Programación Estadística con R

UPCH

2019

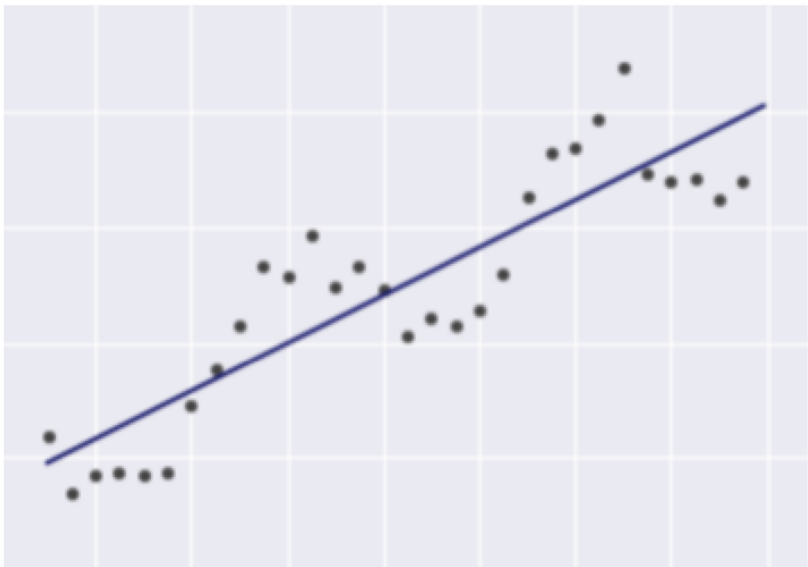
Modulo III

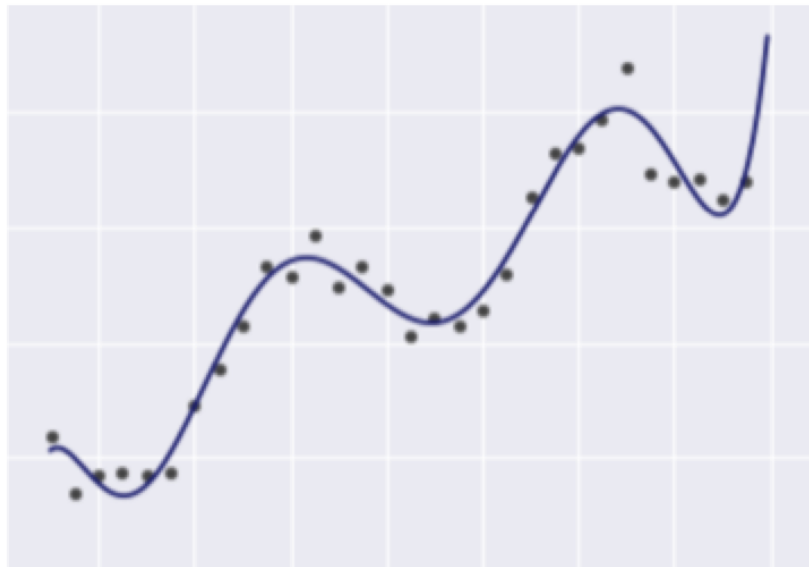
CONSTRUCCIÓN DE MODELOS Y MÉTODOS DE REGRESIÓN

UNIVERSIDAD PERUANA
CAYETANO HEREDIA
ESCUELA DE POSGRADO

Análisis de Regresión

El análisis de regresión es un poderoso método estadístico que le permite examinar la relación entre dos o más variables de interés.





Si bien existen muchos tipos de análisis de regresión, en su núcleo todos examinan la influencia de una o más variables independientes en una variable dependiente.

Análisis de Regresión

Un primer ejemplo

Supongamos que es un gerente de ventas que intenta predecir los números del próximo mes. Usted sabe que docenas, tal vez incluso cientos de factores, desde el clima hasta la promoción de un competidor y el rumor de un modelo nuevo y mejorado, pueden afectar el número. Tal vez las personas en su organización incluso tengan una teoría sobre lo que tendrá el mayor efecto en las ventas. Y escuchas cosas como **Créeme. Cuanta más lluvia tengamos, más vendemos**.Seis semanas después de la promoción de la competencia, las ventas suben ".

Análisis de Regresión

Un primer ejemplo : Razonamiento matemático

El análisis de regresión es una forma de clasificar matemáticamente cuál de esas variables tiene un impacto. Responde las preguntas:

- ¿Qué factores son más importantes?
- ¿Qué podemos ignorar?
- ¿Cómo interactúan esos factores entre sí?

Y, quizás lo más importante,

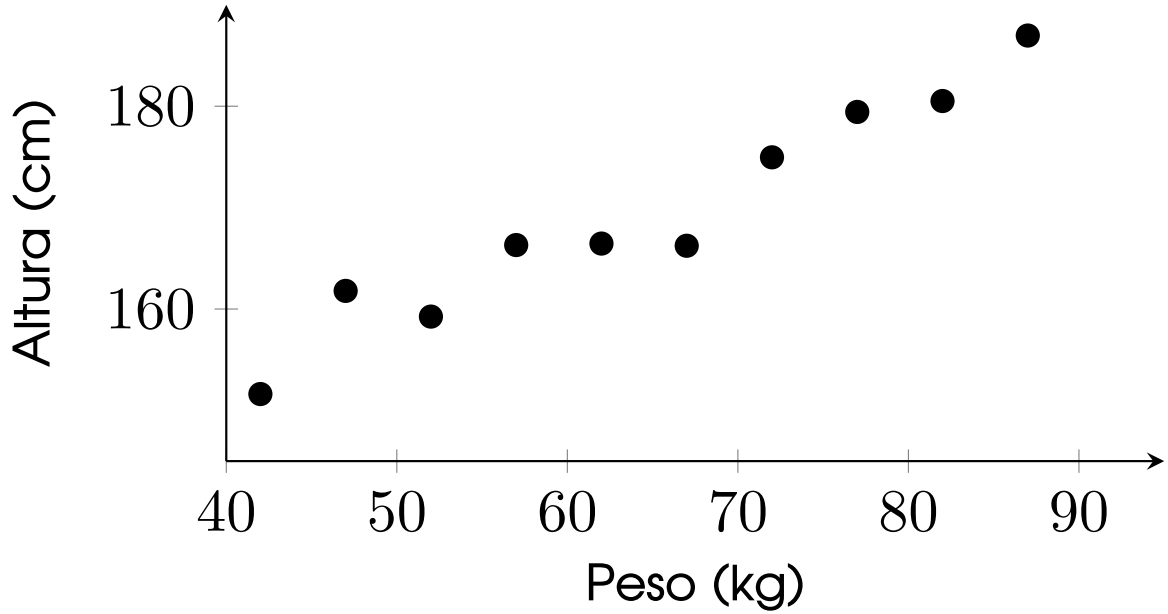
- ¿qué tan seguros estamos de todos estos factores?

Análisis de Regresión

Ejemplo 2 : Peso y Altura

x	y
42	151.62
47	161.79
52	159.25
57	166.31
62	166.46
67	166.24
72	174.96
77	179.44
82	180.51
87	186.97

Consideremos la variable *Peso* en el eje X, es decir esta es la variable independiente. Y luego consideremos la variable *Altura* en el eje Y, es decir, esta es la variable dependiente.



Lo que buscamos es construir un modelo lineal que relacione estas dos variables.

Análisis de Regresión

Ejemplo 2 : Peso y Altura

Para el análisis de una situación de relación entre dos variables se debe:

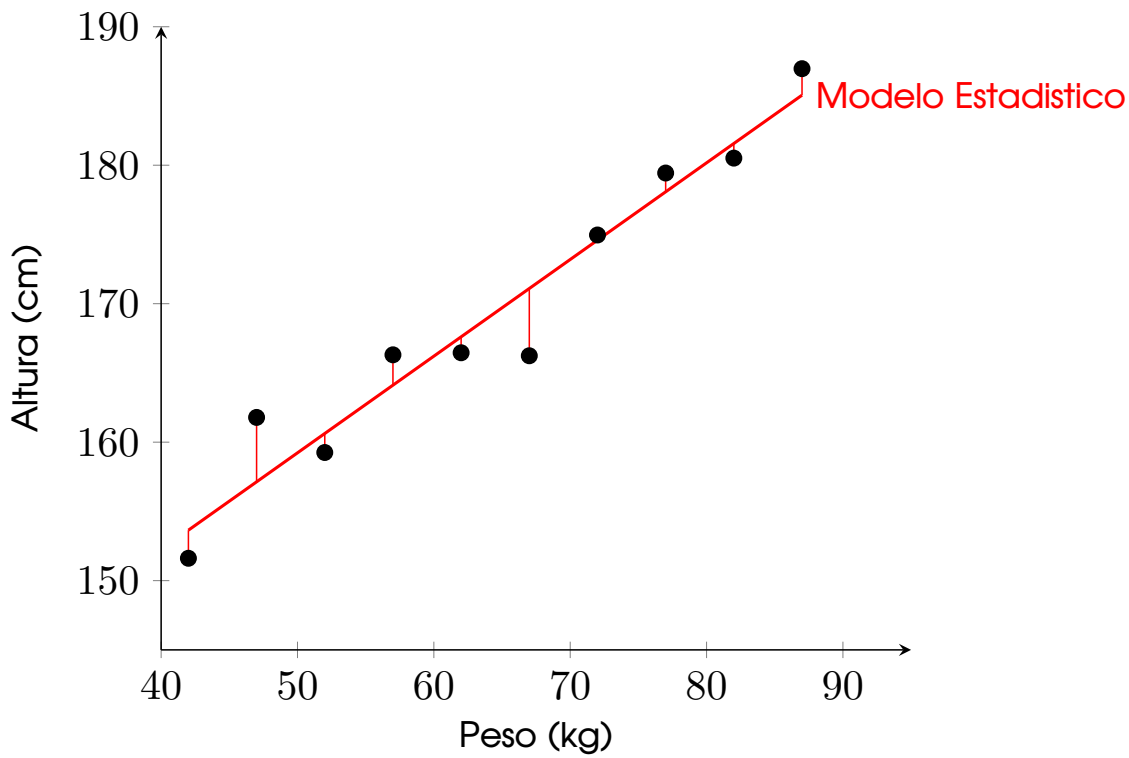
1. Identificar la variable independendiente y la variable dependiente: En este caso la variable dependiente es la **Altura** y la variable independiente es **Peso.**
2. Determinar si existe una relación de dependencia razonable. En la situación presentada puede observarse que en la realidad estas dos características (Pero y Altura) presentan una relación lógica.

Para determinar de manera inicial la relación lineal entre las dos variables se debe elaborar un *diagrama de dispersión,* como el que aparece en la figura del slide anterior.

De acuerdo al gráfico de dispersión se puede asumir que existe una relación lineal y se requiere la línea recta que mejor se ajuste a los datos experimentales.

3. Determinar el modelo estadístico : Como la Altura parece aumentar a medida que aumenta el Peso entonces se debe sugerir un modelo lineal dado por:

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, ..., 10$$



Análisis de Regresión

Ejemplo 2 : Peso y Altura

En el modelo :

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, ..., 10$$

Donde y_i es el valor observado en este caso la Altura para un valor del Peso x_i, β_o corresponde al intercepto del eje Y con la línea de regresión y β_1 representa la pendiente de la línea de regresión o coeficiente de regresión, y ε_i es la variable aleatoria de error.
Para poder utilizar este modelo , se asume que las variables error ε_i cumplen los siguientes supuestos:

- Normales con media cero
- Independientes
- Con igual varianza σ^2.

Estos supuesto deben validarse.

4. Determinar la ecuación de regresión o modelo ajustado: El modelo predicho o ecuación de regresión ajustada es una expresión como la siguiente

$$\widehat{y_i} = \widehat{\beta}_o + \widehat{\beta}_1 x_i$$

Para obtenerla usted debe encontrar los valores estimados de los parámetros: $\widehat{\beta}_o$ y $\widehat{\beta}_1$. Éstos se obtienen aplicando el **método de mínimos cuadrados.**

Análisis de Regresión

El método de mínimos cuadrados

El método de mínimos cuadrado trata de buscar cual es la recta que más se acerca a los puntos; es decir busca la recta que haga que la distancia entre el valor real y_i y el valor obtenido por la recta ajustada $\widehat{y_i}$ sea la más pequeña y así, la suma de todas estas distancias simbolizadas como:

$$\text{Suma de cuadrados del error} = SCE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Sea la más pequeña.Como la mejor recta está determinada por $\widehat{\beta}_o$ y $\widehat{\beta}_1$ entonces matemáticamente, se desea escoger los valores para $\widehat{\beta}_o$ y $\widehat{\beta}_1$ que minimicen la suma de cuadrados del error.

INICIO: ABRIL 2018

CURSO VIRTUAL

PROGRAMACIÓN ESTADÍSTICA CON R

UNIVERSIDAD PERUANA
CAYETANO HEREDIA
ESCUELA DE POSGRADO

Programación Estadística con R

UPCH

Septiembre 2018

Unidad 3.

Construcción de Modelos y Métodos de Regresión (I)

1. Regresión Lineal Simple

Regresión

Regresión Simple

El objetivo del Análisis de regresión es determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables. En el Análisis de regresión simple, se pretende estudiar y explicar el comportamiento de una variable que notamos y, y que llamaremos variable explicada, variable dependiente o variable de interés, a partir de otra variable, que notamos x, y que llamamos variable explicativa, variable independiente o variable de predicción. El principal objetivo de la regresión es encontrar la función que mejor explique la relación entre la variable dependiente y las independientes.

Regresión

Regresión Simple

A menudo se supone que la relación que guardan la variable dependiente y las independientes es lineal. En estos casos, se utilizan los modelos de regresión lineal. Aunque las relaciones lineales aparecen de forma frecuente, también es posible considerar otro tipo de relación entre las variables, que se modelizan mediante otros modelos de regresión, como pueden ser el modelo de regresión cuadrático o parabólico o el modelo de regresión hiperbólico.

Teoría de la Regresión

Consiste en la búsqueda de una **función** que exprese lo mejor posible el tipo de relación entre dos o más variables.

Correlación

Regresión Simple

La correlación está íntimamente ligada con la regresión en el sentido de que se centra en el estudio del grado de asociación entre variables. Por lo tanto, una variable independiente que presente un alto grado de correlación con una variable dependiente será muy útil para predecir los valores de ésta última. Cuando la relación entre las variables es lineal, se habla de correlación lineal. Una de las medidas más utilizadas para medir la correlación lineal entre variables es el coeficiente de correlación lineal de Pearson.

Teoría de la Correlación

Estudia el grado de dependencia entre las variables, es decir su objetivo es medir el grado de ajuste existente entre la función teórica (función ajustada) y la nube de puntos.

Regresión lineal simple

Modelo Lineal

La regresión lineal simple supone que los valores de la variable dependiente, a los que llamaremos y_i, pueden escribirse en función de los valores de una única variable independiente, los cuales notaremos por x_i, según el siguiente modelo lineal:

$$y_i = \beta_0 + \beta_1 x_i \tag{1}$$

donde β_0 y β_1 son los **parámetros** desconocidos que vamos a **estimar**

UNIVERSIDAD PERUANA
CAYETANO HEREDIA
ESCUELA DE POSGRADO

Regresión lineal simple

Habitualmente, al iniciar un estudio de regresión lineal simple se suelen representar los valores de la variable dependiente y de la variable independiente de forma conjunta mediante un diagrama de dispersión para determinar si realmente existe una relación lineal entre ambas. Para realizar un diagrama de dispersión en R y RStudio utilizaremos la orden plot

Función plot

```
1 plot(x,y)
```

donde *x* e *y* son los valores de las variables independiente y dependiente, respectivamente.

UNIVERSIDAD PERUANA CAYETANO HEREDIA
ESCUELA DE POSGRADO

Regresión lineal simple

Después de comprobar gráficamente la relación lineal entre las variables, el siguiente paso es la estimación de los valores de los parámetros β_0 y β_1 que aparecen en la fórmula (1) a partir de un conjunto de datos. Para ello, podemos utilizar la función lm de R, cuya sintaxis es la siguiente

```
lm(formula , data)
```

donde `formula` indica la relación que guardan la variable dependiente y la variable independiente.

Regresión lineal simple

Por ejemplo:

```
1 lm(formula = y ~ x, data=midataset)
```

Algunas peculiaridades sobre este argumento:

- Las partes izquierda (variable dependiente) y derecha (variable independiente) de la fórmula vienen separadas por el símbolo ~ , que puede escribirse con la secuencia de comandos Alt + 126.
- `data` es el conjunto de datos en el que se encuentran las variables que se utilizan en la fórmula.

Regresión lineal simple

Ejemplo 1

En primer lugar, vamos a almacenar los datos de las dos variables en dos vectores.

```
edad <- c(56, 42, 72,  36, 63, 47, 55, 47, 38, 42)
presion <- c(148, 126, 159, 118, 149, 130, 151, 142, 114, 141)
edad
presion
```

Regresión lineal simple

Ejemplo 1

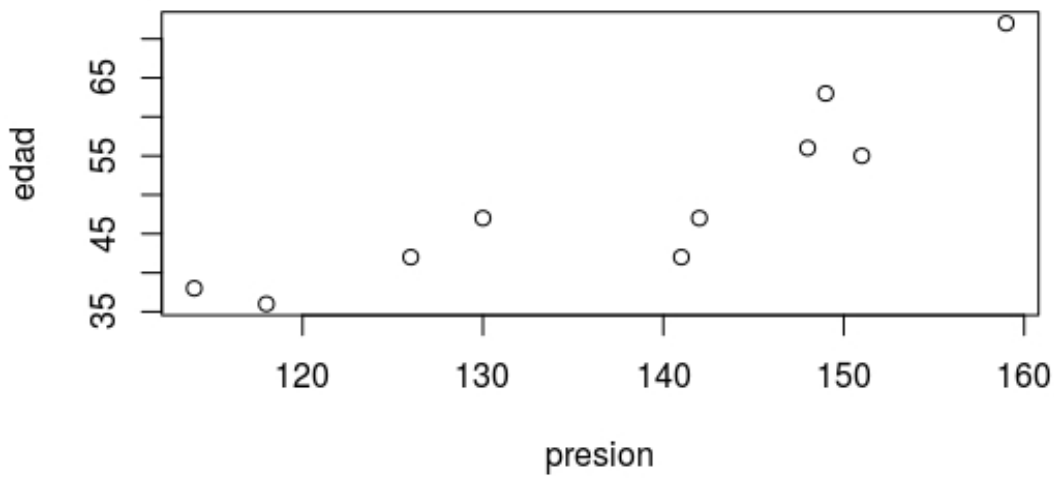
Supongamos que nuestro objetivo es determinar la edad de una mujer a partir de su presión sanguínea o, lo que es lo mismo, supongamos que la variable dependiente es edad y que la variable independiente es presión. Vamos a representar el diagrama de dispersión de las dos variables para determinar si la relación existente entre ambas puede considerarse lineal, y por tanto, si tiene sentido plantear un modelo de regresión lineal simple.

```
1 plot(presion , edad)
```

Regresión lineal simple

Ejemplo 1

```
1 plot(presion , edad)
```



Regresión lineal simple

Ejemplo 1

A la vista del gráfico de dispersión, se puede asumir un cierto grado de relación lineal entre ambas variables, por lo que procedemos al ajuste del modelo lineal.

```
1 reg_lin <- lm(edad ~ presion)
2 reg_lin
```

```
Call:
lm(formula = edad ~ presion)

Coefficients:
(Intercept)      presion
-43.5440          0.6774
```

Regresión lineal simple

Ejemplo 1

Por defecto, la salida que muestra la función lm incluye únicamente las estimaciones para los parámetros, en nuestro caso β_0 y β_1 Por tanto, el modelo lineal puede escribirse del siguiente modo:

$$edad_i = -43.5440 + 0.6774 presion_i$$

Estos dos parámetros pueden interpretarse del siguiente modo:
-43.5440 es el valor de la edad para una persona de presión sanguínea 0, lo cual no tiene sentido. De hecho, en multitud de ocasiones la interpretación del parámetro β_0 no es relevante y todo el interés recae sobre la interpretación del resto de parámetros.

Regresión lineal simple

Ejemplo 1

El parámetro β_1 es igual a 0.6774 indica que, por término medio, cada mmHg (milímetro de mercurio) de incremento en la presión sanguínea de una persona supone un incremento de 0.6774. en su edad.
Podemos obtener más información sobre el modelo de regresión que hemos calculado aplicando la función **summary** al objeto que contiene los datos de la regresión, al cual hemos llamado `reg_lin` en este ejemplo.

```
1 summary(reg_lin)
```

Regresión lineal simple

Verificación Visual

Rstudio permite que dibujemos la recta de regresión lineal sobre el diagrama de dispersión mediante la orden **abline**. De este modo podemos visualizar la distancia existente entre los valores observados y los valores que el modelo pronostica (esto es, los residuos).

```
1 plot(presion , edad)
2 abline(reg_lin)
```

Programación Estadística con R

UPCH

Septiembre 2018

UNIVERSIDAD PERUANA
CAYETANO HEREDIA
Escuela de Posgrado

Programación en R

1. Regresión Lineal
2. Regresión Logística
3. Regresión Polinomial
4. Regresión de Poisson

Regresion Lineal

Ejemplo - Unidad 3 parte 1

Agreguemos ciertas características visuales a nuestro código original :

```
edad <- c(56, 42, 72, 36, 63, 47, 55, 47, 38, 42)
presion <- c(148, 126, 159, 118, 149, 130, 151, 142, 114, 141)
plot(presion, edad)
reg_lin <- lm(edad ~ presion)
summary(reg_lin)
xmin <- 0.9*min(presion)
xmax <- 1.1*max(presion)
ymin <- 0.9*min(edad)
ymax <- 1.1*max(edad)
plot(presion, edad, main="Edad ~ Presion Sanguinea", sub="POb. : 10
    mujeres",xlab="Presion", ylab="Edad",xlim=c(xmin, xmax), ylim=c(
    ymin, ymax))
abline(reg_lin)
```

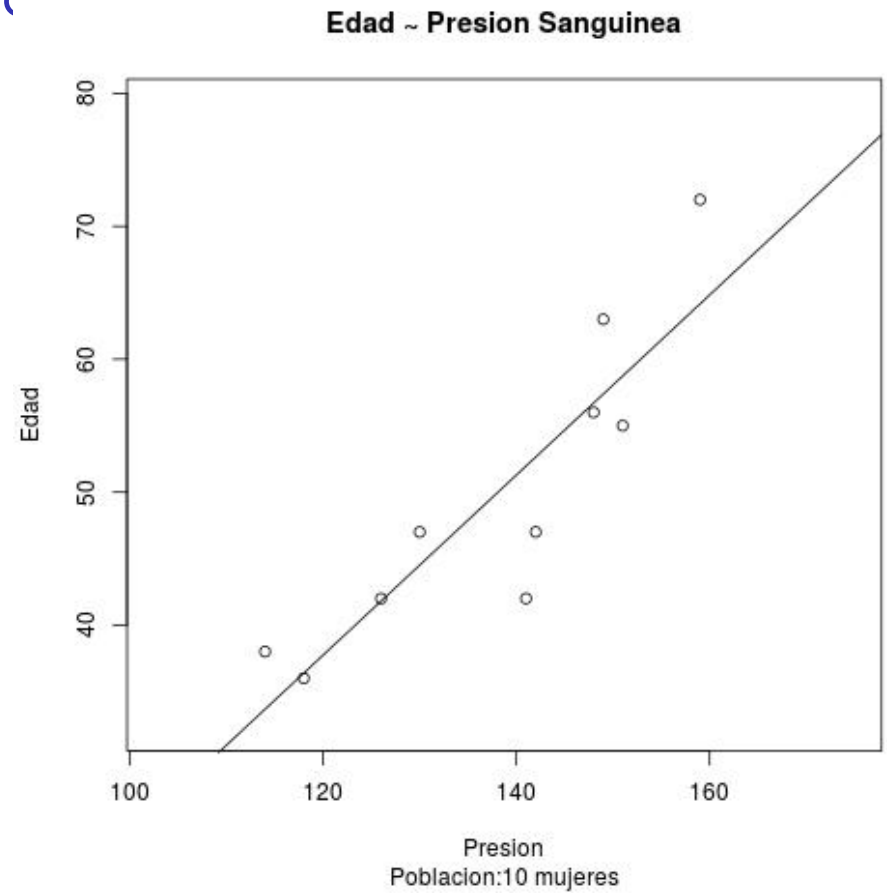
Regresion Lineal

Ejemplo - Unidad 3 Parte 1 ++

```
edad <- c(56, 42, 72,  36, 63, 47, 55, 47, 38, 42)
presion <- c(148, 126, 159, 118, 149, 130, 151, 142, 114, 141)
plot(presion, edad)
reg_lin <- lm(edad ~ presion)
summary(reg_lin)
xmin <- 0.9*min(presion)
xmax <- 1.1*max(presion)
ymin <- 0.9*min(edad)
ymax <- 1.1*max(edad)
png(filename = "PobMujeres10.png")
plot(presion, edad, main="Edad ~ Presion Sanguinea", sub="Poblacion:10
     mujeres",xlab="Presion", ylab="Edad",xlim=c(xmin, xmax), ylim=c(
     ymin, ymax))
abline(reg_lin)
dev.off()
```

Regresion Lineal

Ejemplo - Unid



Regresion Lineal

Ejemplo - Unidad 3 parte 1 ++

```
> reg_lin

Call:
lm(formula = edad ~ presion)

Coefficients:
(Intercept)      presion
   -43.5440       0.6774
```

Por defecto, la salida que muestra la función **lm** incluye únicamente las estimaciones para los parámetros, en nuestro caso β_0 y β_1 . Por tanto, el modelo lineal puede escribirse del siguiente modo:

$$edad_i = -43{,}5440 + 0{,}6774 presion_i$$

UNIVERSIDAD PERUANA CAYETANO HEREDIA
Escuela de Posgrado

Ejemplo - Unidad 3 parte 1 ++

- Estos dos parámetros pueden interpretarse del siguiente modo: -43.5440 es el valor de la edad para una persona de presión sanguínea 0, lo cual no tiene sentido. De hecho, en multitud de ocasiones la interpretación del parámetro β_0 no es relevante y todo el interés recae sobre la interpretación del resto de parámetros.
- El parámetro β_1 es igual a 0.6774 indica que, por término medio, cada mmHg (milímetro de mercurio) de incremento en la presión sanguínea de una persona supone un incremento de 0.6774. en su edad.

Regresion Lineal

Podemos obtener más información sobre el modelo de regresión que hemos calculado aplicando la función **summary** al objeto que contiene los datos de la regresión, al cual hemos llamado `reg_lin` en este ejemplo.

Ejemplo - Unidad 3 parte 1 ++

```
1 summary(reg_lin)
```

Regresion Lineal

Ejemplo - Unidad 3 parte 1 ++

```
summary(reg_lin)
Call:
lm(formula = edad ~ presion)

Residuals:
Min      1Q  Median      3Q     Max
-9.9676 -2.9835 -0.0973  3.8623  7.8394

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.5440    17.6126  -2.472 0.038571 *
presion       0.6774     0.1271   5.328 0.000704 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 5.742 on 8 degrees of freedom
Multiple R-squared:  0.7802,    Adjusted R-squared:  0.7527
F-statistic: 28.39 on 1 and 8 DF,  p-value: 0.000704
```

Regresion Lineal : summary(reg_lin)

Funcion **summary**

Esta salida contiene una información más completa sobre el análisis. Así, por ejemplo, encontramos información sobre los residuos (en el apartado **Residuals**), que se definen como la diferencia entre el verdadero valor de la variable dependiente y el valor que pronostica el modelo de regresión. Cuanto más pequeños sean estos residuos mejor será el ajuste del modelo a los datos y más acertadas serán las predicciones que se realicen a partir de dicho modelo. En la tabla **Coefficients** encontramos los

```
Call:
lm(formula = edad ~ presion)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9676 -2.9835 -0.0973  3.8623  7.8394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.5440    17.6126  -2.472 0.038571 *
presion       0.6774     0.1271   5.328 0.000704 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.742 on 8 degrees of freedom
Multiple R-squared:  0.7802,    Adjusted R-squared:  0.7527
F-statistic: 28.39 on 1 and 8 DF,  p-value: 0.000704
```

Regresion Lineal

Es decir, para resolver los siguientes **contrastes de hipótesis**:

$$H_0 \equiv \beta_0 = 0 \qquad \text{vs} \qquad H_1 \equiv \beta_0 \neq 0$$

$$H_0 \equiv \beta_1 = 0 \qquad \text{vs} \qquad H_1 \equiv \beta_1 \neq 0$$

Lo que se pretende mediante estos contrastes, es determinar si los efectos de la constante y de la variable independiente son realmente importantes para explicar la variable dependiente o si, por el contario, pueden considerarse nulos.

Regresion Lineal

En nuestro ejemplo, los p-valores que nos ayudan a resolver estos contrastes son 0.038571 y 0.000704 ambos menores que 0.05. Así, considerando un nivel del significación del 5 %, rechazamos la hipótesis nula en ambos contrastes, de manera que podemos suponer ambos parámetros significativamente distintos de 0.

Por último, en la parte final de la salida, encontramos el valor de R^2 (Multiple R-squared) y de R^2 ajustado (Adjusted R-squared), que son indicadores de la bondad del ajuste de nuestro modelo a los datos. R^2 oscila entre 0 y 1, de manera que, valores de R^2 próximos a 1 indican un buen ajuste del modelo lineal a los datos. Por otro lado, R^2 ajustado es similar a R^2, pero penaliza la introducción en el modelo de variables independientes poco relevantes a la hora de explicar la variable dependiente. Por tanto, R^2 ajustado $\leq$ R^2. En nuestro ejemplo, R^2 = 0.7802

y R^2 ajustado = 0.7527, por lo que podemos concluir que el modelo lineal se ajusta de forma aceptable a nuestros datos.

Regresion Lineal

La última línea de la salida incluye un estadístico F de **Snedecor** y el p-valor correspondiente que se utilizan para resolver el siguiente contraste:

$$H_0 \equiv \beta_i \qquad \forall i$$
$$H_1 \equiv \beta_i \neq 0 \qquad \text{para al menos un i}$$

que se conoce habitualmente como contraste **ómnibus**. Mediante este contraste se comprueba si, de forma global, el modelo lineal es apropiado para modelizar los datos. En nuestro ejemplo, el p-valor asociado a este contraste es inferior a 0.05 por lo que, al 5 % de significación podemos rechazar la hipótesis nula y afirmar que, efectivamente, el modelo lineal es adecuado para nuestro conjunto de datos.

Regresion Lineal

Un aspecto importante cuando se trabaja con modelos de regresión lineal es la comprobación de las hipótesis que deben de cumplirse para poder utilizar este tipo de modelos. Estas suposiciones hacen referencia a los residuos y pueden resumirse en los siguientes puntos:

- Normalidad de los residuos
- Independencia de los residuos
- Homocedasticidad (igualdad de las varianzas de los residuos)
- Linealidad de los residuos

Regresion Lineal

Al aplicar la función **plot** sobre el objeto que contiene la información del modelo obtenemos 4 gráficos que nos ayudan para la validación del modelo. Estos gráficos son:

- Valores predichos frente a residuos
- Gráfico Q-Q de normalidad
- Valores predichos frente a raíz cuadrada de los residuos estandarizados (en valor absoluto)
- Residuos estandarizados frente a leverages

```
1 plot(reg_lin)
```

Regresion Lineal :summary(reg_lin)

plot(reg_lin)

El gráfico de residuos estandarizados frente a leverages se utiliza para detectar puntos con una influencia importante en el cálculo de las estimaciones de los parámetros. En caso de detectarse algún punto fuera de los límites que establecen las líneas discontinuas debe estudiarse este punto de forma aislada para detectar, por ejemplo, si la elevada importancia de esa observación se debe a un error.

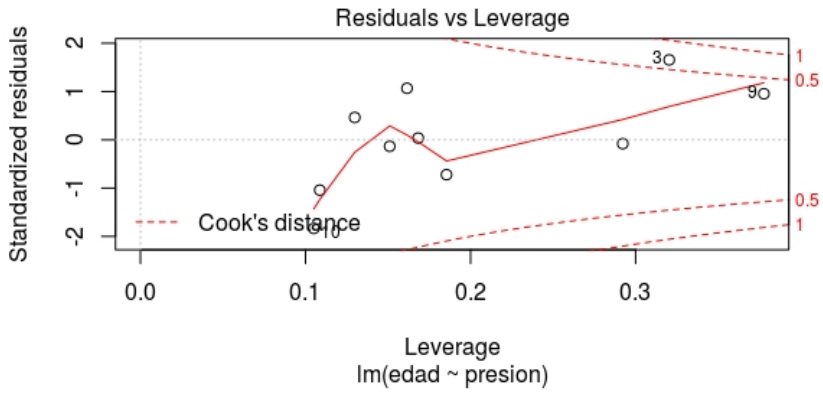


Figura: Gráfico:Validación de Modelo

Regresion Lineal : summary(reg_lin)

plot(reg_lin)

Los gráficos que vemos a continuación se utilizan para contrastar gráficamente la independencia, la homocedasticidad y la linealidad de los residuos. Idealmente, los residuos deben estar aleatoriamente distribuidos a lo largo del gráfico, sin formar ningún tipo de patrón.

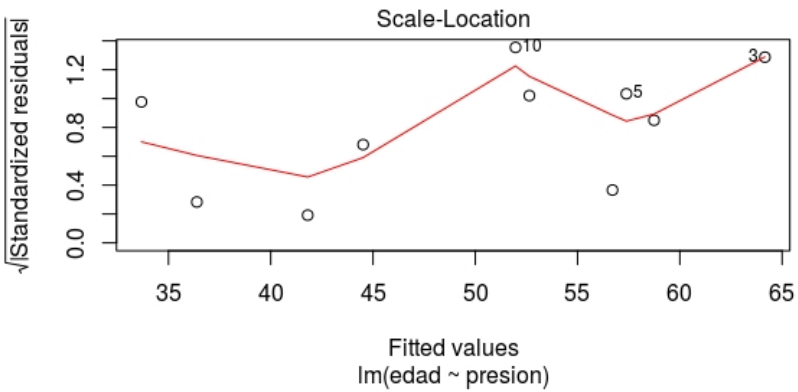


Figura: Gráfico:Validación de Modelo

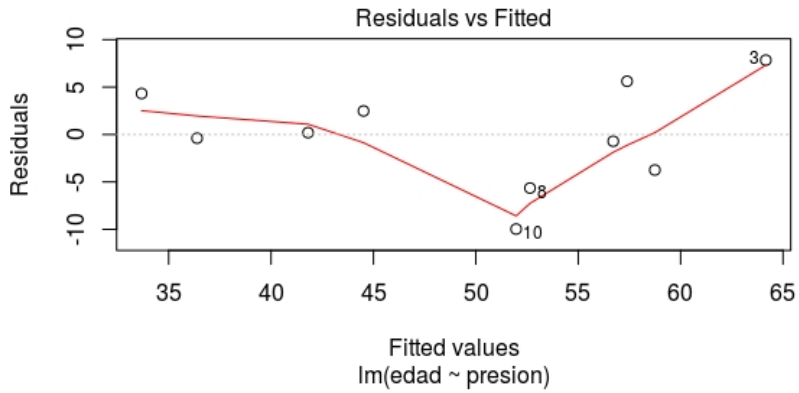


Figura: Gráfico:Validación de Modelo

Regresion Lineal: summary(reg_lin)

plot(reg_lin)

El gráfico Q- Q, por su parte, se utiliza para contrastar la normalidad de los residuos. Lo deseable es que los residuos estandarizados estén lo más cerca posible a la línea punteada que aparece en el gráfico.

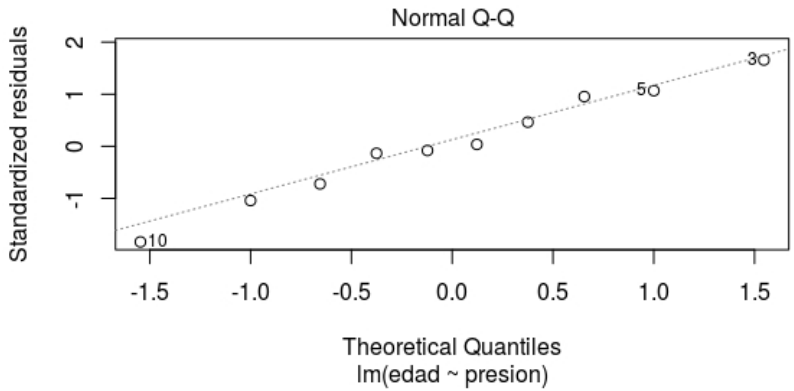


Figura: Gráfico:Validación de Modelo

Regresion Lineal

Los gráficos parecen indicar que los residuos son aleatorios, independientes y homocedásticos. Sin embargo, no parece que los residuos sigan una distribución Normal. Vamos a confirmar si esto es así mediante métodos analíticos.

Para comprobar la normalidad, aplicaremos a los residuos el test de normalidad de Kolmogorov-Smirnov, que en R y Rstudio se calcula a través de la función ks.test, cuya sintaxis es la siguiente:

ks.test (x, distrib)

donde:

- **x** es un vector numérico con los datos a los que vamos a aplicar el test (en nuestro caso, los residuos)
- **distrib** indica la distribución de referencia que se usará en el contraste (en nuestro caso, la distribución Normal, por lo que **distrib = pnorm**)

Regresion Lineal

test de normalidad de Kolmogorov-Smirnov

Al realizar un análisis de regresión lineal, R y RStudio guardan automáticamente los residuos en el objeto que almacena la información de la regresión (y que nosotros hemos llamado reg_lin). Para acceder a estos residuos, escribiremos `$residuals` a continuación del nombre del objeto que contiene la información del análisis. Por tanto, podemos realizar el contraste de Kolmogorov-Smirnov del siguiente modo:

```
ks.test(reg_lin$residuals, "pnorm")

One-sample Kolmogorov-Smirnov test

data: reg_lin$residuals
D = 0.3935, p-value = 0.06608
alternative hypothesis: two-sided
```

Regresion Lineal

test de normalidad de Kolmogorov-Smirnov

Los resultados del test nos confirman lo que se intuía en el gráfico Q-Q: a un 10 % de significación los residuos no siguen una distribución normal, puesto que el p-valor que se obtiene (0.06608) es menor que 0.1. Sin embargo para una significación del 5 % no se debe rechazar la hipótesis nula.

Test de Durbin-Watson

Por último, contrastemos la independencia de los residuos mediante el **test de Durbin-Watson**. La función que calcula este test se llama **dwtest** y se encuentra dentro del paquete **lmtest**. Por lo que lo primero que tenemos que hacer es instalar y cargar dicho paquete.

```
1 library(lmtest)
2 dwtest(edad~presion)
```

Regresion Lineal

Test de Durbin-Watson

```
1 library(lmtest)
2 dwtest(edad~presion)
```

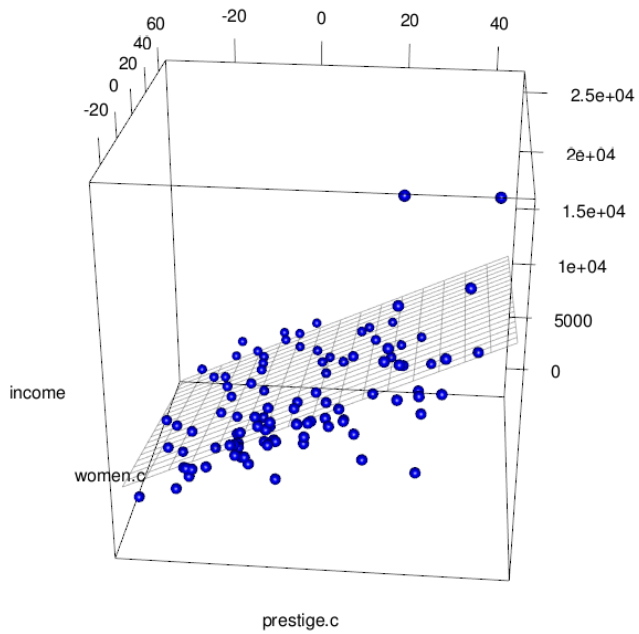
```
Durbin-Watson test

data: edad ~ presion
DW = 1.9667, p-value = 0.5879
alternative hypothesis: true autocorrelation
is greater than 0
```

En este caso, con un p-valor de 0.5879 no podemos rechazar la hipótesis de que los residuos son independientes.

Regresión lineal múltiple

El modelo de regresión múltiple es la extensión a k variables explicativas del modelo de regresión simple. En general, una variable de interés y dependiente de varias variables independientes $x_1, x_2, \ldots, x_k$, y no sólo de una única variable de



Regresión lineal múltiple

Por ejemplo, para estudiar la contaminación atmosférica, parece razonable considerar más de una variable explicativa, como pueden ser la temperatura media anual, el número de fábricas, el número de habitantes, etc. Además de las variables observables, la variable de interés puede depender de otras desconocidas para el investigador. Un modelo de regresión representa el efecto de estas variables en lo que se conoce como error aleatorio o perturbación.

Un modelo de regresión teórico en el que las variables se pueden relacionar mediante una función de tipo lineal, podemos expresarlo de la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{1}$$

Regresión lineal múltiple

Modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{2}$$

donde :

- y es la variable de interés que vamos a predecir también llamada **variable respuesta** o variable dependiente.
- $x_1, x_2, \cdots, x_k$ son variables independientes, explicativas o de predicción.
- $\beta_1, \beta_2, \cdots, \beta_k$ son los parámetros desconocidos que vamos a estimar.
- ϵ es el error aleatorio o perturbación, que representa el efecto de todas las variables que pueden afectar a la variable dependiente y no están incluidas en el modelo de regresión.

Regresión lineal múltiple

En R y Rstudio, el paso de un modelo de regresión lineal simple a un modelo de regresión lineal múltiple es muy sencillo: basta con añadir variables independientes al argumento formula de la función **lm** separadas por el signo +. Veamos un ejemplo:

Ejemplo 2

Se desea conocer el gasto de alimentación mensual de una familia en función del ingreso mensual, el tamaño de la familia y el número de hijos en la universidad. Vamos a ajustar un nuevo modelo de regresión lineal múltiple que explique el gasto de alimentación mensual en función de las variables descritas anteriormente. En primer lugar, vamos a crear cuatro vectores numéricos, uno para cada variable .

```
gastos <- c(1000, 580, 520, 500, 600, 550, 400)
ingresos <- c(50000, 2500, 2000, 1900, 3000, 4000, 2000)
tamano <- c(7, 4, 3, 3, 6, 5, 2)
hijosU <- c(3,1,1,0,1,2,0)
```

Regresión lineal múltiple

Ejemplo 2

Y vamos a agrupar la información relativa a las 4 variables de las que disponemos en un data frame al que pondremos por nombre datos2:

```
datos2 <- data.frame(gastos, ingresos, tamaño, hijosU)
```

A continuación ajustamos el modelo de regresión lineal múltiple

```
reg_lin_mul <- lm(gastos ~ ingresos + tamaño + hijosU)
summary(reg_lin_mul)
```

Regresión lineal múltiple

Ejemplo 2 : **summary(reg_lin_mul)**

```
Call:
lm(formula = gastos ~ ingresos + tamano + hijosU)
Residuals:
      1       2       3       4       5       6       7
  1.216  48.164  29.125  15.209 -10.134 -35.402 -48.178
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.590e+02  6.291e+01   5.706   0.0107 *
ingresos    7.247e-03  1.802e-03   4.021   0.0276 *
tamano       3.734e+01  2.046e+01   1.825   0.1655
hijosU      5.359e+00  4.061e+01   0.132   0.9034
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 48.57 on 3 degrees of freedom
Multiple R-squared:  0.9677,    Adjusted R-squared:  0.9353
F-statistic: 29.93 on 3 and 3 DF,  p-value: 0.009772
```

Regresión lineal múltiple

Ejemplo 2 : **summary(reg_lin_mul)**

En `Coefficients` se muestran los parámetros estimados de regresión : $\beta_0 = 3{,}590e+02$, $\beta_1 = 7{,}247e-03$, $\beta_2 = 3{,}734e+01$ y $\beta_3 = 5{,}359e+00$.
La ecuación de regresión estará ajustada por

$$gastos_i = (3{,}590e{+}02){+}(7{,}247e{-}03)ingresos_i{+}(3{,}734e{+}01)tamano_i{+}(5{,}359e{+}00)$$

donde :

- Los gastos estimados son iguales a $3{,}590e+02$ euros (constantes las demás variables).
- Por cada mil euros de ingresos, los gastos aumentan en $7{,}247e-03$, supuesto que permanecen constantes las otras variables.
- Por cada aumento del tamaño de la familia en un familiar, los gastos estimados aumentan en $3{,}734e+01$, suponiendo que se mantienen constantes las otras variables
- Por cada aumento del número de hijos estudiando en la Universidad, los gastos estimados aumentan en $5{,}359e+00$, suponiendo que se mantienen constantes las otras variables

Regresión lineal múltiple

Ejemplo 2 : **summary(reg_lin_mul)**

Tanto la interpretación como la comprobación de la significación de los parámetros se realizan de forma similar al caso en que se cuenta con una única variable independiente. Igualmente, la validación se lleva a cabo del mismo modo que para la regresión lineal simple.
El p-valor asociado al contraste (0.009772) es menor que a = 0.05 , por lo que rechazamos la hipótesis nula. Esto implica que al menos una de las variables independientes contribuye de forma significativa a la explicación de la variable respuesta.
Para las variables tamaño familiar y número de hijos en la Universidad, los p-valores son 0.1655 y 0.9034, respectivamente. Ambos mayores que 0.05, por lo que no rechazamos la hipótesis nula de significación de ambas variables. Estas variables no son válidas para predecir los gastos alimentación mensual de una familia y por tanto se pueden eliminar del modelo.

Regresión lineal múltiple

Ejemplo 3

El conjunto de datos a usar en este ejemplo se llama **Prestige** y proviene del paquete llamado **car**, este es un data.frame de 102 filas y 6 columnas . Cada fila es una profesión (u ocupación), las columnas se refieren a predictores como el promedio de años de educación, el porcentaje de mujeres en la ocupación,el prestigio de la ocupación, etc .

Regresión lineal múltiple

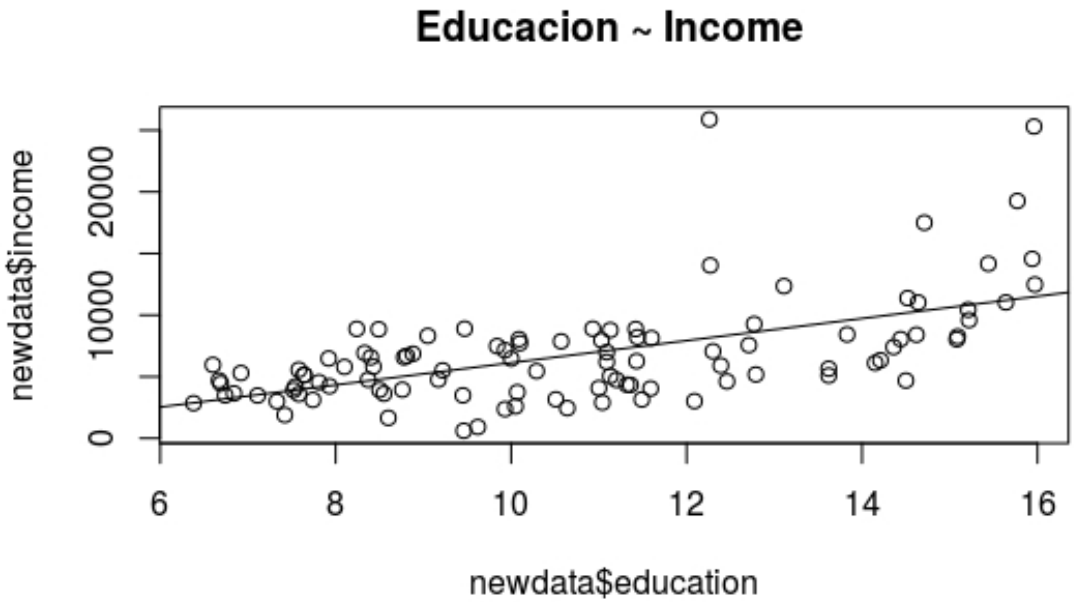
Ejemplo 3

Para obtener el data.frame tenemos que instalar la librería **car** y luego cargarla en la sesión de R

```
library(car)
head(Prestige,5)
newdata = Prestige(,c(1:2))
summary(newdata)
modelo = lm(income ~ education, data = newdata)
plot(newdata$education, newdata$income, main="Educacion ~ Income ")
abline(modelo)
summary(modelo)
```

Regresión lineal múltiple

Ejemplo 3



Regresión lineal múltiple

Ejemplo 3

```
summary(modelo)
Call:
lm(formula = income ~ education, data = newdata)

Residuals:
Min      1Q  Median      3Q     Max
-5493.2 -2433.8   -41.9  1491.5 17713.1

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2853.6     1407.0  -2.028   0.0452 *
education      898.8      127.0   7.075 2.08e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3483 on 100 degrees of freedom
Multiple R-squared:  0.3336,    Adjusted R-squared:  0.3269
F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10
```

Regresión lineal múltiple

Ejemplo 3

La ecuación para el modelo lineal es la siguiente :

$$income = 898{,}8(education) - 2853{,}6$$

Obs.

(`education=0`) $\Longrightarrow$ (**NO tener educación** $\equiv$ **debes dinero**)

Modelo de Regresión Logística

Parte de la hipótesis que los datos siguen el siguiente modelo:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \mu = b\, x + \mu$$

Con el fin de simplificar notaciones, definamos :

$$z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Por consiguiente el modelo se puede expresar como

$$ln\left(\frac{p}{1-p}\right) = z + \mu$$

donde p es la probabilidad de que ocurra el suceso en estudio.

Modelo de Regresión Logística

Operando algebráicamente sobre el modelo:

$$ln\left(\frac{p}{1-p}\right) = z$$

$$\frac{p}{1-p} = e^z$$

$$p = \frac{e^z}{1+e^z}$$

Regresión Logistica

Modelo de Regresión Logística

De donde se deduce que el modelo de regresión logística es, en principio, un modelo de regresión no lineal, pero es lineal en escala logarítmica atendiendo a su definición original:

$$ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p) = z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Es decir, la diferencia de la probabilidad de que ocurra un suceso respecto de que no ocurra es lineal pero en escala logarítmica. Por tanto, el significado de los coeficientes, aunque guardando una cierta relación con el modelo de regresión lineal, va a ser algo más complejo de interpretar.

Regresión Logistica

Modelo de Regresión Logística

Recordemos las dos formas más importantes de expresar el modelo de regresión logística:

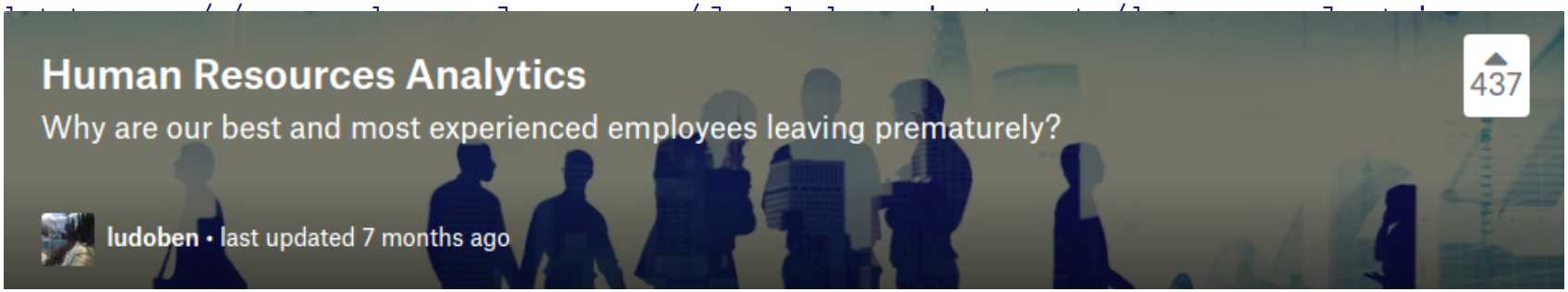
- $\ln(p) - \ln(1-p) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$
- $\dfrac{p}{1-p} = e^{b_0}\ e^{b_1 x_1}\ e^{b_2 x_2} \ldots e^{b_k x_k}$

La primera expresión se llama **logit** y la segunda **odds ratio**.

Regresión Logistica

Ejemplo : Modelo de Regresión Logística

El conjunto de datos a utilizar proviene de la plataforma de concursos de análisis de datos kaggle.



Regresión Logistica

Ejemplo : Modelo de Regresión Logística

Para este ejemplo ilustrativo se han seleccionado aleatoriamente 100 registros

```
datos <- read.csv("HumanResourcesAnalytics.csv",T)
muestra <- dim(datos)[1]
datos <- datos[sample(muestra,100,replace=TRUE),]
class(datos)
str(datos)
head(datos)
# View(datos)
colnames(datos) = c("nivel_satisfaccion","ultima_evaluacion","numero_
    proyectos","promedio_horas_mensuales","antiguedad","accidente","
    abandona","promocionado","departamento","salario")
```

Regresión Logistica

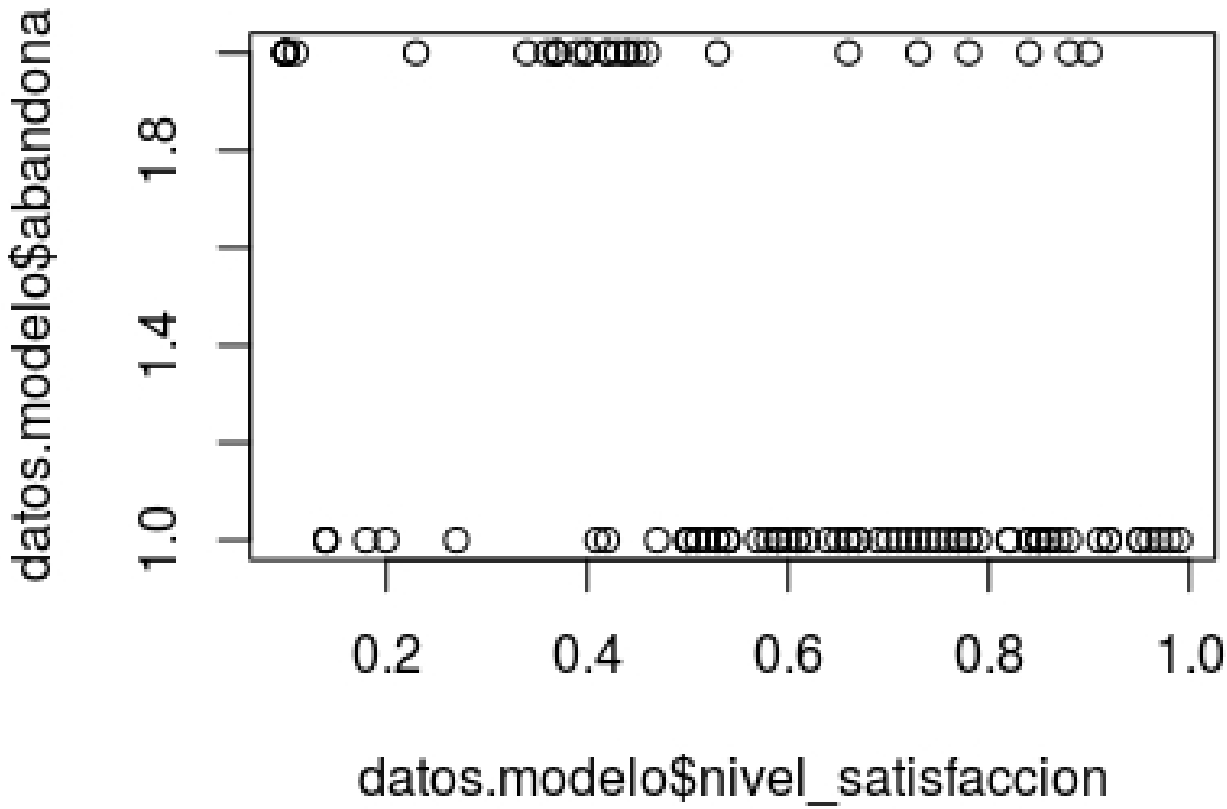
Ejemplo : Modelo de Regresión Logística

En este conjunto de datos tenemos tres variables sobre las que podemos estimar un modelo logit: **accidente**, **abandona** y **promocionado**. Un estudio de interés puede ser intentar explicar/predecir si un empleado abandonará o no la empresa en función de las puntuaciones en las variables **nivel_satisfaccion** y **ultima_evaluacion**. Para facilitar la explicación vamos a seleccionar solo las variables del modelo:

```
1 datos.modelo <- subset(datos, select = c(abandona, nivel_satisfaccion,
      ultima_evaluacion))
2 datos.modelo$abandona <- factor(datos.modelo$abandona)
3 head(datos.modelo)
4 plot(datos.modelo$nivel_satisfaccion,datos.modelo$abandona)
```

Regresión Logistica

Ejempl



Regresión Logistica

Ejemplo : Modelo de Regresión Logística

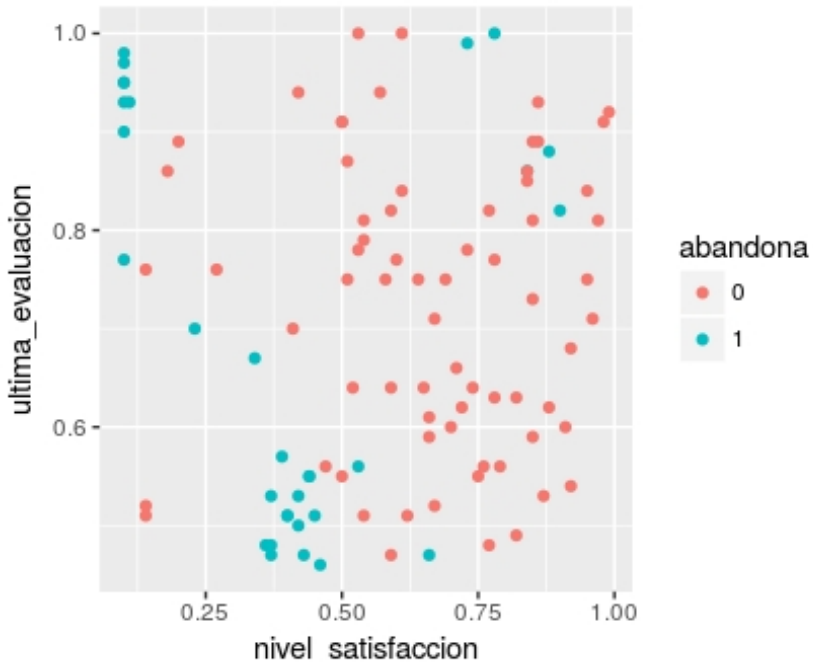
Además, debemos asegurarnos de que la variable respuesta factor solo toma dos valores1. Descriptivamente, podemos hacer un resumen de cada categoría:

```
table(datos.modelo$abandona)
summary(datos.modelo$nivel_satisfaccion)
summary(datos.modelo$ultima_evaluacion)
```

Regresión Logistica

Ejemplo : Modelo de Regresión Logística

Además, debemos asegurarnos de que la variable respuesta factor solo toma dos valores1. Gráfic



Regresión Logistica

Ejemplo : Modelo de Regresión Logística

En R, los GLMs se ajustan con la función **glm**. La principal diferencia con la función **lm** para ajustar modelos lineales es que le tenemos que proporcionar la familia de la distribución. En nuestro caso, como es una variable dicotómica, la familia es la binomial:

```
1 modelo.logit <- glm(abandona ~ ultima_evaluacion + nivel_satisfaccion ,
      data = datos.modelo, family = "binomial")
2 summary(modelo.logit)
```

Regresión Logistica

Ejemplo : Modelo de Regresión Logística

La interpretación de los p-valores es similar a la del modelo lineal. Podemos ver que la variable **ultima_evaluacion** no es significativa en el modelo ,el p-valor debe ser mucho mayor de 0.05, mientras que la variable **nivel_satisfaccion** es moderadamente significativa si es que su p-valor entre 0.01 y 0.05.

Ejemplo : Modelo de Regresión Logística

En cuanto a los coeficientes, la interpretación cambia. El modelo GLM no ajusta la variable respuesta sino una función de enlace. En el caso del modelo logit esta función es:

$$\eta = \ln\left(\frac{p}{1-p}\right)$$

siendo p la probabilidad de que el individuo tome el valor de 1 en la variable dicotómica.Al cociente $\left(\frac{p}{1-p}\right)$ se le conoce como **odds ratio**. Por tanto , los coeficientes del modelo logit se interpretan como el logaritmo del odds ratio. Si nos fijamos en el coeficiente de la variable **nivel_satisfaccion** (-2.163), nos está indicando que el logaritmo del odds ratio de abandonar la empresa disminuye 2.163 unidades por cada unidad que aumenta la puntuación en el nivel de satisfacción.

Regresión Logistica

Ejemplo : Modelo de Regresión Logística

Una forma de facilitar la interpretación de los coeficientes es evaluando en la exponencial :

```
exp(coefficients(modelo.logit))
## (Intercept)  ultima_evaluacion  nivel_satisfaccion
##   0.9713175       1.3630515          0.1150215
```

Que se corresponde con este modelo:

$$odds = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2}$$

Lo interpretaremos de la siguiente manera, que aumentar en la última evaluación un punto aumenta un 36 % las posibilidades de abandonar la empresa, mientras que aumentar un punto en el nivel de satisfacción las reduce casi un 90 %.

Ejemplo : Modelo de Regresión Logística

Por último, una interpretación probabilista sería estimar la probabilidad p de que un individuo abandone la empresa:

$$p = \frac{e^{\eta}}{1 - e^{\eta}}$$

Así, podemos predecir la función η para un individuo que tenga, por ejemplo, una evaluación de 0.75 y un nivel de satisfacción de 0.6:

```
log.odds <- predict(modelo.logit, data.frame(nivel_satisfaccion = 0.6,
    ultima_evaluacion = 0.75))
log.odds
##          1
## -1.094389
```

Regresión Logistica

Ejemplo : Modelo de Regresión Logística

La probabilidad de abandonar la empresa sería:

```
exp(log.odds)/(1-exp(log.odds))
##         1
##  0.5031808
```

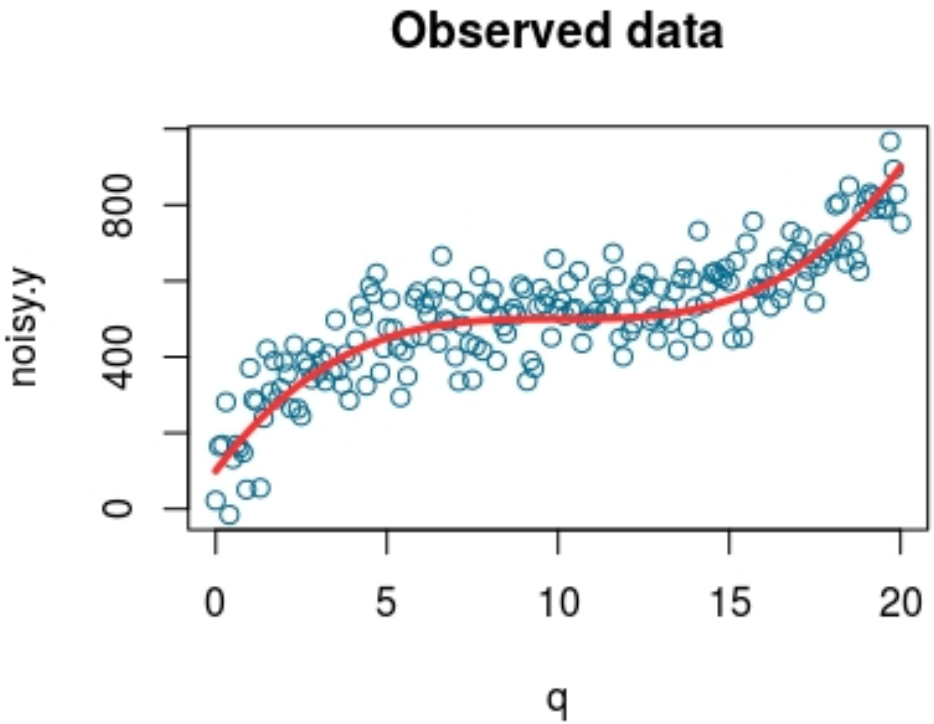
Regresión Polinomial

Ejemplo : Modelo de Regresión Polinomial

```r
q <- seq(from=0, to=20, by=0.1)
y <- 500 + 0.4 * (q-10)^3
noise <- rnorm(length(q), mean=10, sd=80)
noisy.y <- y + noise
plot(q,noisy.y,col='deepskyblue4',xlab='q',main='Observed data')
lines(q,y,col='firebrick1',lwd=3)
model <- lm(noisy.y ~ poly(q,3))
```

Regresión Polinomial

Ejemplo : Mc



Regresión Polinomial

Ejemplo : Modelo de Regresión Polinomial

Utilizando la función **confint** podemos obtener los intervalos de confianza de los parámetros de nuestro modelo. Intervalos de confianza para los parámetros del modelo:

```
confint(model, level=0.95)

2.5 %       97.5 %
(Intercept)    495.3353    517.7783
poly(q, 3)1    1897.4477   2215.6316
poly(q, 3)2    -199.6034   118.5805
poly(q, 3)3    805.0122    1123.1960
```

Regresión de Poisson

DATOS DE RECUENTO

Se denominan variables de recuento **(count data)** a aquéllas que toman valores positivos, enteros (incluido el cero).

- Economía de la salud: Número de veces que los individuos acudieron a un determinado servicio médico; número de episodios de enfermedad durante un periodo de tiempo.
- Economía del transporte: El número de viajes efectuados en un determinado medio de transporte, o a un determinado lugar.
- Economía industrial: Número de patentes registradas por las empresas
- Economía de la familia: Número de hijos
- Finanzas: Número de clientes embargados por impago de hipotecas en diferentes entidades bancarias.

Regresión de Poisson

DATOS DE RECUENTO : ¿POR QUÉ UTILIZAMOS MODELOS ESPECÍFICOS PARA DATOS DE RECUENTO?

- MODELO DE REGRESIÓN LINEAL
  1. Las predicciones de Y pueden salirse del rango de valores en el que está definido.
  2. Las estimaciones pueden ser inconsistentes.
  3. Puede tener validez para hacer una exploración previa de las relaciones
- MODELOS DE ELECCIÓN BINARIA
  1. Si la variable Y toma muchos valores, plantear un modelo de elección binaria nos conduce a un pérdida de eficiencia (porque perdemos información) ya que agregamos todos los valores mayores que 0 en un solo valor.

Regresión de Poisson

En una regresión logística se predice una respuesta que viene en una de dos formas, cara o sello, varón o niña, hubo terremoto o no hubo. La generalización a un tipo de respuesta que viene en varios eventos discretos que pueden ser más de dos se llama la regresión de Poisson.
Analicemos los datos de una clase, dado que nuestra tarea es predecir la nota, que puede tomar los valores 1,2,3,4,5 a partir de dos descriptores, el primero mide el trabajo en casa y el segundo la asistencia a clase:

```
rm(list = ls())
NumResueltos <- c(0,1,2,1,5,3,2,5,7,8,12,13,12,11,10,12,10,15 )
HorasClase <-   c(1,3,4,1,3,5,1,3,5,2, 3, 5, 0, 3, 5, 4, 3, 5)
Nota <-       c(0,2,3,0,4,3,1,3,4,3, 4, 5, 4, 5, 4, 5, 5,5)
tabla <- data.frame(NumResueltos, HorasClase, Nota)
```

Regresión de Poisson

```
regPoisson <- glm(Nota ~ NumResueltos + HorasClase, data = tabla, family
    = poisson())
summary(regPoisson)
predict(regPoisson, type="response")
```