Hola

Enfoque Estadístico del Aprendizaje

Linealización, evaluación y diagnóstico

Bernardo L. Español · 5 de Septiembre

Linealización

LINEALIZACIÓN

¿Qué significa linealizar?

LINEALIZACIÓN

¿Qué significa linealizar?

En muchos casos, una ecuación **no lineal** puede **linealizarse** usando un poco de matemática.

LINEALIZACIÓN

¿Qué significa linealizar?

En muchos casos, una ecuación **no lineal** puede **linealizarse** usando un poco de matemática.

Esto permite utilizar métodos de regresión lineal sobre las **variables transformadas** y en base a los parámetros del ajuste recuperar los **parámetros originales**.

LINEALIZACIÓN

¿Qué significa linealizar?

En muchos casos, una ecuación **no lineal** puede **linealizarse** usando un poco de matemática.

Esto permite utilizar métodos de regresión lineal sobre las **variables transformadas** y en base a los parámetros del ajuste recuperar los **parámetros originales**.

$$y = f(x \mid a, b) \qquad \longrightarrow \qquad Y = \beta_0 + \beta_1 X$$

LINEALIZACIÓN

¿Qué significa linealizar?

En muchos casos, una ecuación **no lineal** puede **linealizarse** usando un poco de matemática.

Esto permite utilizar métodos de regresión lineal sobre las **variables transformadas** y en base a los parámetros del ajuste recuperar los **parámetros originales**.

$$y = f(x \mid a, b) \qquad \longrightarrow \qquad Y = \beta_0 + \beta_1 X$$

donde Y y X son transformaciones de las variables originales x, y.

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Si aplicamos logaritmo natural a cada lado vemos que

$$\log(y) = \log\left(ae^{bx}\right)$$

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Si aplicamos logaritmo natural a cada lado vemos que

$$\log(y) = \log\left(ae^{bx}\right) = \log(a) + \log\left(e^{bx}\right)$$

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Si aplicamos logaritmo natural a cada lado vemos que

$$\log(y) = \log\left(ae^{bx}\right) = \log(a) + \log\left(e^{bx}\right) = \log(a) + bx$$

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Si aplicamos logaritmo natural a cada lado vemos que

$$\log(y) = \log\left(ae^{bx}\right) = \log(a) + \log\left(e^{bx}\right) = \log(a) + bx$$

$$\implies \qquad \log(y) = \log(a) + b\,x$$

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Si aplicamos logaritmo natural a cada lado vemos que

$$\log(y) = \log\left(ae^{bx}\right) = \log(a) + \log\left(e^{bx}\right) = \log(a) + bx$$

$$\Longrightarrow \qquad \log(y) = \log(a) + b\,x$$

Ajustando $Y = \beta_0 + \beta_1 X, \quad$ con $\quad Y = \log(y), X = x, \quad \beta_0 = \log(a), \beta_1 = b.$

Ejemplo 1: Exponencial

Supongamos que nuestros datos están bien modelados por una relación exponencial:

$$y = ae^{bx}$$

donde a y b son parámetros desconocidos.

Si aplicamos logaritmo natural a cada lado vemos que

$$\log(y) = \log\left(ae^{bx}\right) = \log(a) + \log\left(e^{bx}\right) = \log(a) + bx$$

$$\Longrightarrow \qquad \log(y) = \log(a) + b\,x$$

Ajustando $Y = \beta_0 + \beta_1 X, \quad$ con $\quad Y = \log(y), X = x, \quad \beta_0 = \log(a), \beta_1 = b.$

Podemos recuperar los parámetros originales invirtiendo la transformación:

$$a = e^{\beta_0} \qquad b = \beta_1$$

Ejemplos 2 y 3: Leyes de potencia e inversa

Ejemplos 2 y 3: Leyes de potencia e inversa

- Leyes de potencias: $y = ax^b$

Tomando logaritmo a ambos lados obtenemos una relación lineal entre $\log(y)$ y $\log(x)$:

$$\log(y) = \log(a) + b\log(x)$$

Podemos recuperar los parámetros originales tomando $a = e^{\beta_0}, b = \beta_1$

Ejemplos 2 y 3: Leyes de potencia e inversa

- Leyes de potencias: $y = ax^b$

Tomando logaritmo a ambos lados obtenemos una relación lineal entre $\log(y)$ y $\log(x)$:

$$\log(y) = \log(a) + b\log(x)$$

Podemos recuperar los parámetros originales tomando $a = e^{\beta_0}, b = \beta_1$

- Inversa: $y = \frac{1}{a+bx}$

Tomando la inversa a ambos lados obtenemos una relación lineal entre $\frac{1}{y}$ y x:

$$\frac{1}{y} = a + bx$$

En este caso los parámetros originales coinciden con los transformados: $a = \beta_0, b = \beta_1$

LINEALIZACIÓN

Ejemplo en código

```r
# Creo data exponencial artificialmente
set.seed(0)
x <- seq(1, 10, length.out = 50)
y <- 2 * exp(0.8 * x)
y <- y * (1 + rnorm(length(x), 0, 0.2))
df <- data.frame(x = x, y = y)
```

LINEALIZACIÓN

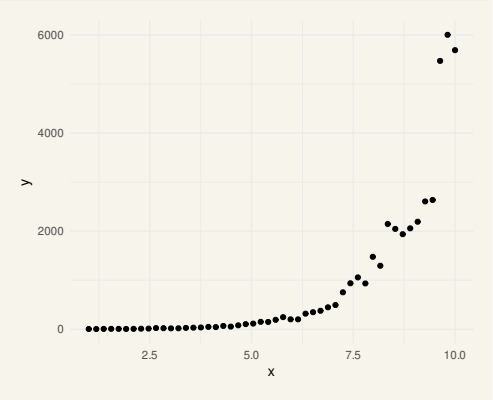
Ejemplo en código

```r
# Creo data exponencial artificialmente
set.seed(0)
x <- seq(1, 10, length.out = 50)
y <- 2 * exp(0.8 * x)
y <- y * (1 + rnorm(length(x), 0, 0.2))
df <- data.frame(x = x, y = y)

# La grafico
ggplot(df, aes(x, y)) +
  geom_point(color = "black") +
  theme_minimal()
```

LINEALIZACIÓN

Ejemplo en código

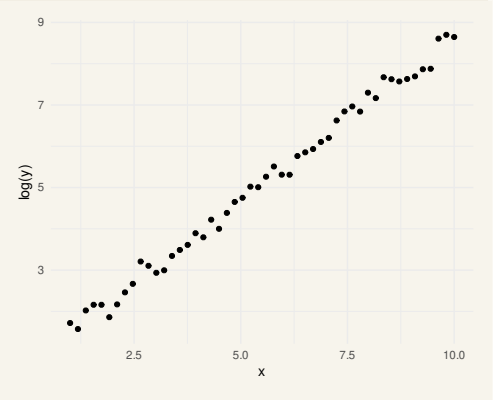
```r
# Creo data exponencial artificialmente
set.seed(0)
x <- seq(1, 10, length.out = 50)
y <- 2 * exp(0.8 * x)
y <- y * (1 + rnorm(length(x), 0, 0.2))
df <- data.frame(x = x, y = y)

# La grafico
ggplot(df, aes(x, y)) +
  geom_point(color = "black") +
  theme_minimal()
```



LINEALIZACIÓN

Ejemplo en código

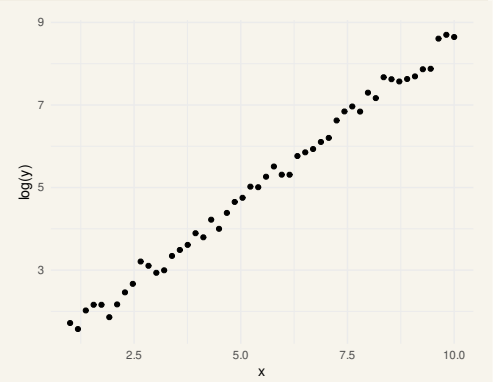
```
# Mismo gráfico pero en semi-logy
ggplot(df, aes(x, log(y))) +
  geom_point(color = "black") +
  theme_minimal()
```



LINEALIZACIÓN

Ejemplo en código

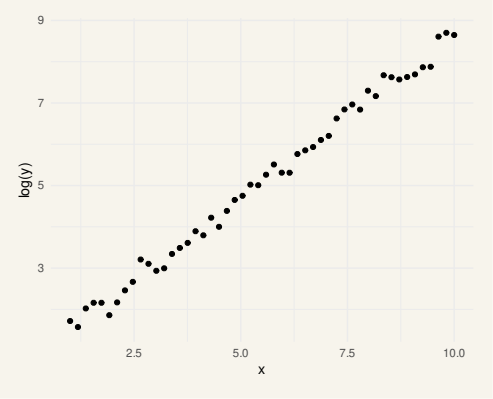
```r
# Modelo linealizado
mod <- lm(log(y) ~ x)
a <- exp(mod$coef[1])
b <- mod$coef[2]
```



LINEALIZACIÓN

Ejemplo en código

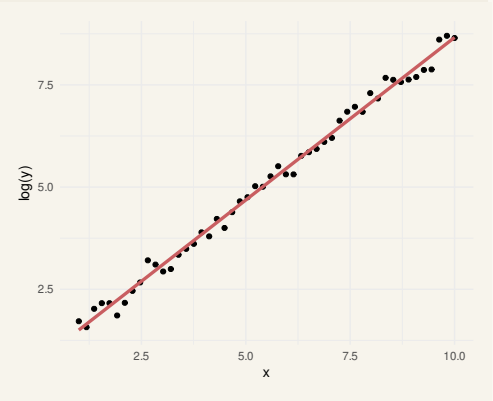
```r
# Modelo linealizado
mod <- lm(log(y) ~ x)
a <- exp(mod$coef[1])
b <- mod$coef[2]

# Plot en semi-logy
ggplot(df, aes(x, log(y))) +
  geom_point(color = "black") +
```



LINEALIZACIÓN

Ejemplo en código

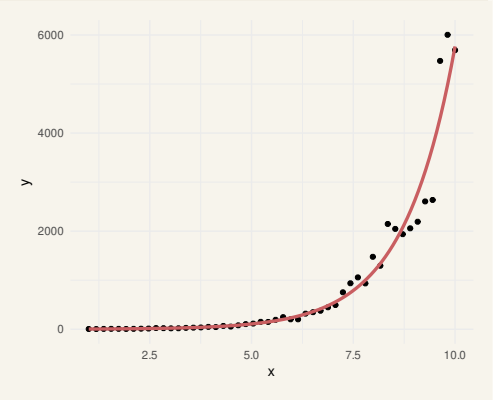
```r
# Modelo linealizado
mod <- lm(log(y) ~ x)
a <- exp(mod$coef[1])
b <- mod$coef[2]

# Plot en semi-logy
ggplot(df, aes(x, log(y))) +
  geom_point(color = "black") +
  stat_function(
    fun = \(x) log(a) + b * x,
    col = "#C95E61",
  ) +
  theme_minimal()
```



LINEALIZACIÓN

Ejemplo en código

```
# Modelo linealizado
mod <- lm(log(y) ~ x)
a <- exp(mod$coef[1])
b <- mod$coef[2]

# Plot en cords. originales
ggplot(df, aes(x = x, y = y)) +
  geom_point(color = "black") +
  stat_function(
   fun = function(x) a*exp(b*x),
   col = "#C95E61"
  ) +
  theme_minimal()
```



Evaluación y Diagnóstico

INTRODUCCIÓN

Librerías y datos

```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

INTRODUCCIÓN

Librerías y datos

```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

```
# Igual que la clase pasada
df_filt_18_40 <- read_delim("usu_individual_T126.txt", delim = ";") %>%
```

INTRODUCCIÓN

Librerías y datos

```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

```
# Igual que la clase pasada
df_filt_18_40 <- read_delim("usu_individual_T126.txt", delim = ";") %>%
  rename(edad = CH06, salario = P21) %>%
```

INTRODUCCIÓN

Librerías y datos

```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

```
# Igual que la clase pasada
df_filt_18_40 <- read_delim("usu_individual_T126.txt", delim = ";") %>%
  rename(edad = CH06, salario = P21) %>%
  filter(salario > 0, ESTADO == 1, between(edad, 18, 40))
```

INTRODUCCIÓN

Librerías y datos

```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

```
# Igual que la clase pasada
df_filt_18_40 <- read_delim("usu_individual_T126.txt", delim = ";") %>%
  rename(edad = CH06, salario = P21) %>%
  filter(salario > 0, ESTADO == 1, between(edad, 18, 40))

edad_salario_18_40 <- df_filt_18_40 %>%
  group_by(edad) %>%
```

INTRODUCCIÓN

Librerías y datos

```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

```
# Igual que la clase pasada
df_filt_18_40 <- read_delim("usu_individual_T126.txt", delim = ";") %>%
  rename(edad = CH06, salario = P21) %>%
  filter(salario > 0, ESTADO == 1, between(edad, 18, 40))

edad_salario_18_40 <- df_filt_18_40 %>%
  group_by(edad) %>%
  summarise(salario_prom = mean(salario)) %>%
```

INTRODUCCIÓN

Librerías y datos

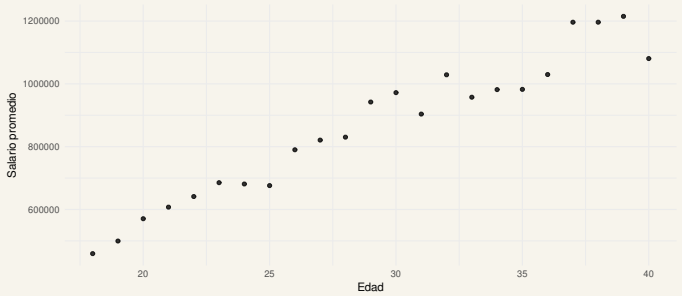
```
library(tidyverse)
library(broom) # mejores métricas para los modelos
```

```
# Igual que la clase pasada
df_filt_18_40 <- read_delim("usu_individual_T126.txt", delim = ";") %>%
  rename(edad = CH06, salario = P21) %>%
  filter(salario > 0, ESTADO == 1, between(edad, 18, 40))

edad_salario_18_40 <- df_filt_18_40 %>%
  group_by(edad) %>%
  summarise(salario_prom = mean(salario)) %>%
  ungroup() %>%
  arrange(edad)
```

GRÁFICOS CON GGPLOT

Edad vs salario promedio

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  labs(x = "Edad", y = "Salario promedio") +
  theme_minimal()
```



Ajuste

```
mod <- lm(formula = salario_prom ~ edad, data = edad_salario_18_40)
```

Evaluación del modelo

EVALUACIÓN DEL MODELO

Visual

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
```

EVALUACIÓN DEL MODELO

Visual

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  geom_abline(
    intercept = coef(mod)[1],
    slope = coef(mod)[2],
    linewidth = 1.2, col = "#C95E61"
  ) +
```

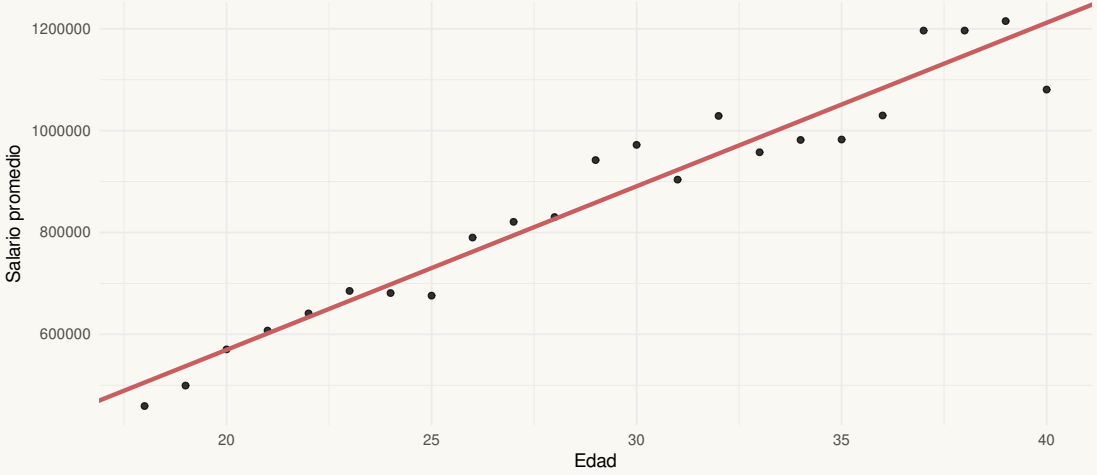
EVALUACIÓN DEL MODELO

Visual

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  geom_abline(
    intercept = coef(mod)[1],
    slope = coef(mod)[2],
    linewidth = 1.2, col = "#C95E61"
  ) +
  labs(x = "Edad", y = "Salario promedio") +
  theme_minimal()
```

EVALUACIÓN DEL MODELO

Visual



EVALUACIÓN DEL MODELO

`summary()`: resumen por defecto de R

```
summary(mod)
```

`summary()`: resumen por defecto de R

```
summary(mod)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-131218  -37793    3782   31569   83566
```

`summary()`: resumen por defecto de R

```
summary(mod)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-131218  -37793    3782   31569   83566

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -72558      52216   -1.39    0.179
edad           32110       1755   18.29 2.21e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`summary()`: resumen por defecto de R

```
summary(mod)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-131218  -37793    3782   31569   83566

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -72558      52216   -1.39    0.179
edad           32110       1755   18.29 2.21e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55840 on 21 degrees of freedom
Multiple R-squared:  0.941,     Adjusted R-squared:  0.9381
F-statistic: 334.7 on 1 and 21 DF,  p-value: 2.212e-14
```

EVALUACIÓN DEL MODELO

`broom::tidy()`: métricas de los parámetros

```
tidy(mod, conf.int = TRUE)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	−72560	52220	−1.39	1.792e−01	−181100	36030
edad	32110	1755	18.29	2.212e−14	28460	35760

EVALUACIÓN DEL MODELO

`broom::glance()`: métricas globales

```
glance(mod)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual	nobs
0.941	0.9381	55840	334.7	2.212e−14	1	21	23

`predict()`: la incerteza de la recta

`tidy` y `glance` nos dieron la incerteza de los parámetros.
Para la de un punto de la recta está `predict`:

```
nuevos <- tibble(edad = c(25, 35, 40))

predict(mod, newdata = nuevos, interval = "confidence")
```

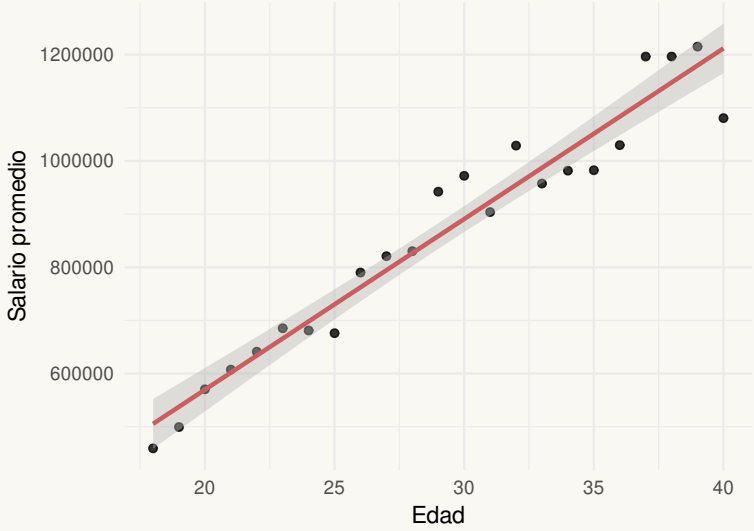
```
        fit       lwr       upr
1  730201.6  701927.7  758475.6 # 25
2 1051305.3 1018657.4 1083953.3 # 35
3 1211857.2 1164970.1 1258744.2 # 40
```

`predict()`: la incerteza de la recta

En `ggplot` esa banda es lo que dibuja `geom_smooth(method = "lm")`:

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  geom_smooth(method = "lm", col = "#C95E61") +
  theme_minimal()
```

`predict()`: la incerteza de la recta



EVALUACIÓN DEL MODELO

Promedios vs personas

Variabilidad de los 23 promedios respecto a la recta:

```
predict(mod, newdata = nuevos, interval = "confidence", level=95)
```

```
        fit       lwr       upr
1  730201.6  701927.7  758475.6 # 25
2 1051305.3 1018657.4 1083953.3 # 35
3 1211857.2 1164970.1 1258744.2 # 40
```

EVALUACIÓN DEL MODELO

Promedios vs personas

Variabilidad de los 23 promedios respecto a la recta:

```
predict(mod, newdata = nuevos, interval = "confidence", level=95)
```

```
        fit       lwr       upr
1  730201.6  701927.7  758475.6 # 25
2 1051305.3 1018657.4 1083953.3 # 35
3 1211857.2 1164970.1 1258744.2 # 40
```

Variabilidad dentro de cada edad y cuánta gente hay en cada una:

```
mod_personas <- lm(salario ~ edad, data = df_filt_18_40)
predict(mod_personas, newdata = nuevos, interval = "confidence", level=95)
```

```
      fit       lwr       upr
1  738411  706168.4  770653.6 # 25
2 1046726 1018354.1 1075098.5 # 35
3 1200884 1158429.3 1243338.7 # 40
```

EVALUACIÓN DEL MODELO

Promedios vs personas

Promedios vs personas

El segundo es el que corresponde para un intervalo de la media.

EVALUACIÓN DEL MODELO

Promedios vs personas

El segundo es el que corresponde para un intervalo de la media.

En los dos casos la banda es de **la recta**: no dice nada sobre cuánto gana una persona.

EVALUACIÓN DEL MODELO

confidence vs prediction

EVALUACIÓN DEL MODELO

confidence vs prediction

`interval = "confidence"` contesta *¿dónde está la recta en* $x = 35$*?*

EVALUACIÓN DEL MODELO

confidence vs prediction

`interval = "confidence"` contesta *¿dónde está la recta en $x = 35$?*

`interval = "prediction"` contesta *¿cuánto gana una persona de 35?*

EVALUACIÓN DEL MODELO

¿Cuánto gana una persona al azar?

```
predict(mod_personas, newdata = nuevos, interval = "prediction")
```

```
      fit        lwr     upr
1  738411 -1281570.3 2758392 # 25
2 1046726  -973197.0 3066650 # 35
3 1200884  -819286.2 3221054 # 40
```

EVALUACIÓN DEL MODELO

¿Cuánto gana una persona al azar?

```
predict(mod_personas, newdata = nuevos, interval = "prediction")
```

```
      fit        lwr     upr
1  738411 -1281570.3 2758392 # 25
2 1046726  -973197.0 3066650 # 35
3 1200884  -819286.2 3221054 # 40
```

Para una persona de 25 años el intervalo va de −$1.281.570 a $2.758.392.

EVALUACIÓN DEL MODELO

¿Cuánto gana una persona al azar?

```
predict(mod_personas, newdata = nuevos, interval = "prediction")
```

```
      fit        lwr     upr
1  738411 -1281570.3 2758392 # 25
2 1046726  -973197.0 3066650 # 35
3 1200884  -819286.2 3221054 # 40
```

Para una persona de 25 años el intervalo va de −$1.281.570 a $2.758.392.

Sobre datos individuales el supuesto de normalidad ya no se sostiene.

EVALUACIÓN DEL MODELO

El salario no es normal, su log sí

```
# El salario, con la normal de su misma media y desvío encima
m_sal <- mean(df_filt_18_40$salario); s_sal <- sd(df_filt_18_40$salario)
```

EVALUACIÓN DEL MODELO

El salario no es normal, su log sí

```
# El salario, con la normal de su misma media y desvío encima
m_sal <- mean(df_filt_18_40$salario); s_sal <- sd(df_filt_18_40$salario)

ggplot(df_filt_18_40, aes(x = salario)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 2e5,
                 fill = "grey80", colour = "white") +
```

EVALUACIÓN DEL MODELO

El salario no es normal, su log sí

```
# El salario, con la normal de su misma media y desvío encima
m_sal <- mean(df_filt_18_40$salario); s_sal <- sd(df_filt_18_40$salario)

ggplot(df_filt_18_40, aes(x = salario)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 2e5,
                 fill = "grey80", colour = "white") +
  stat_function(fun = dnorm, args = list(mean = m_sal, sd = s_sal),
                col = "#C95E61", linewidth = 1.2) +
  theme_minimal()
```

EVALUACIÓN DEL MODELO

El salario no es normal, su log sí

```
# El salario, con la normal de su misma media y desvío encima
m_sal <- mean(df_filt_18_40$salario); s_sal <- sd(df_filt_18_40$salario)

ggplot(df_filt_18_40, aes(x = salario)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 2e5,
                 fill = "grey80", colour = "white") +
  stat_function(fun = dnorm, args = list(mean = m_sal, sd = s_sal),
                col = "#C95E61", linewidth = 1.2) +
  theme_minimal()

# Lo mismo sobre el log
m_log <- mean(log(df_filt_18_40$salario)); s_log <- sd(log(df_filt_18_40$salario))
```

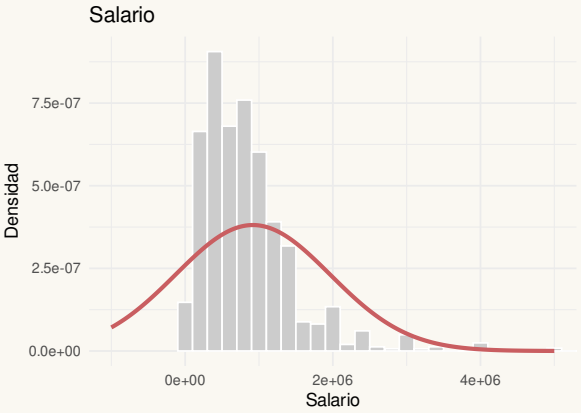
EVALUACIÓN DEL MODELO

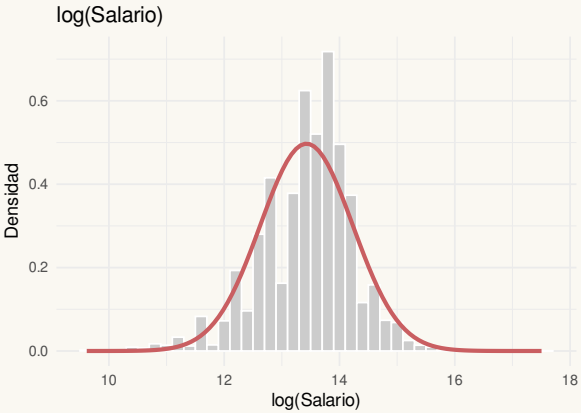
El salario no es normal, su log sí

```r
# El salario, con la normal de su misma media y desvío encima
m_sal <- mean(df_filt_18_40$salario); s_sal <- sd(df_filt_18_40$salario)

ggplot(df_filt_18_40, aes(x = salario)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 2e5,
                 fill = "grey80", colour = "white") +
  stat_function(fun = dnorm, args = list(mean = m_sal, sd = s_sal),
                col = "#C95E61", linewidth = 1.2) +
  theme_minimal()

# Lo mismo sobre el log
m_log <- mean(log(df_filt_18_40$salario)); s_log <- sd(log(df_filt_18_40$salario))

ggplot(df_filt_18_40, aes(x = log(salario))) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.2,
                 fill = "grey80", colour = "white") +
  stat_function(fun = dnorm, args = list(mean = m_log, sd = s_log),
                col = "#C95E61", linewidth = 1.2) +
  theme_minimal()
```

EVALUACIÓN DEL MODELO

El salario no es normal, su log...





EVALUACIÓN DEL MODELO

¿Y sigue siendo una recta?

```
ggplot(df_filt_18_40, aes(x = edad, y = log(salario))) +
  geom_point(alpha = 0.05) +
```

EVALUACIÓN DEL MODELO

¿Y sigue siendo una recta?

```
ggplot(df_filt_18_40, aes(x = edad, y = log(salario))) +
  geom_point(alpha = 0.05) +

  # la recta (ahora para y = log(salario))
  geom_smooth(method = "lm", se = FALSE,
              col = "#C95E61", linewidth = 1.2) +
```

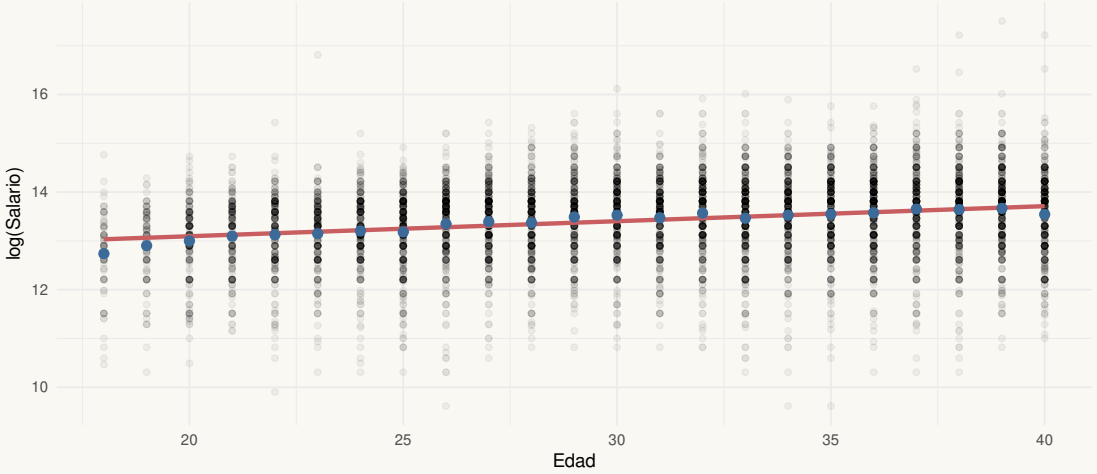
EVALUACIÓN DEL MODELO

¿Y sigue siendo una recta?

```
ggplot(df_filt_18_40, aes(x = edad, y = log(salario))) +
  geom_point(alpha = 0.05) +

  # la recta (ahora para y = log(salario))
  geom_smooth(method = "lm", se = FALSE,
              col = "#C95E61", linewidth = 1.2) +

  # el promedio de log(salario) en cada edad, para ver la forma
  stat_summary(fun = mean, geom = "point",
               col = "#3C6B9A", size = 2.4) +
  theme_minimal()
```

EVALUACIÓN DEL MODELO

¿Sigue siendo una recta?



EVALUACIÓN DEL MODELO

El salario no es normal, su log ...

```
mod_log <- lm(log(salario) ~ edad, data = df_filt_18_40)
exp(predict(mod_log, newdata = nuevos, interval = "prediction"))
```

```
       fit      lwr     upr
1 567580.5 122870.4 2621849 # 25
2 772556.7 167251.2 3568548 # 35
3 901325.9 195092.0 4164129 # 40
```

El salario no es normal, su log ...

```
mod_log <- lm(log(salario) ~ edad, data = df_filt_18_40)
exp(predict(mod_log, newdata = nuevos, interval = "prediction"))
```

```
       fit      lwr     upr
1 567580.5 122870.4 2621849 # 25
2 772556.7 167251.2 3568548 # 35
3 901325.9 195092.0 4164129 # 40
```

Para esa misma persona de 25 el intervalo ahora va de $122.870 a $2.621.849.

EVALUACIÓN DEL MODELO

¿Y cuál es el intervalo de verdad?

```
salarios_25 <- df_filt_18_40 %>% filter(edad == 25) %>% pull(salario)
quantile(salarios_25, c(0.025, 0.975))
```

```
   2.5%   97.5% 
  80000 1708750   # n = 354
```

EVALUACIÓN DEL MODELO

¿Y cuál es el intervalo de verdad?

```
salarios_25 <- df_filt_18_40 %>% filter(edad == 25) %>% pull(salario)
quantile(salarios_25, c(0.025, 0.975))
```

```
   2.5%   97.5%
  80000 1708750   # n = 354
```

`lm(y)` daba −$1.281.570 a $2.758.392.

EVALUACIÓN DEL MODELO

¿Y cuál es el intervalo de verdad?

```
salarios_25 <- df_filt_18_40 %>% filter(edad == 25) %>% pull(salario)
quantile(salarios_25, c(0.025, 0.975))
```

```
   2.5%   97.5%
  80000 1708750   # n = 354
```

`lm(y)` daba −$1.281.570 a $2.758.392.

`lm(log(y))` daba $122.870 a $2.621.849.

Diagnóstico

DIAGNÓSTICO

Supuestos

Los residuos deberían ser:

DIAGNÓSTICO

Supuestos

Los residuos deberían ser:

- aproximadamente normales

DIAGNÓSTICO

Supuestos

Los residuos deberían ser:

- aproximadamente normales
- con media cero

DIAGNÓSTICO

Supuestos

Los residuos deberían ser:

- aproximadamente normales
- con media cero
- varianza constante

DIAGNÓSTICO

Supuestos

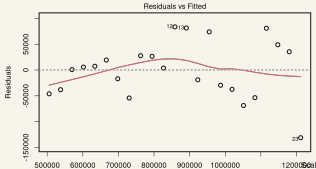
Los residuos deberían ser:

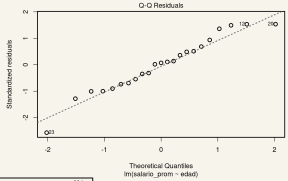
- aproximadamente normales
- con media cero
- varianza constante
- independientes

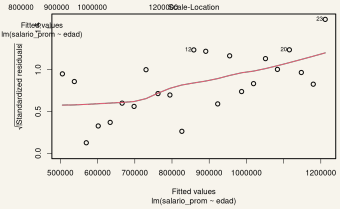
DIAGNÓSTICO

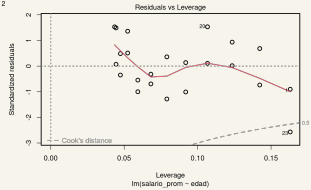
(muchos) Gráficos

```
plot(mod)
```



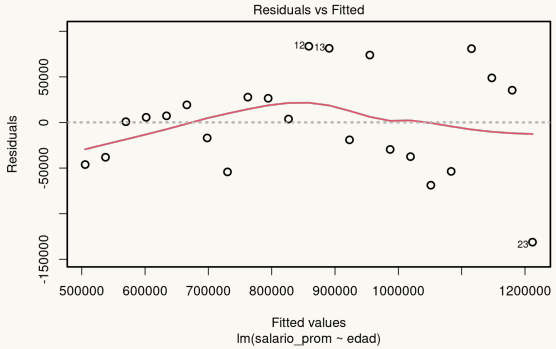






DIAGNÓSTICO

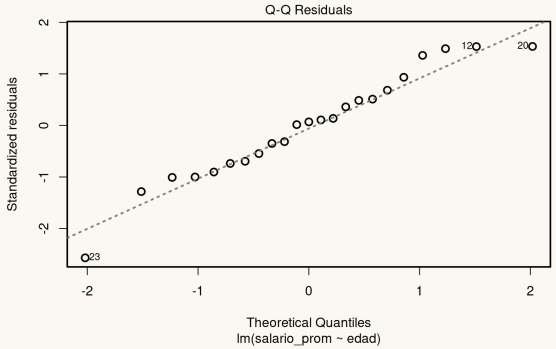
Residuos



- **Se busca:** que la nube no tenga forma.
- **Te dice:** si falta algo no lineal.

DIAGNÓSTICO

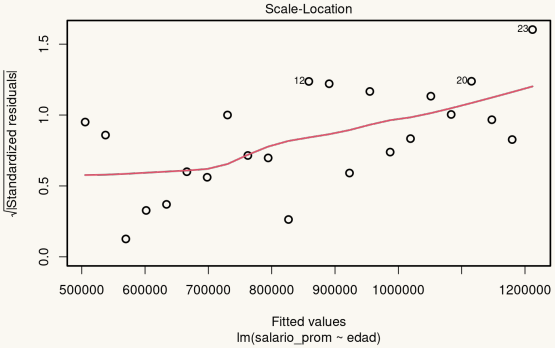
Residuos Q-Q



- **Se busca:** que los puntos sigan la diagonal.
- **Te dice:** si los residuos son normales.

DIAGNÓSTICO

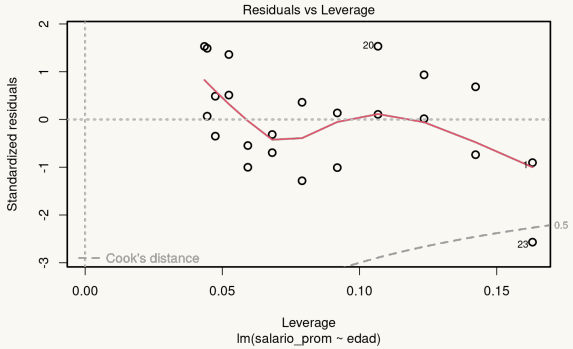
Dispersión de residuos



- **Se busca:** que la línea roja se mantenga plana.
- **Te dice:** si la varianza es constante.

DIAGNÓSTICO

Leverage y Curvas de Cook



- **Se busca:** quién se escapa de las curvas.
- **Te dice:** qué punto manda en el ajuste.

DIAGNÓSTICO

El punto 23

Los cuatro gráficos venían señalando la misma fila:

```
edad_salario_18_40[23, ]
```

```
# A tibble: 1 × 2
   edad salario_prom
  <int>        <dbl>
1    40     1080639.
```

DIAGNÓSTICO

El punto 23

Los cuatro gráficos venían señalando la misma fila:

```
edad_salario_18_40[23, ]
```

```
# A tibble: 1 × 2
   edad salario_prom
  <int>        <dbl>
1    40     1080639.
```

Es el último punto del rango, el que más lejos está del promedio de las edades. Por eso pesa tanto en la pendiente.

DIAGNÓSTICO

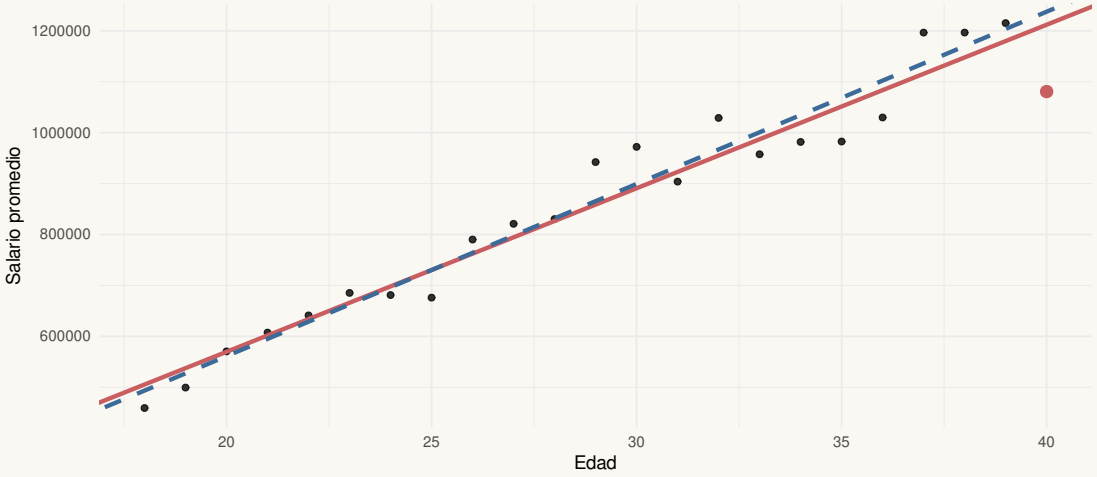
El punto 23

```
mod_sin_23 <- lm(
  formula = salario_prom ~ edad,
  data = edad_salario_18_40[-23, ]
)
mod_sin_23
```

```
Coefficients:
(Intercept)         edad
    -115161        33815
```

DIAGNÓSTICO

El punto 23



DIAGNÓSTICO

El punto 23

DIAGNÓSTICO

El punto 23

Sacando un solo punto de 23, la pendiente pasa de \$32.110 a \$33.815 y el R^2 de 0.941 a 0.958.

DIAGNÓSTICO

El punto 23

Sacando un solo punto de 23, la pendiente pasa de \$32.110 a \$33.815 y el R^2 de 0.941 a 0.958.

No es un dato erróneo que haya que tirar: es el borde del rango que elegimos mirar.
Pero conviene saber que la conclusión se apoya bastante en él.

DIAGNÓSTICO

`augment()`: métricas sobre los residuos

```
df_aug <- augment(mod, data = edad_salario_18_40)
```

DIAGNÓSTICO

`augment()`: métricas sobre los residuos

```
df_aug <- augment(mod, data = edad_salario_18_40)
head(df_aug)
```

edad	salario_prom	.fitted	.resid	.hat	.sigma	.cooksd	.std.resid
18	459300	505400	−46150	0.163	56090	0.07951	−0.9035
19	499400	537500	−38130	0.1423	56470	0.04511	−0.7374
20	570500	569600	831.3	0.1235	57220	1.782e−05	0.0159
21	607400	601800	5650	0.1067	57200	0.0006846	0.1071
22	641200	633900	7284	0.0919	57190	0.0009481	0.1369
23	685300	666000	19300	0.07905	57040	0.00557	0.3602

DIAGNÓSTICO CON GGPLOT

Residuos

```
ggplot(df_aug, aes(.fitted, .resid)) +
  geom_point() +
```

DIAGNÓSTICO CON GGPLOT

Residuos

```
ggplot(df_aug, aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed",
             colour = "grey", linewidth = 1.2) +
```

DIAGNÓSTICO CON GGPLOT

Residuos

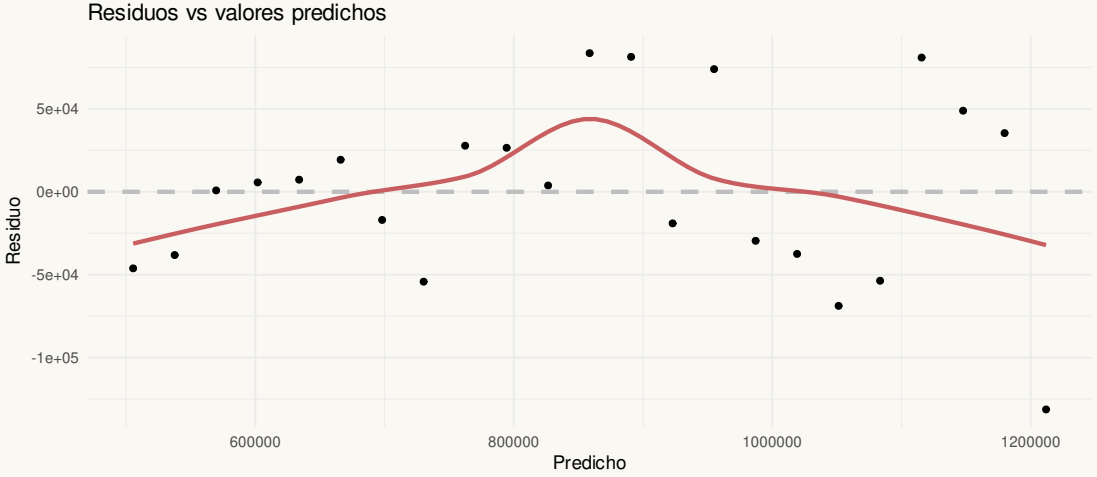
```
ggplot(df_aug, aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed",
             colour = "grey", linewidth = 1.2) +
  geom_smooth(se = FALSE, col = "#C95E61", linewidth = 1.2) +
```

Residuos

```
ggplot(df_aug, aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed",
             colour = "grey", linewidth = 1.2) +
  geom_smooth(se = FALSE, col = "#C95E61", linewidth = 1.2) +
  labs(title = "Residuos vs valores predichos",
       x = "Predicho", y = "Residuo") +
  theme_minimal()
```

DIAGNÓSTICO CON GGPLOT

Residuos



DIAGNÓSTICO CON GGPLOT

Residuos Q-Q

```
ggplot(df_aug, aes(sample = .std.resid)) +
  stat_qq() +
```

DIAGNÓSTICO CON GGPLOT

Residuos Q-Q

```
ggplot(df_aug, aes(sample = .std.resid)) +
  stat_qq() +
  geom_abline() +
```

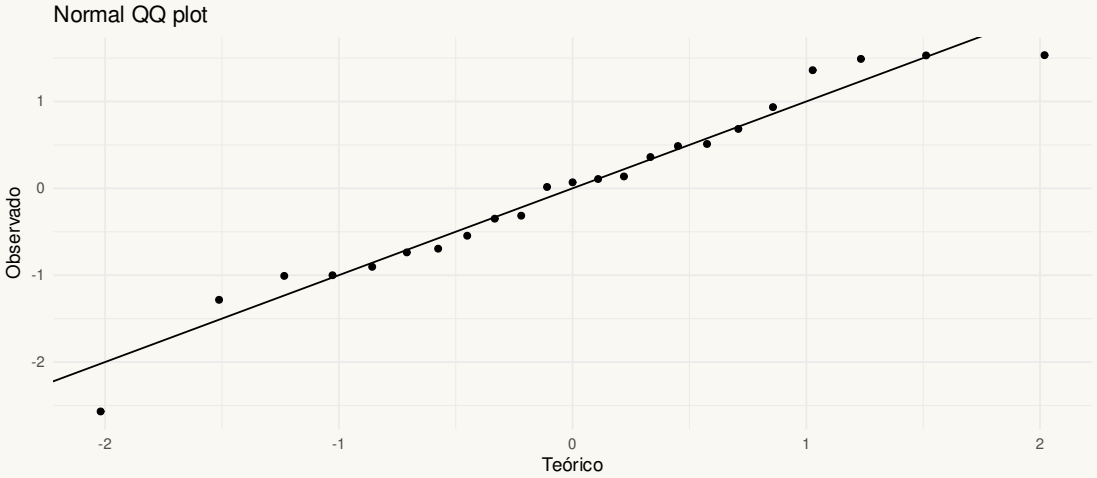
DIAGNÓSTICO CON GGPLOT

Residuos Q-Q

```
ggplot(df_aug, aes(sample = .std.resid)) +
  stat_qq() +
  geom_abline() +
  labs(title = "Normal QQ plot",
       x = "Teórico", y = "Observado") +
  theme_minimal()
```

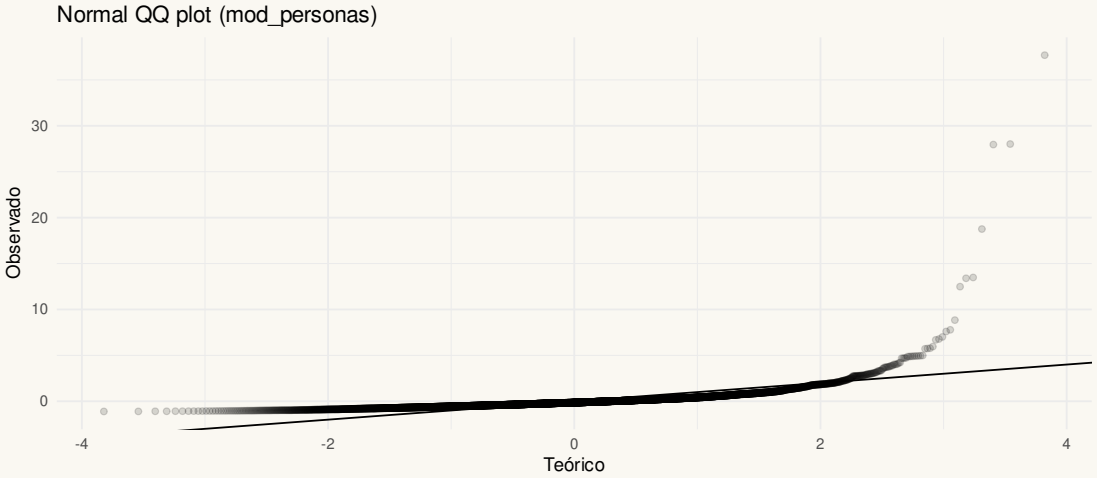
DIAGNÓSTICO CON GGPLOT

Residuos Q-Q



DIAGNÓSTICO CON GGPLOT

Residuos Q-Q: las personas



Pongamos en práctica lo que aprendimos sobre modelos lineales

Ejercicio

EJERCICIO

Linealización y ajuste exponencial

Para poner un poco en práctica las cosas que estuvimos viendo sobre regresiones lineales vamos a trabajar con datos reales sobre consumo energético de fuentes renovables.

EJERCICIO

Linealización y ajuste exponencial

Para poner un poco en práctica las cosas que estuvimos viendo sobre regresiones lineales vamos a trabajar con datos reales sobre consumo energético de fuentes renovables.

El objetivo del ejercicio va a ser visualizar los datos, ajustar un modelo exponencial a las observaciones, evaluar su validez y automatizar el análisis.

EJERCICIO

Consignas

EJERCICIO

Consignas

- Descarguen el dataset desde la página web de Our World in Data y levántenlo como un dataframe.

`https://ourworldindata.org/grapher/global-energy-substitution`

EJERCICIO

Consignas

- Descarguen el dataset desde la página web de Our World in Data y levántenlo como un dataframe.

`https://ourworldindata.org/grapher/global-energy-substitution`

- Exploren gráficamente la evolución del consumo energético para las fuentes renovables del dataset (Biofuels, Solar, Wind, Other renewables).

EJERCICIO

Consignas

- Descarguen el dataset desde la página web de Our World in Data y levántenlo como un dataframe.

`https://ourworldindata.org/grapher/global-energy-substitution`

- Exploren gráficamente la evolución del consumo energético para las fuentes renovables del dataset (Biofuels, Solar, Wind, Other renewables).
- Elijan una de esas fuentes y ajusten un modelo exponencial del tipo $E(t) = a \cdot e^{bt}$, donde $E(t)$ representa el consumo de la fuente elegida en función del tiempo.

EJERCICIO

Consignas

- Descarguen el dataset desde la página web de Our World in Data y levántenlo como un dataframe.

`https://ourworldindata.org/grapher/global-energy-substitution`

- Exploren gráficamente la evolución del consumo energético para las fuentes renovables del dataset (Biofuels, Solar, Wind, Other renewables).
- Elijan una de esas fuentes y ajusten un modelo exponencial del tipo $E(t) = a \cdot e^{bt}$, donde $E(t)$ representa el consumo de la fuente elegida en función del tiempo.
- Evalúen el ajuste gráficamente y con las métricas que vimos en clase. ¿Les parece razonable el modelo?

EJERCICIO

Consignas

- Descarguen el dataset desde la página web de Our World in Data y levántenlo como un dataframe.

`https://ourworldindata.org/grapher/global-energy-substitution`

- Exploren gráficamente la evolución del consumo energético para las fuentes renovables del dataset (Biofuels, Solar, Wind, Other renewables).
- Elijan una de esas fuentes y ajusten un modelo exponencial del tipo $E(t) = a \cdot e^{bt}$, donde $E(t)$ representa el consumo de la fuente elegida en función del tiempo.
- Evalúen el ajuste gráficamente y con las métricas que vimos en clase. ¿Les parece razonable el modelo?
- Generalicen el proceso para todas las fuentes: escriban una función que, dada una fuente, devuelva los coeficientes del ajuste y alguna métrica de calidad, y apliquen esa función al dataset completo usando alguna herramienta de programación funcional.

EJERCICIO

Consignas

- Descarguen el dataset desde la página web de Our World in Data y levántenlo como un dataframe.

`https://ourworldindata.org/grapher/global-energy-substitution`

- Exploren gráficamente la evolución del consumo energético para las fuentes renovables del dataset (Biofuels, Solar, Wind, Other renewables).
- Elijan una de esas fuentes y ajusten un modelo exponencial del tipo $E(t) = a \cdot e^{bt}$, donde $E(t)$ representa el consumo de la fuente elegida en función del tiempo.
- Evalúen el ajuste gráficamente y con las métricas que vimos en clase. ¿Les parece razonable el modelo?
- Generalicen el proceso para todas las fuentes: escriban una función que, dada una fuente, devuelva los coeficientes del ajuste y alguna métrica de calidad, y apliquen esa función al dataset completo usando alguna herramienta de programación funcional.

Nota: Este punteo no es más que una guía de los pasos que pueden seguir para analizar un dataset como este, pero siéntanse libres de calcular y graficar los estadísticos que le parezcan más interesantes.

Esto es todo.