

Segunda forma

```
library(ggplot2)
library(ggExtra)
data(mtcars)
p <- ggplot(mtcars, aes(x = hp, y = mpg)) + geom_point(shap
e = 2) +
  theme_minimal() +
  labs(title = "Relación entre Caballos de Fuerza y Millas
por Galón",
       x = "Caballos de Fuerza (hp)",
       y = "Millas por Galón (mpg)")
ggMarginal(p, type = "boxplot")
```

La ecuación que relaciona a estas dos variables es:

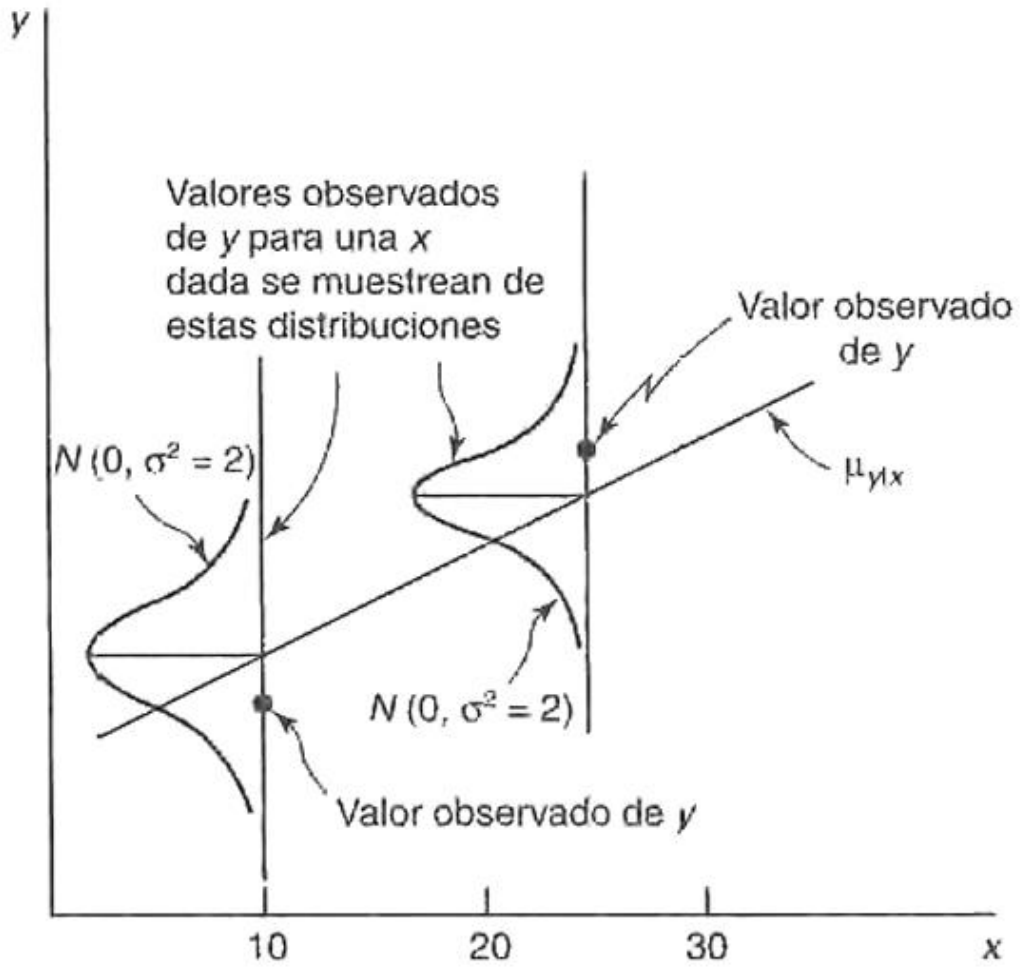
$$y = \beta_0 + \beta_1 X$$

donde β_0 es la ordenada en el origen y β_1 la pendiente. Ahora se puede observar que los datos no caen exactamente sobre la recta por lo que la ecuación anterior se debe modificar para tomar en cuenta esto.
Se define la diferencia entre el valor observado de *y* y el de la línea recta ($\beta_0 + \beta_1 X$) un error ε que es una variable aleatoria que explica porque el modelo no ajusta exactamente a los datos. Este error puede estar formado por los efectos de otras variables. Un modelo más real sería:

$$y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

Donde $\varepsilon_i \sim N(0, \sigma^2)$
En una primera fase del análisis de regresión se debe estimar los parámetros desconocidos del modelo de regresión. También se le llama a este proceso ajuste del modelo con los datos. Uno de los métodos que permite cumplir con este objetivo es el Método de Mínimos Cuadrados.

La siguiente fase del análisis de regresión se llama comprobación de la adecuación del modelo en donde se estudia lo apropiado del modelo y la calidad del ajuste determinado. Mediante esos análisis se puede determinar la utilidad del modelo de regresión. El resultado de la comprobación de adecuación puede indicar que el modelo es razonable, o que debe modificarse el ajuste original. Por lo anterior, el análisis de regresión es un procedimiento iterativo, en el que los datos conducen a un modelo, y se produce un ajuste del modelo a los datos. A continuación, se investiga la calidad del ajuste y se pasa a modificar el modelo, o el ajuste, o a adoptar el modelo.



3. El Modelo de Regresión Lineal Simple

Es un modelo con un solo regresor x donde la relación con la variable respuesta y es una línea recta. Este modelo es:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

donde β_0 es la ordenada en el origen y β_1 la pendiente son constantes desconocidas denominados coeficientes de regresión y ε es un componente aleatorio de error.

La pendiente β_1 es el cambio de la media de la distribución de *y* producido por un cambio unitario en x.
Si el intervalo de los datos incluye a x = 0, entonces la ordenada al origen, β_0, es la media de la distribución de la respuesta *y* cuando x = 0. Si no incluye al cero, β_0 no tiene interpretación práctica.

Se supone que los errores no están correlacionados, tienen media cero y varianza σ^2 desconocida.
La media de y para cada valor posible de x es

$$E(y/x) = \beta_0 + \beta_1 X$$

Y la varianza es:

$$V(y/x) = var(\beta_0 + \beta_1 X + \varepsilon) = \sigma^2$$

4. Estimación de los parámetros por Mínimos Cuadrados

Supongamos que tenemos n pares de datos (y_1,x_1), (y_2,x_2),…,(y_n,x_n), para estimar β_0 y β_1 se busca que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima
Podemos definir el modelo estimado por los n pares de la siguiente manera:

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{i=1,2,…,n}$$

Por lo que se quiere minimizar

$$\sum_{i=1}^{n} \varepsilon_i^2$$

Donde: $\varepsilon_i = y_i - \beta_0 - \beta_1 X_i$, es decir:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$$

Recordar:
Para encontrar el máximo o mínimo local, primero se debe encontrar la primera derivada de la función. Los valores de x que hacen que la primera derivada sea igual a 0 son puntos críticos. Si la segunda derivada en x=c es positiva, entonces f(c) es un mínimo. Cuando la segunda derivada es negativa en x=c, entonces f(c) es máxima.

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0 \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0 \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i) x_i = 0$$

De la primera ecuación se tiene:

$$-\sum_{i=1}^{n} \mathrm{y}_i + n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} \mathrm{y}_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

De la segunda ecuación se tiene

$$-\sum_{i=1}^{n} x_i \mathrm{y}_i + \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i \mathrm{y}_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

Las ecuaciones anteriores son llamadas ecuaciones normales de mínimos cuadrados

De la primera ecuación si se divide entre n a ambos lados se tiene

$$\frac{\sum_{i=1}^{n} y_i}{n} = \frac{n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i}{n}$$

De donde se obtiene $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ por lo tanto:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Por otro lado,

$$\sum_{i=1}^{n} y_i\, x_i = \left(\frac{\sum_{i=1}^{n} y_i - \widehat{\beta_1} \sum_{i=1}^{n} x_i}{n} \right) \sum_{i=1}^{n} x_i + \widehat{\beta_1} \sum_{i=1}^{n} x_i^2$$

$$n \sum_{i=1}^{n} y_i\, x_i = n \left(\frac{\sum_{i=1}^{n} y_i - \widehat{\beta_1} \sum_{i=1}^{n} x_i}{n} \right) \sum_{i=1}^{n} x_i + n\widehat{\beta_1} \sum_{i=1}^{n} x_i^2$$

Multiplicamos por n para no trabajar con fracciones

$$n \sum_{i=1}^{n} y_i\, x_i = n \left(\frac{\sum_{i=1}^{n} y_i - \widehat{\beta_1} \sum_{i=1}^{n} x_i}{n} \right) \sum_{i=1}^{n} x_i + n\widehat{\beta_1} \sum_{i=1}^{n} x_i^2$$

$$n \sum_{i=1}^{n} y_i\, x_i = \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \widehat{\beta_1} \left(\sum_{i=1}^{n} x_i \right)^2 + n\widehat{\beta_1} \sum_{i=1}^{n} x_i^2$$

$$n \sum_{i=1}^{n} y_i\, x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i = \widehat{\beta_1} \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right]$$

Dividimos nuevamente sobre n para darle forma:

$$\sum_{i=1}^{n} y_i\, x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i \frac{1}{n} = \widehat{\beta_1}\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \frac{1}{n}\right]$$

Despejando $\widehat{\beta_1}$:

$$\frac{\sum_{i=1}^{n} y_i\, x_i - \frac{(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i)}{n}}{\left[\sum_{i=1}^{n} x_i^2 - (\frac{\sum_{i=1}^{n} x_i)^2}{n}\right]} = \widehat{\beta_1}$$

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

Una forma práctica de escribir la pendiente es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(\mathrm{x}_i - \bar{x})}{\sum_{i=1}^{n}(\mathrm{x}_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

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

Operando el numerador:

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{x} y_i \\ &= \sum_{i=1}^{n} y_i (x_i - \bar{x})\end{aligned}$$

Operando el denominador:

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} &= \sum_{i=1}^{n} x_i^2 - \frac{(\bar{x}n)^2}{n} \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} (x - \bar{x})^2\end{aligned}$$

Finalmente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

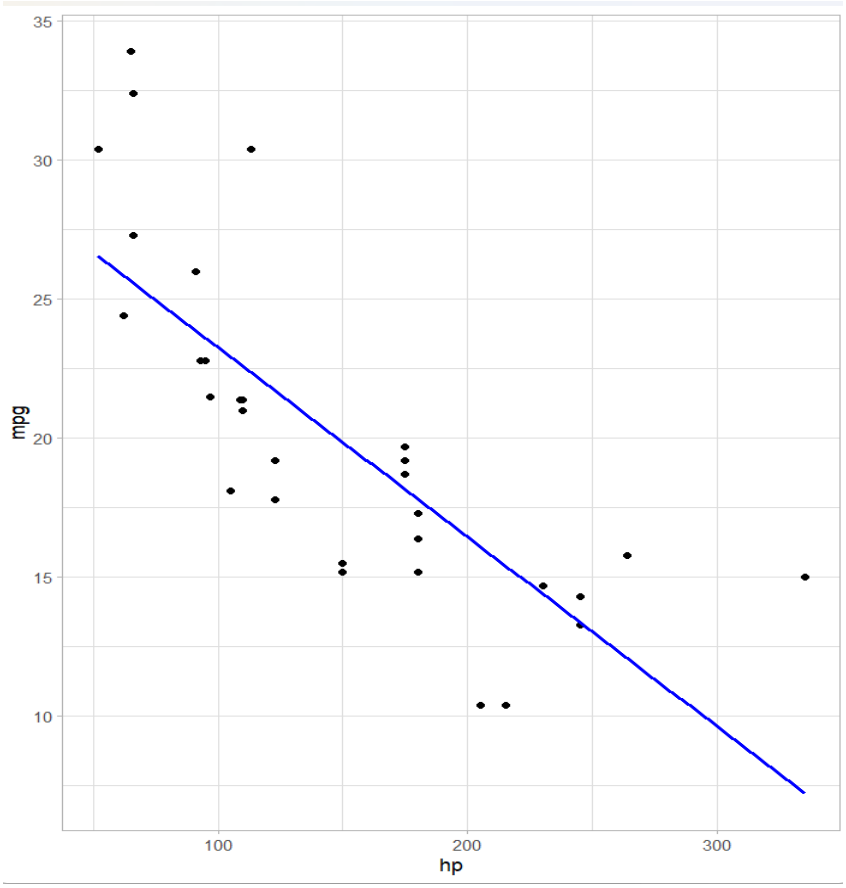
Con lo cual el modelo estimado queda representado:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

La diferencia entre el valor observado y_i y el valor estimado correspondiente $\hat{y}_i$ se llama residual

$$e_i = y_i - \hat{y}_i \quad i = 1,2,\ldots,n$$

```
library(ggplot2)
data("mtcars")
ggplot(mtcars, aes(x= mtcars$hp, y= mtcars$mpg)) +
  geom_point() +
  geom_smooth(method='lm', formula=y~x, se=FALSE, col='blue
') +
  theme_light()
```

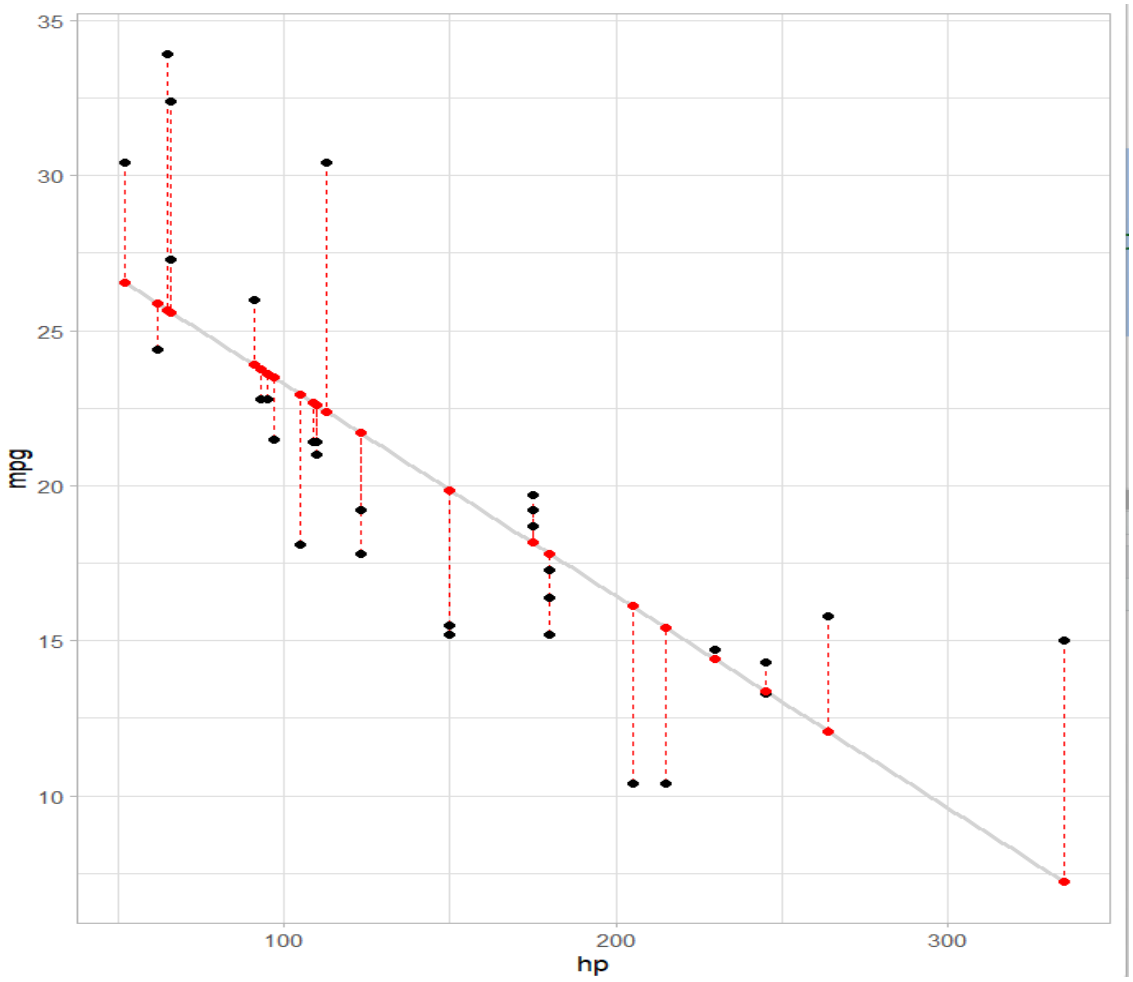



```
x<-mtcars$hp
y<-mtcars$mpg
n<-length(x)
b1<-(sum(x*y)-n*mean(x)*mean(y))/(sum(x^2)-n*mean(x)^2)
b0<-mean(y)-b1*mean(x)

modelo<-lm(mpg~hp,data=mtcars)
modelo
```

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

```
Coefficients:
(Intercept)           hp
   30.09886     -0.06823
```

```
ggplot(mtcars, aes(x=hp, y=mpg)) +
  geom_smooth(method="lm", se=FALSE, color="lightgrey") +
  geom_segment(aes(xend=hp, yend=modelo$fit), col='red', lty='dashed') +
  geom_point() +
  geom_point(aes(y=modelo$fit), col='red') +
  theme_light()
```



4.1 Propiedades de los estimadores mínimos cuadráticos

$\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las observaciones y_i

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} y_i c_i$$

Donde $c_i = \frac{(x_i - \bar{x})}{S_{xx}}$ para i=1,2,…,n

Se demostrará que $\hat{\beta}_1$ es insesgado

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} y_i c_i\right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

Se puede demostrar que $\sum_{i=1}^{n} c_i = 0$ y $\sum_{i=1}^{n} c_i x_i = 1$, entonces

$$E(\hat{\beta}_1) = \beta_1$$

De igual manera se demuestra que:

$$E(\hat{\beta}_0) = \beta_0$$

$$\begin{aligned} E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) &= E\left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \bar{x}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(y_i) - E(\hat{\beta}_1)\bar{x} \\ &= \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\ &= \beta_0 \end{aligned}$$

La varianza de $\hat{\beta}_1$ se calcula de la siguiente manera:

$$V(\hat{\beta}_1) = V\left(\sum_{i=1}^{n} y_i c_i\right) = \sum_{i=1}^{n} c_i^2 V(y_i)$$

Ya que las observaciones y_i son no correlacionadas, la varianza de la suma es igual a la suma de las varianzas y se ha supuesto que $V(y_i) = \sigma^2$ se tiene:

$$V(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

$$V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1 \bar{x}) = V(\bar{y}) + x^2(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} y_i\right) + \bar{x}^2 \frac{\sigma^2}{S_{xx}} - 0$$

$$\frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Demostrando:

$$\begin{aligned} Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\sum_{i=1}^{n} \frac{y_i}{n}, \sum_{j=1}^{n} c_j y_j\right) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{c_j}{n} Cov(y_i, y_j) = \sum_{k=1}^{n} \frac{c_k}{n} V(y_k) \\ &= \frac{\sigma^2}{n} \sum_{k=1}^{n} c_k \end{aligned}$$

Y como:

$$\sum_{i=1}^{n} c_i = 0$$

Entonces:

$$Cov(\bar{y}, \hat{\beta}_1) = 0$$

4.2 Propiedades del ajuste de mínimos cuadrados

a) La suma de los residuales en cualquier modelo de regresión que contenga una ordenada al origen β_0 siempre es igual a cero, esto es:

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

```
modelo<-lm(mpg~hp,data=mtcars)
sum(modelo$residuals)
```

b) La suma de los valores observados y_i es igual a la suma de los valores estimados $\hat{y}_i$

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

```
sum(mtcars$mpg)
sum(modelo$fitted.values)
```

c) La línea de regresión de mínimos cuadrados siempre pasa por el centroide de los datos que es el punto $(\bar{x}, \bar{y})$

```
xbar<-mean(mtcars$hp)
sum(modelo$coefficients*c(1,xbar))
ybar<-mean(mtcars$mpg)
```

d) La suma de los residuales, ponderados por el valor correspondiente de la variable predictora, siempre es igual a cero

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

```
sum(mtcars$hp*modelo$residuals)
```

e) La suma de los residuales, ponderados por el valor estimado correspondiente siempre es igual a cero

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

```
sum(modelo$fitted.values*modelo$residuals)
```

4.3 Estimación de σ^2

Se requiere esta estimación para probar hipótesis y obtener intervalos de confianza.

El estimado de σ^2 se obtiene de la suma de cuadrados de residuales

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

La suma de cuadrados de residuales tiene n-2 grados de libertad, porque dos grados de libertad se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{y}_i$. Se puede demostrar que el valor esperado de SCRes es E(SCRes) = (n-2)σ^2 por lo que un estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = \frac{SCRes}{n-2} = CMRes$$

La cantidad anterior es llamada el Cuadrado Medio del Residual y su raíz cuadrada es denominada el error estándar de la regresión y tiene las mismas unidades que la variable respuesta y.

5. Prueba de Hipótesis de la pendiente

Con frecuencia interesa probar hipótesis y establecer intervalos de confianza de los parámetros del modelo. Estos procedimientos requieren hacer la hipótesis adicional de que los errores ε_i del modelo estén distribuidos normalmente. Así, las hipó tesis completas son: que los errores estén distribuidos en forma normal e independiente con media 0 y varianza σ^2.

5.1 Uso de la prueba t

Si se desea probar la hipótesis que la pendiente es igual a una constante β_{10}, las hipótesis correspondientes son:

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

$\hat{\beta}_1$ es una combinación lineal de las observaciones de modo que $\hat{\beta}_1$ está distribuido normalmente con media β_1 y varianza σ^2 / S_{xx} , por lo tanto el estadístico

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

Si se conociera σ^2 se podría usar Z para probar la hipótesis propuesta, pero comúnmente σ^2 es desconocido. Como CMRes es un estimador insesgado de σ^2 , por lo que

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{CMRes/S_{xx}}} \sim t_{(n-2)}$$

Este procedimiento rechaza H_0 si

$$|t_0| > t_{(1-\alpha/2,n-2)}$$

```
modelo <- lm(mpg ~ hp, data = mtcars)
car::linearHypothesis(modelo, "hp = -0.09")
```

Al denominador $\sqrt{CMRes/S_{xx}}$ del estadístico t_0 se le llama frecuentemente el error estándar de la pendiente $(se(\hat{\beta}_1))$.

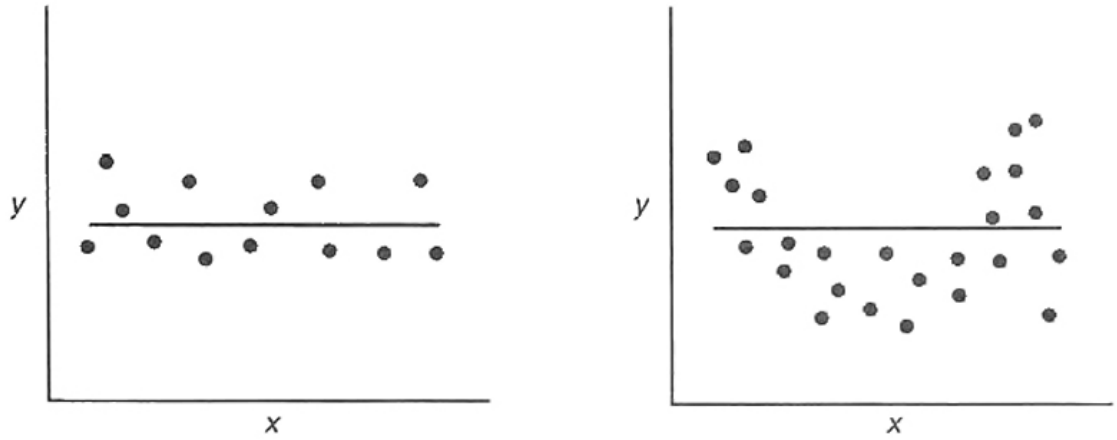
```
summary(modelo)
```

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


Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

5.2 Prueba de significancia de la regresión

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

El no rechazar H_0 implica que x no explica a y



Si usamos el estadístico propuesto anteriormente cuando $\beta_{10} = 0$

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{(n-2)}$$

H_0 se rechaza si

$$|t_0| > t_{(1-\alpha/2,n-2)}$$

5.3 Análisis de Varianza

Para probar que X explica a Y también se puede usar el método de Análisis de Varianza. Este análisis se basa en una partición de la variabilidad total de la variable *y.* Para obtener esta partición se comienza con la identidad

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

Se elevan al cuadrado ambos lados de la ecuación anterior y se suma para las n observaciones. Se tiene

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

El tercer termino del lado derecho es cero porque puede escribirse de la siguiente manera:

$$2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n}e_i = 0$$

Por lo que:

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

El lado izquierdo de la igualdad mide la suma corregida de cuadrados de las observaciones (SCT) es decir la variabilidad total en las observaciones. El primer componente del lado derecha mide la variabilidad en las observaciones y_i explicada por la línea de regresión (SCReg) y el segundo componente la variación residual que queda sin explicar por la línea de regresión (SCRes)

$$SCT = SCReg + SCRes$$

La suma de cuadrados de la regresión también puede ser calculada como:

$$SReg = \hat{\beta}_1 S_{xy} = \hat{\beta}_1{}^2 S_{xx}$$

La cantidad de grados de libertad queda determinada como sigue:

La SCT tiene n-1 grados de libertad porque se perdió un grado de libertad como resultado de la restricción $\sum_{i=1}^{n}(y_i - \bar{y})$ para las desviaciones $(y_i - \bar{y})$
La SCReg tiene 1 grado de libertad porque SCReg queda determinado por el estimador de un parámetro que es $\hat{\beta}_1$

La SCRes tiene n-2 grados de libertad porque se imponen dos restricciones a las desviaciones $(y_i - \hat{y}_i)$ como resultado de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$

$$GLT = GLReg + GLRes$$

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

Se puede aplicar la prueba F del análisis de varianza para probar la hipótesis H_0: $\beta_1=0$, de la siguiente manera:

$$F_0 = \frac{SCReg/GLReg}{SCRes/GLRes} = \frac{SCReg/1}{SCRes\ /\ (n-2)} = \frac{CMReg}{CMRes} \sim F_{(1,n-2)}$$

Cuadro ANOVA

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	SCReg=$\hat{\beta}_1 S_{xy}$	1	CMReg	CMReg/CMRes
Residual	SCRes = SCT - SCReg	n-2	CMRes	
Total	SCT	n-1		

Se rechaza H_0 si:

$$F_0 > F_{(1-\alpha,1,n-2)}$$

Por lo que el pvalor se calcula de la siguiente manera:

Pvalor = $P(F_{(1,n-2)} > F_0)$

t_0^2 es equivalente a F_0, esto debido a que el cuadro de una variable aleatoria t con f grados de libertad es una variable aleatoria F con uno y f grados de libertad.

```
n<-nrow(mtcars)
sct<-var(mtcars$mpg)*(n-1)
screg<-var(mtcars$hp)*(n-1)*modelo$coefficients[2]^2
sce<-sct-screg

anova(modelo)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
hp         1 678.37  678.37   45.46 1.788e-07 ***
Residuals 30 447.67   14.92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

#Cálculo del pvalor
pf(45.46,1,30,lower.tail=F)
[1] 1.787764e-07

#Valor crítico
qf(0.05,1,30,lower.tail=F)
[1] 4.170877
```

Aunque la prueba t para H_0: $\beta_1 = 0$ equivale a la prueba F en la regresión lineal simple, la prueba t es más adaptable, porque se podría usar para probar hipótesis alternativas unilaterales (H_1: $\beta_1 < 0$ o H_1: $\beta_1 > 0$), mientras que la prueba F solo considera la alterna bilateral.
La incapacidad de demostrar que $\beta_1 \neq 0$ no necesariamente quiere decir que *y* y x no están relacionadas. Puede indicar que la capacidad de detectar esta relación se ha confundido por la varianza del proceso de medición, o que el intervalo de valores de x es inadecuado. Se requiere una gran cantidad de evidencia no estadística y conocimiento del problema, para llegar a la conclusión que $\beta_1 = 0$.

5.4 Estimación de Intervalo en la Regresión Lineal Simple

Intervalos de confianza para β_1 y σ^2
El ancho de los intervalos es una medida de calidad general de la recta de regresión. Si los errores se distribuyen normalmente e independientemente

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{SE(\hat{\beta}_1)} \sim t_{(n-2)}$$

Por lo tanto, un intervalo del 100(1-α)% para β_1 se determina con:

$$\hat{\beta}_1 - t_{(1-\alpha/2,n-2)}SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{(1-\alpha/2,n-2)}SE(\hat{\beta}_1)$$

```
#Intervalo de confianza para los coeficientes
res<-summary(modelo)
LI<-modelo$coefficients[2]-qt(1-0.05/2,n-2)*res$coefficients[2,2]
LS<-modelo$coefficients[2]+qt(1-0.05/2,n-2)*res$coefficients[2,2]
c(LI,LS)

confint(modelo,level = 0.95)
                   2.5 %     97.5 %
(Intercept) 26.76194879 33.4357723
hp          -0.08889465 -0.0475619
```

Por otro lado,

$$\frac{(n-2)CMRes}{\sigma^2} \sim \chi^2_{(n-2)}$$

En consecuencia, un intervalo del 100(1-α)% para σ^2

$$\frac{(n-2)CMRes}{\chi^2_{(1-\alpha/2,n-2)}} \leq \sigma^2 \leq \frac{(n-2)CMRes}{\chi^2_{(\alpha/2,n-2)}}$$

A continuación se presenta una propuesta para estimar un intervalo de confianz a para σ^2

```
cisigma <- function(model, level){
  alfa <- 1-level; n <- length(model$residuals)
  LI <- sum(model$residuals^2)/qchisq(1-alfa/2, n-2)
  LS <- sum(model$residuals^2)/qchisq(alfa/2, n-2)
  data <- matrix(c(LI, LS), 1, 2)
  rownames(data) <- c("sigma^2")
  colnames(data) <- c(paste(alfa/2*100, "%"), paste(100-alfa/2*100, "%"))
  data <- as.table(data)
  data
}
modelo <- lm(mpg ~ hp, data = mtcars)
cisigma(modelo, level = 0.99)
```

6. Coeficiente de determinación

Es una medida de bondad de ajuste definido por:

$$R^2 = \frac{SCReg}{SCT} = 1 - \frac{SCRes}{SCT}$$

Como SCT es una medida de variabilidad de *y* sin considerar el efecto de la variable regresara x Y SCRes es una medida de la variabilidad de y que queda después de haber tenido en consideración a x, R^2 se llama, con frecuencia, la proporción de la variación explicada por el regresor x.
$0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y está explicada por el modelo de regresión.
Aunque R^2 sea grande, eso no necesariamente implica que el modelo de regresión sea un predictor exacto.

```
summary(modelo)$r.sq*100
[1] 60.24373
```

7. Estimación de intervalos de la respuesta media

Una aplicación importante de un modelo de regresión es estimar la respuesta media, E(y), para determinado valor de la variable regresora x.
Sea x_o el valor, o "nivel", de la variable regresora para el que se desea estimar la respuesta media, es decir, E(yl x_o) Se supone que x_o es cualquier valor de la variable regresora dentro del intervalo de los datos originales de x que se usaron para ajustar el modelo. Un estimador insesgado de E(yl x_o) se determina a partir del modelo ajustado como sigue:

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Para obtener un intervalo de confianza del 100(1-α)% para E(yl x_o) se debe notar primero que $\hat{\mu}_{y|x_0}$ es una variable aleatoria normalmente distribuida porque es una combinación lineal de las observaciones y_i. La varianza de $\hat{\mu}_{y|x_0}$ es:

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

Así la distribución de muestreo de

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{CME\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{(n-2)}$$

Por lo que un intervalo 100(1-α)% para la respuesta media en el punto x=x_0 es

$$\hat{\mu}_{y|x_0} - t_{(1-\alpha/2,n-2)}\sqrt{CME\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \leq E(y|x_0)$$

$$\leq \hat{\mu}_{y|x_0} + t_{(1-\alpha/2,n-2)}\sqrt{CME\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

```
summary(mtcars$hp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   52.0    96.5   123.0   146.7   180.0   335.0


#fit
sum(modelo$coefficients*c(1,330))
predict(modelo, data.frame(hp =330), level = 0.95,interval = "confiden
ce")
       fit      lwr      upr
1 7.583529 3.546574 11.62048
```

8. Predicción de nuevas observaciones

Una aplicación importante del modelo de regresión es predecir nuevas observaciones *y* que corresponda a un nivel especificado de la variable regresora x.

Si x_0 es el valor de interés de la variable regresora, entonces

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

Es el estimado puntual del nuevo valor de la respuesta y_0.

Si se define la variable

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

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

Por lo que un intervalo 100(1-α)% para una observación futura en el punto $x=x_0$ es:

$$\hat{y}_0 - t_{(1-\alpha/2,n-2)}\sqrt{CME\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \leq y_0$$

$$\leq \hat{y}_0 + t_{(1-\alpha/2,n-2)}\sqrt{CME\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

```
predict(modelo, data.frame(hp = 3.252), level = 0.95,
interval = "prediction")
```

9. Regresión por el origen

Algunos casos de regresión parecen implicar que una recta que pase por el origen debe ajustarse a los datos. El modelo sin ordenada al origen es

$$y = \beta_1 X + \varepsilon$$

Dada n pares de datos (y_1,x_1), (y_2,x_2),…,(y_n,x_n) la función de mínimos cuadrados es:

$$S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 X_i)^2$$

La única ecuación normal es:

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

Y el estimador de la pendiente por mínimos cuadrados es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

El estimador $\hat{\beta}_1$ es insesgado para β_1 y el modelo de regresión ajustado es:

$$\hat{y} = \hat{\beta}_1 X$$

El estimador de σ^2 es:

$$\hat{\sigma}^2 \equiv CMRes = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i}{n-1}$$

El intervalo de confianza de 100(1-α)% para β_1 es:

$$\hat{\beta}_1 - t_{1-\alpha/2,n-1}\sqrt{\frac{CMRes}{\sum_{i=1}^{n} x_i^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2,n-1}\sqrt{\frac{CMRes}{\sum_{i=1}^{n} x_i^2}}$$

Un intervalo 100(1-α)% para la respuesta media en el punto $x=x_0$ es

$$\hat{\mu}_{y|x_0} - t_{(1-\alpha/2,n-1)}\sqrt{\frac{CMRes}{\sum_{i=1}^{n} x_i^2}} \leq E(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{(1-\alpha/2,n-1)}\sqrt{\frac{CMRes}{\sum_{i=1}^{n} x_i^2}}$$

Un intervalo de predicción 100(1-α)% para una observación futura en $x=x_0$ es

$$\hat{y}_0 - t_{(1-\alpha/2,n-1)}\sqrt{CMRes\left(1+\frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)} \leq y_0 \leq \hat{y}_0 + t_{(1-\alpha/2,n-1)}\sqrt{CMRes\left(1+\frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}$$

El análogo de R^2 en un modelo que pasa por el origen es:

$$R_0^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2}$$

```
#Estimación del modelo
modelo0<-lm(mpg~0+hp,data=mtcars)
modelo0

Call:
lm(formula = mpg ~ 0 + hp, data = mtcars)

Coefficients:
    hp
0.1011

#Análisis de Varianza

anova(modelo0)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
hp         1 8530.8  8530.8  47.982 9.062e-08 ***
Residuals 31 5511.5   177.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(modelo0)

Call:
lm(formula = mpg ~ 0 + hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-18.875  -2.102   6.062  13.244  27.327

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
hp   0.1011     0.0146   6.927 9.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.33 on 31 degrees of freedom
Multiple R-squared:  0.6075,   Adjusted R-squared:  0.5948
F-statistic: 47.98 on 1 and 31 DF,  p-value: 9.062e-08

#Intervalo de confianza para la pendiente
confint(modelo0,level = 0.95)
        2.5 %    97.5 %
```

```
hp 0.07134742 0.1308938

summary(modelo0)$r.sq
[1] 0.6075074
#Verificación de r2
sum(modelo0$fitted.values^2)/sum(mtcars$mpg^2)
[1] 0.6075074
#Prediciones
predict(modelo0, data.frame(hp =100), level = 0.95, interval = "confid
ence")
       fit      lwr      upr
1 10.11206 7.134742 13.08938
predict(modelo0, data.frame(hp =100), level = 0.95, interval = "predic
tion")
       fit       lwr      upr
1 10.11206 -17.24491 37.46904
```