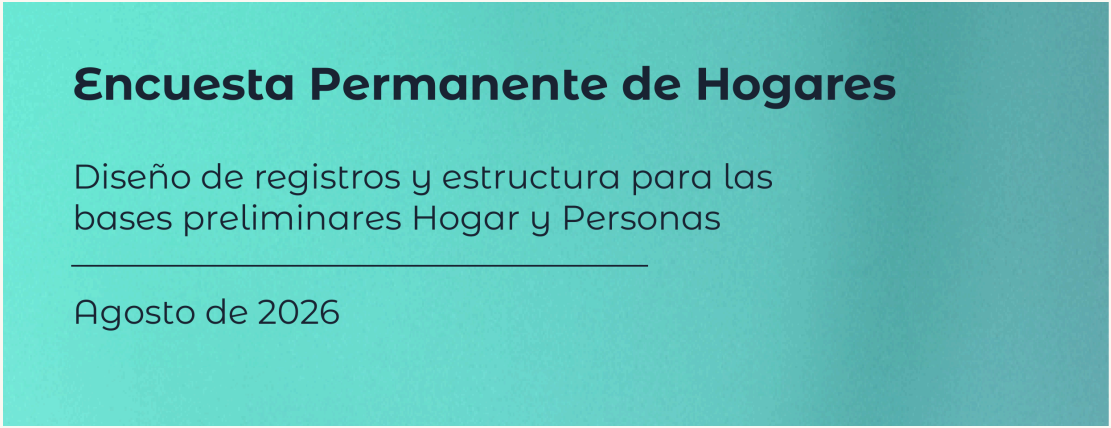
Hola

Enfoque Estadístico del Aprendizaje

Introducción al análisis de datos y los modelos lineales

Bernardo L. Español · 29 de Agosto

Encuesta Permanente de Hogares



https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos

Antes de empezar, una aclaración:

Antes de empezar, una aclaración:

Todas las conclusiones que se puedan llegar a desprender de este análisis deben interpretarse con cautela. Se trata únicamente de un ejercicio pedagógico cuyo propósito es ilustrar un posible flujo de trabajo con datos reales.

Ahora sí, **empecemos**

TRATAMIENTOS DE DATOS

Librerías

```
# para hacer todo lo que aprendieron la clase pasada
library(tidyverse)
```

Librerías

```r
# para hacer todo lo que aprendieron la clase pasada
library(tidyverse)

# para hacer todo lo que aprendieron la clase pasada,
# aún con objetos de R nativo
library(broom)
```

Lectura de datos

```
df <- read_delim(
  file = "usu_individual_T126.txt",
  delim = ";",
)
```

Lectura de datos

```
df <- read_delim(
  file = "usu_individual_T126.txt",
  delim = ";",
)
```

CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION
TQRMNORURHLLKUCDEOJAH00893620	2026	1	1	2	1	44
TQRMNORURHLLKUCDEOJAH00893620	2026	1	1	1	1	44
TQRMNORURHLLKUCDEOJAH00893620	2026	1	1	3	1	44
TQRMNORURHLLKUCDEOJAH00893620	2026	1	1	4	1	44
TQRMNOTUQHJOLOCDEHLEH00875560	2026	1	1	3	1	40
TQRMNOQUPHLOLPCDEHLEH00893322	2026	1	1	1	1	40

```
nrow(df)
```

```
[1] 43739
```

TRATAMIENTOS DE DATOS

Variables de interés

```
df <- df %>%
  rename(
    edad = CH06,
    salario = P21
    # ...
  )
```

Características de los miembros del hogar Cuestionario Hogar		
Campo	**Tipo (longitud)**	**Descripción**
CH03	N (2)	Relación de parentesco 1 = Jefe/a 2 = Cónyuge/pareja 3 = Hijo/a/hijastro/a 4 = Yerno/nuera 5 = Nieto/a 6 = Madre/padre 7 = Suegro/a 8 = Hermano/a 9 = Otros familiares 10 = No familiares
CH04	N (1)	Sexo 1 = Varón 2 = Mujer
CH05	date	Fecha de nacimiento (día, mes y año)
CH06	N (2)	¿Cuántos años cumplidos tiene?

Filtros

Nos quedamos con la gente que tiene algún trabajo (formal o informal)

```
df_filt <- df %>%
  filter(
    ESTADO == 1,
    salario > 0,
  )
```

Filtros

Nos quedamos con la gente que tiene algún trabajo (formal o informal)

```
df_filt <- df %>%
  filter(
    ESTADO == 1,
    salario > 0,
  )
```

Podemos ver cuánto filtramos comparando

```
nrow(df)
```

```
[1] 43739
```

```
nrow(df_filt)
```

```
[1] 15448
```

Gráficos

GRÁFICOS

GGPlot

```
ggplot(df_filt,                 # dataframe con los datos
  aes(x = edad, y = salario)) + # vars del df quiero "mapear"
```

GRÁFICOS

GGPlot

```
ggplot(df_filt,                 # dataframe con los datos
  aes(x = edad, y = salario)) + # vars del df quiero "mapear"
  geom_point(alpha = 0.2) +     # representa las vars con puntos
```

GRÁFICOS

GGPlot

```
ggplot(df_filt,                  # dataframe con los datos
  aes(x = edad, y = salario)) + # vars del df quiero "mapear"
  geom_point(alpha = 0.2) +     # representa las vars con puntos
  labs(                         # modificá los labels
    x = "Edad",
    y = "Salario",
  ) +
```

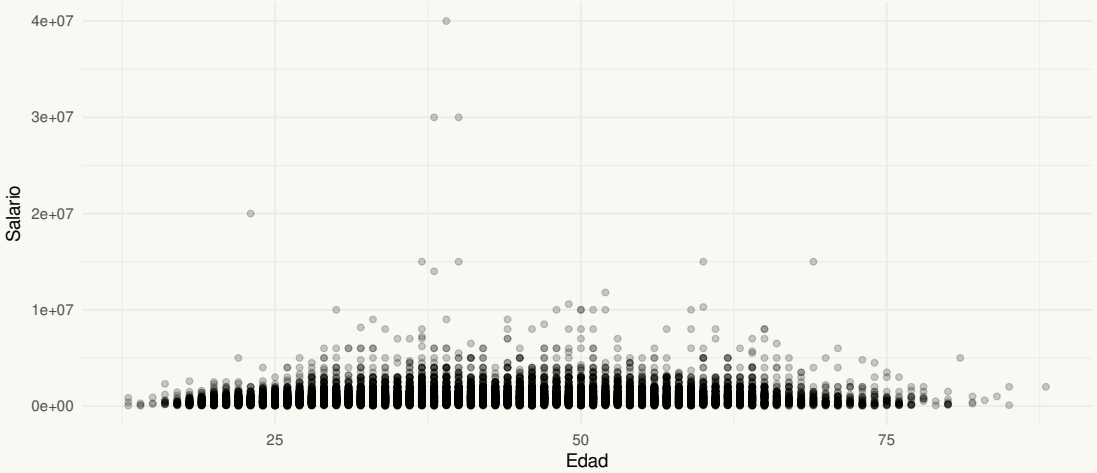
GRÁFICOS

GGPlot

```
ggplot(df_filt,                  # dataframe con los datos
  aes(x = edad, y = salario)) + # vars del df quiero "mapear"
  geom_point(alpha = 0.2) +     # representa las vars con puntos
  labs(                         # modificá los labels
    x = "Edad",
    y = "Salario",
  ) +
  theme_minimal()               # que sea más "limpio" (lindo)
```

GRÁFICOS

Edad vs Salario



EXPLORACIÓN GRÁFICA

Salarios menores a 5M

```
ggplot(df_filt,                  # dataframe con los datos
  aes(x = edad, y = salario)) + # vars del df quiero "mapear"
  geom_point(alpha = 0.2) +     # representa las vars con puntos
  labs(                         # modificá los labels
    x = "Edad",
    y = "Salario",
  ) +
  ylim(0, 5000000) +            # solo salarios menores a 5M
  theme_minimal()               # que sea más "limpio" (lindo)
```

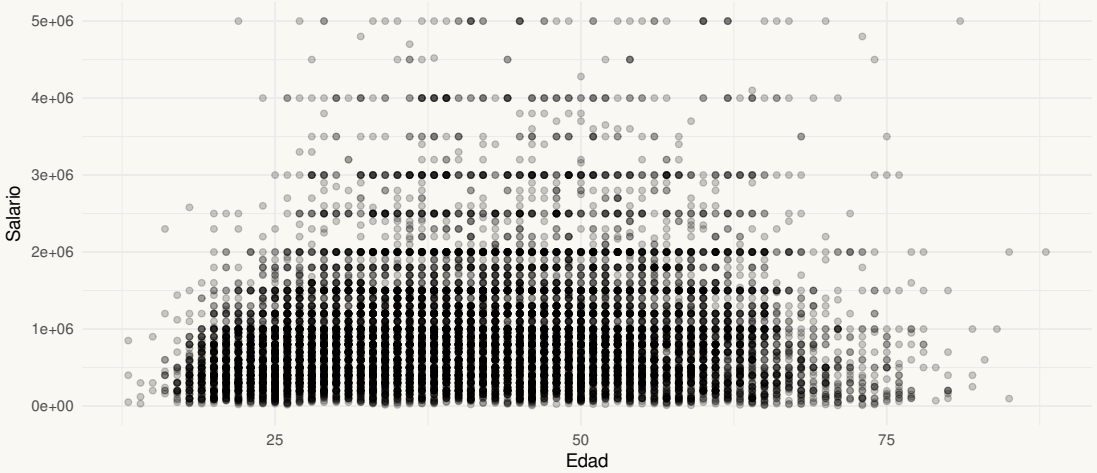
EXPLORACIÓN GRÁFICA

Salarios menores a 5M

```r
ggplot(df_filt,                  # dataframe con los datos
  aes(x = edad, y = salario)) + # vars del df quiero "mapear"
  geom_point(alpha = 0.2) +     # representa las vars con puntos
  labs(                         # modificá los labels
    x = "Edad",
    y = "Salario",
  ) +
  ylim(0, 5000000) +            # solo salarios menores a 5M
  theme_minimal()               # que sea más "limpio" (lindo)
```

EXPLORACIÓN GRÁFICA

Salarios menores a 5M



EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```
edad_salario <- df_filt %>%
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```
edad_salario <- df_filt %>%
  # Agrupo todas las filas por edades
  group_by(edad) %>%
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```r
edad_salario <- df_filt %>%
  # Agrupo todas las filas por edades
  group_by(edad) %>%

  # Calculo el salario promedio de cada grupo
  summarise(salario_prom = mean(salario)) %>%
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```
edad_salario <- df_filt %>%
  # Agrupo todas las filas por edades
  group_by(edad) %>%

  # Calculo el salario promedio de cada grupo
  summarise(salario_prom = mean(salario)) %>%

  # Desarmo el grupo
  ungroup() %>%
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```
edad_salario <- df_filt %>%
  # Agrupo todas las filas por edades
  group_by(edad) %>%

  # Calculo el salario promedio de cada grupo
  summarise(salario_prom = mean(salario)) %>%

  # Desarmo el grupo
  ungroup() %>%

  # Ordeno el df por edad de forma creciente
  arrange(edad)
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```
# Calculo el salario promedio por edad
edad_salario <- df_filt %>%
  group_by(edad) %>%
  summarise(salario_prom = mean(salario)) %>%
  ungroup() %>%
  arrange(edad)
```

EN BUSCA DE LA CORRELACIÓN

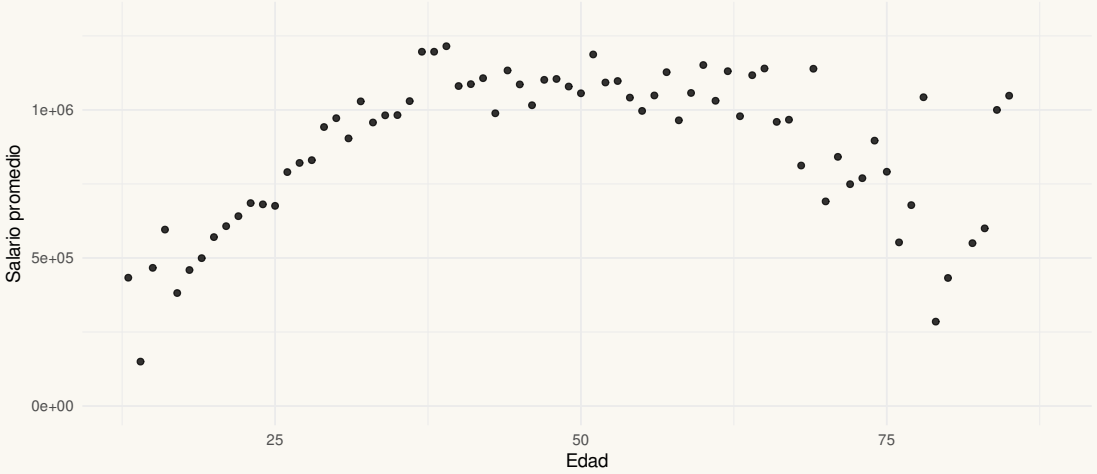
Edad vs salario promedio

Graficamos igual que antes:

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

EN BUSCA DE LA CORRELACIÓN

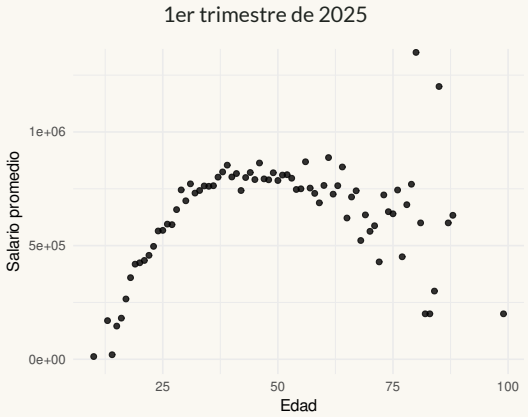
Edad vs salario promedio

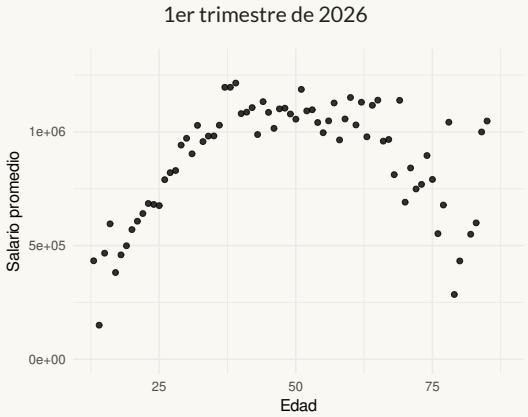


EN BUSCA DE LA CORRELACIÓN

2025 vs 2026

(Como curiosidad, el mismo gráfico con la base del año pasado)





EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

Filtrando por edad nos queda…

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

Filtrando por edad nos queda...

```
# Personas entre 18 y 40 años que sean asalariadas
df_filt_18_40 <- df %>%
  filter(
    salario > 0,
    ESTADO == 1,
    between(edad, 18, 40)
  )
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

Filtrando por edad nos queda...

```
# Personas entre 18 y 40 años que sean asalariadas
df_filt_18_40 <- df %>%
  filter(
    salario > 0,
    ESTADO == 1,
    between(edad, 18, 40)
  )

# Promedio de su salario por edad
edad_salario_18_40 <- df_filt_18_40 %>%
  group_by(edad) %>%
  summarise(salario_prom = mean(salario)) %>%
  ungroup() %>%
  arrange(edad)
```

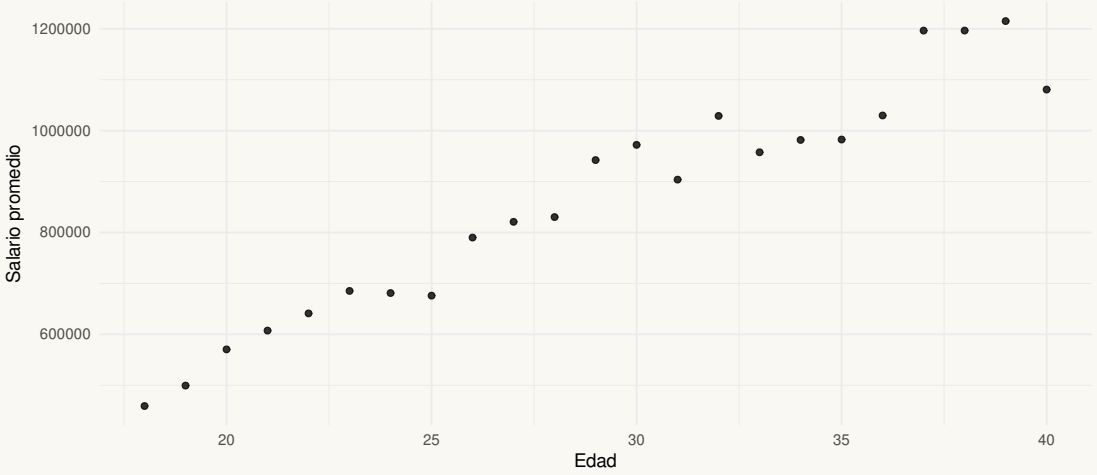
EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio

```
# Grafico el salario promedio vs la edad
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  labs(
    x = "Edad",
    y = "Salario promedio",
  ) +
  theme_minimal()
```

EN BUSCA DE LA CORRELACIÓN

Edad vs salario promedio



Modelo lineal simple

Introducción

Viendo los datos resulta razonable proponer un modelo lineal:

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

donde β_i son los parámetros de nuestro modelo y ε el error.

Introducción

Viendo los datos resulta razonable proponer un modelo lineal:

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

donde β_i son los parámetros de nuestro modelo y ε el error.

Otra forma de decir lo mismo es que Y es una variable aleatoria tal que su esperanza condicionada a X está data por

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$$

Introducción

Viendo los datos resulta razonable proponer un modelo lineal:

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

donde β_i son los parámetros de nuestro modelo y ε el error.

Otra forma de decir lo mismo es que Y es una variable aleatoria tal que su esperanza condicionada a X está data por

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$$

En cualquier caso, dados pares de datos (x_i, y_i) , los *mejores* estimadores para β_0 y β_1 que podemos construir son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$$

A mano

En nuestro código esta cuenta:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$$

se escribe

A mano

En nuestro código esta cuenta:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$$

se escribe

```
x_i <- edad_salario_18_40$edad         # [18, 19, ..., 40]
y_i <- edad_salario_18_40$salario_prom # [460K, 500K, ..., 1.1M]
```

A mano

En nuestro código esta cuenta:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$$

se escribe

```
x_i <- edad_salario_18_40$edad         # [18, 19, ..., 40]
y_i <- edad_salario_18_40$salario_prom # [460K, 500K, ..., 1.1M]
x_prom <- mean(x_i)
y_prom <- mean(y_i)
```

A mano

En nuestro código esta cuenta:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$$

se escribe

```
x_i <- edad_salario_18_40$edad         # [18, 19, ..., 40]
y_i <- edad_salario_18_40$salario_prom # [460K, 500K, ..., 1.1M]
x_prom <- mean(x_i)
y_prom <- mean(y_i)
b1 <- sum((x_i - x_prom) * (y_i - y_prom))/sum((x_i - x_prom)^2)
b0 <- y_prom - b1 * x_prom
```

A mano

En nuestro código esta cuenta:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}$$

se escribe

```
x_i <- edad_salario_18_40$edad          # [18, 19, ..., 40]
y_i <- edad_salario_18_40$salario_prom # [460K, 500K, ..., 1.1M]
x_prom <- mean(x_i)
y_prom <- mean(y_i)
b1 <- sum((x_i - x_prom) * (y_i - y_prom))/sum((x_i - x_prom)^2)
b0 <- y_prom - b1 * x_prom
cat(b0, b1)
```

```
>>> -72557.64 32110.37
```

En R

Podemos pedirle R *que haga la cuenta por nosotros*

En R

Podemos pedirle R *que haga la cuenta por nosotros*

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

En R

Podemos pedirle R *que haga la cuenta por nosotros*

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

```
Coefficients:
(Intercept)         edad
     -72558        32110
```

En R

Podemos pedirle R *que haga la cuenta por nosotros*

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

```
Coefficients:
(Intercept)         edad
     -72558        32110
```

```
b0 = mod$coefficients[1]
b1 = mod$coefficients[2]
cat(b0, b1)
```

```
>>> -72557.64 32110.37
```

MODELO LINEAL SIMPLE

Datos + LM

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

MODELO LINEAL SIMPLE

Datos + LM

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  labs(
    x = "Edad",
    y = "Salario promedio",
  ) +
  geom_function(fun = \(x) b0 + b1*x, # Agregamos una recta
    linewidth = 1.5, col = "#C95E61"  # b0 + b1*edad
  ) +
```

MODELO LINEAL SIMPLE

Datos + LM

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  labs(
    x = "Edad",
    y = "Salario promedio",
  ) +
  geom_function(fun = \(x) b0 + b1*x, # Agregamos una recta
    linewidth = 1.5, col = "#C95E61"  # b0 + b1*edad
  ) +
  theme_minimal()
```

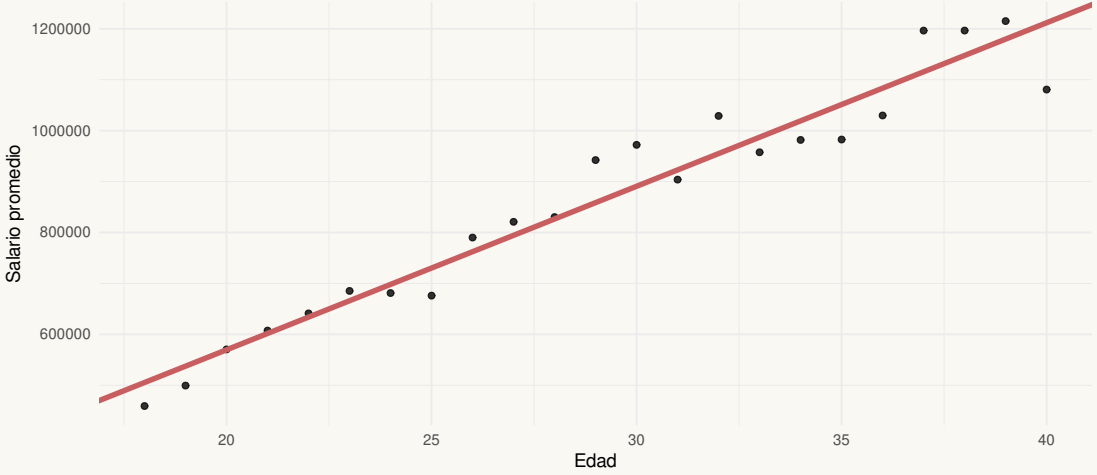
MODELO LINEAL SIMPLE

Datos + LM

```
ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  labs(
    x = "Edad",
    y = "Salario promedio",
  ) +
  geom_abline(intercept = b0, slope = b1,  # También se puede
    linewidth = 1.5, col = "#C95E61"         # hacer así
  ) +
  theme_minimal()
```

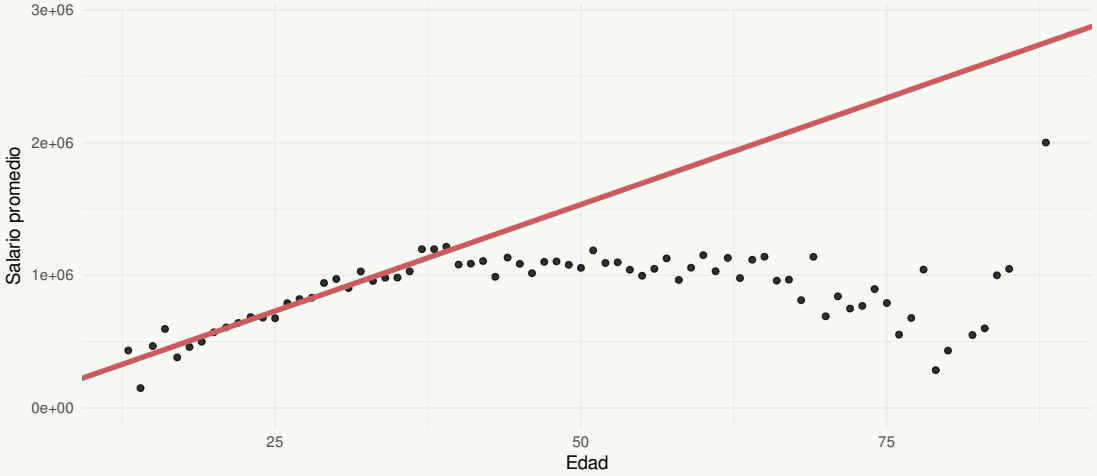
MODELO LINEAL SIMPLE

Datos + LM



MODELO LINEAL SIMPLE

Datos + LM



Predicciones

Predicciones

Recordemos que

$$\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X$$

Predicciones

Recordemos que

$$\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X$$

Lo que quiere decir que el modelo me puede decir cuánto predice que es el *sueldo promedio* Y para una persona de determinada de *edad* X:

$$\mathrm{E}(\text{sueldo}|\text{edad}) = -72558 + 32110\ \text{edad}$$

Predicciones

Recordemos que

$$\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X$$

Lo que quiere decir que el modelo me puede decir cuánto predice que es el *sueldo promedio* Y para una persona de determinada de *edad* X:

$$\mathrm{E}(\text{sueldo}|\text{edad}) = -72558 + 32110\ \text{edad}$$

ej: $\mathrm{E}(\text{sueldo} \mid 35) = -72558 + 32110 \cdot 35 = \$1051305$

Predicciones (y limitaciones)

```
library(glue)

glue("Alguien con 35 años gana en promedio ${round(b0 + 35 * b1)}")
glue("Alguien recién nacido gana en promedio ${round(b0 + 0 * b1)}")
glue("Alguien con 90 años gana en promedio ${round(b0 + 90 * b1)}")
```

Predicciones (y limitaciones)

```
library(glue)

glue("Alguien con 35 años gana en promedio ${round(b0 + 35 * b1)}")
glue("Alguien recién nacido gana en promedio ${round(b0 + 0 * b1)}")
glue("Alguien con 90 años gana en promedio ${round(b0 + 90 * b1)}")
```

```
Alguien con 35 años gana en promedio $1051305
```

Predicciones (y limitaciones)

```
library(glue)

glue("Alguien con 35 años gana en promedio ${round(b0 + 35 * b1)}")
glue("Alguien recién nacido gana en promedio ${round(b0 + 0 * b1)}")
glue("Alguien con 90 años gana en promedio ${round(b0 + 90 * b1)}")
```

```
Alguien con 35 años gana en promedio $1051305
Alguien recién nacido gana en promedio $-72558
```

Predicciones (y limitaciones)

```
library(glue)

glue("Alguien con 35 años gana en promedio ${round(b0 + 35 * b1)}")
glue("Alguien recién nacido gana en promedio ${round(b0 + 0 * b1)}")
glue("Alguien con 90 años gana en promedio ${round(b0 + 90 * b1)}")
```

```
Alguien con 35 años gana en promedio $1051305
Alguien recién nacido gana en promedio $-72558
Alguien con 90 años gana en promedio $2817376
```

MODELO LINEAL SIMPLE

¿Qué significa β_1?

A veces no nos interesa ajustar un modelo para poder predecir un valor que no tenemos, sino para poder entender una tendencia:

¿Qué significa β_1?

A veces no nos interesa ajustar un modelo para poder predecir un valor que no tenemos, sino para poder entender una tendencia:

Entre dos edades que se llevan **un año**, el salario promedio difiere en **$32.110**

¿Qué significa β_1?

A veces no nos interesa ajustar un modelo para poder predecir un valor que no tenemos, sino para poder entender una tendencia:

Entre dos edades que se llevan **un año**, el salario promedio difiere en **$32.110**

Cuidado que esto *no* quiere decir que una persona gane $32.110 más cuando cumple un año: la EPH es una foto de un trimestre, estamos comparando personas distintas, no siguiendo a gente en el tiempo (lamentablemente no tenemos esa *data*).

MODELO LINEAL SIMPLE

¿Qué significa β_1?

A veces no nos interesa ajustar un modelo para poder predecir un valor que no tenemos, sino para poder entender una tendencia:

Entre dos edades que se llevan **un año**, el salario promedio difiere en **$32.110**

Cuidado que esto *no* quiere decir que una persona gane $32.110 más cuando cumple un año: la EPH es una foto de un trimestre, estamos comparando personas distintas, no siguiendo a gente en el tiempo (lamentablemente no tenemos esa *data*).

Ah, y estas afirmaciones son válidas solo en el rango entre los 18 y los 40 años de edad, que es el único tramo que el modelo vio.

Promedios vs individuos

PROMEDIOS VS INDIVIDUOS

El modelo lineal

PROMEDIOS VS INDIVIDUOS

El modelo lineal

Nuestro `lm` nunca vio personas individuales: cada fila que le dimos era el promedio de una edad.

PROMEDIOS VS INDIVIDUOS

El modelo lineal

Nuestro `lm` nunca vio personas individuales: cada fila que le dimos era el promedio de una edad.
En particular

```
nrow(edad_salario_18_40)
```

```
[1] 23
```

```
nrow(df_filt_18_40)
```

```
[1] 7534
```

El modelo lineal

Si nos interesase podemos correr el mismo modelo sin promediar:

```
mod_personas <- lm(
  formula = salario ~ edad,
  data = df_filt_18_40    # una fila por persona
)
mod_personas
```

```
Coefficients:
(Intercept)         edad
     -32377        30832
```

El modelo lineal

Si nos interesase podemos correr el mismo modelo sin promediar:

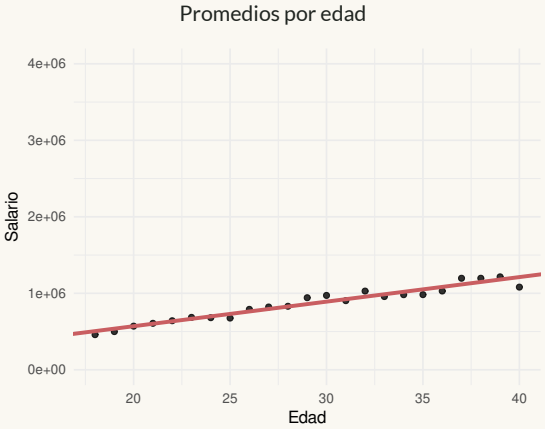
```
mod_personas <- lm(
  formula = salario ~ edad,
  data = df_filt_18_40    # una fila por persona
)
mod_personas
```

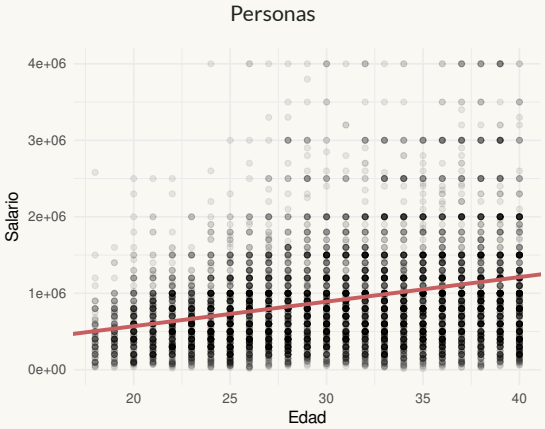
```
Coefficients:
(Intercept)         edad
     -32377        30832
```

La pendiente es casi la misma que la que sacamos de los promedios: 30.832 contra 32.110.

PROMEDIOS VS INDIVIDUOS

El modelo lineal





PROMEDIOS VS INDIVIDUOS

Variabilidad en los datos

La predicción del sueldo promedio de una persona de 35 años era $1051305.

Comparemos esto con los datos explícitamente.

PROMEDIOS VS INDIVIDUOS

Variabilidad en los datos

La predicción del sueldo promedio de una persona de 35 años era $1051305.

Comparemos esto con los datos explícitamente.

```
df_filt %>%
  filter(edad == 35) %>%
  summarise(media = mean(salario),
            desvio = sd(salario),
            p25 = quantile(salario, 0.25),
            p75 = quantile(salario, 0.75))
```

PROMEDIOS VS INDIVIDUOS

Variabilidad en los datos

La predicción del sueldo promedio de una persona de 35 años era $1051305.

Comparemos esto con los datos explícitamente.

```
df_filt %>%
  filter(edad == 35) %>%
  summarise(media = mean(salario),
            desvio = sd(salario),
            p25 = quantile(salario, 0.25),
            p75 = quantile(salario, 0.75))
```

```
     media   desvio   p25     p75
1 982530.3 722841.4 5e+05 1300000
```

PROMEDIOS VS INDIVIDUOS

Variabilidad en los datos

La predicción del sueldo promedio de una persona de 35 años era $1051305.

Comparemos esto con los datos explícitamente.

```
df_filt %>%
  filter(edad == 35) %>%
  summarise(media = mean(salario),
            desvio = sd(salario),
            p25 = quantile(salario, 0.25),
            p75 = quantile(salario, 0.75))
```

```
     media   desvio   p25     p75
1 982530.3 722841.4 5e+05 1300000
```

La media real de esos 413 sueldos es $982.530.

El desvío es de $722.841.

La mitad de los sueldos cae en el intervalo [$500.000, $1.300.000].

Como es de esperar, entre personas de la misma edad hay muchísima variación.

Extendiendo el análisis

EXTENDIENDO EL ANÁLISIS

Otra variable: horas trabajadas

EXTENDIENDO EL ANÁLISIS

Otra variable: horas trabajadas

`PP3E_TOT` son las horas que la persona trabajó en la semana

```
salario_horas <- df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
```

EXTENDIENDO EL ANÁLISIS

Otra variable: horas trabajadas

`PP3E_TOT` son las horas que la persona trabajó en la semana

```
salario_horas <- df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
  group_by(horas = PP3E_TOT) %>%
  filter(n() >= 30) %>%           # solo horas con muchos casos
```

EXTENDIENDO EL ANÁLISIS

Otra variable: horas trabajadas

`PP3E_TOT` son las horas que la persona trabajó en la semana

```
salario_horas <- df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
  group_by(horas = PP3E_TOT) %>%
  filter(n() >= 30) %>%           # solo horas con muchos casos
  summarise(salario_prom = mean(salario)) %>%
  ungroup()
```

EXTENDIENDO EL ANÁLISIS

Otra variable: horas trabajadas

`PP3E_TOT` son las horas que la persona trabajó en la semana

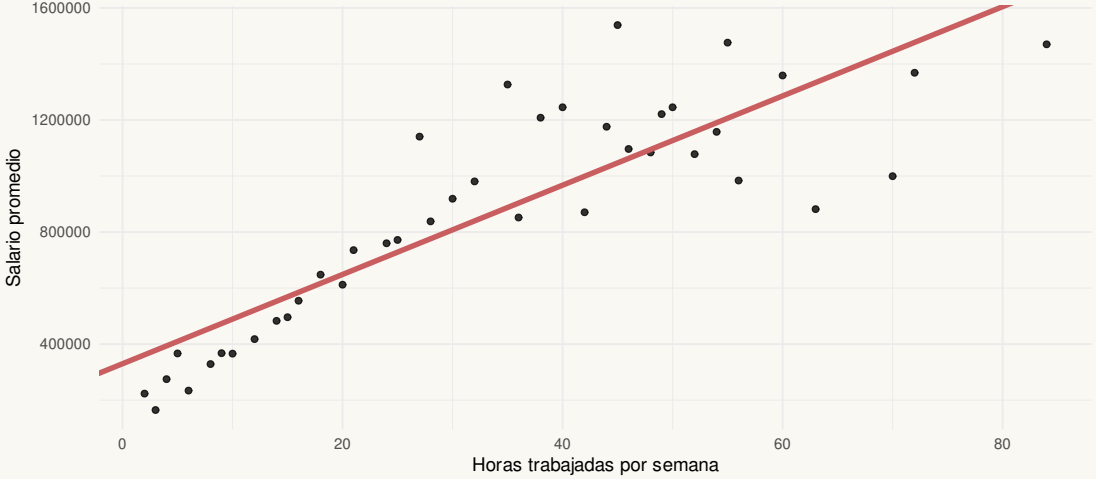
```
salario_horas <- df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
  group_by(horas = PP3E_TOT) %>%
  filter(n() >= 30) %>%           # solo horas con muchos casos
  summarise(salario_prom = mean(salario)) %>%
  ungroup()

mod_horas <- lm(salario_prom ~ horas, data = salario_horas)
mod_horas
```

```
Coefficients:
(Intercept)        horas
     329954        15932
```

EXTENDIENDO EL ANÁLISIS

Horas trabajadas



EXTENDIENDO EL ANÁLISIS

Horas trabajadas

EXTENDIENDO EL ANÁLISIS

Horas trabajadas

"Las personas que trabajan **una hora semanal más** ganan en promedio unos **$16.000 más por mes**."

EXTENDIENDO EL ANÁLISIS

Horas trabajadas

"Las personas que trabajan **una hora semanal más** ganan en promedio unos **$16.000 más por mes**."

Nuevamente, eso **no** es lo que le pasaría a alguien si trabajara una hora más: estamos comparando personas distintas, y quien trabaja más horas también difiere en oficio, en formalidad, en sexo. Toda esa variabilidad queda promediada en la pendiente; más adelante vamos a ver cómo separar estas variables.

EXTENDIENDO EL ANÁLISIS

Varones y mujeres

```
salario_sexo <- df_filt %>%
  filter(between(edad, 18, 55)) %>%
  mutate(sexo = ifelse(CH04 == 1, "Varones", "Mujeres")) %>%
  group_by(sexo, edad) %>%
  summarise(salario_prom = mean(salario))
```

EXTENDIENDO EL ANÁLISIS

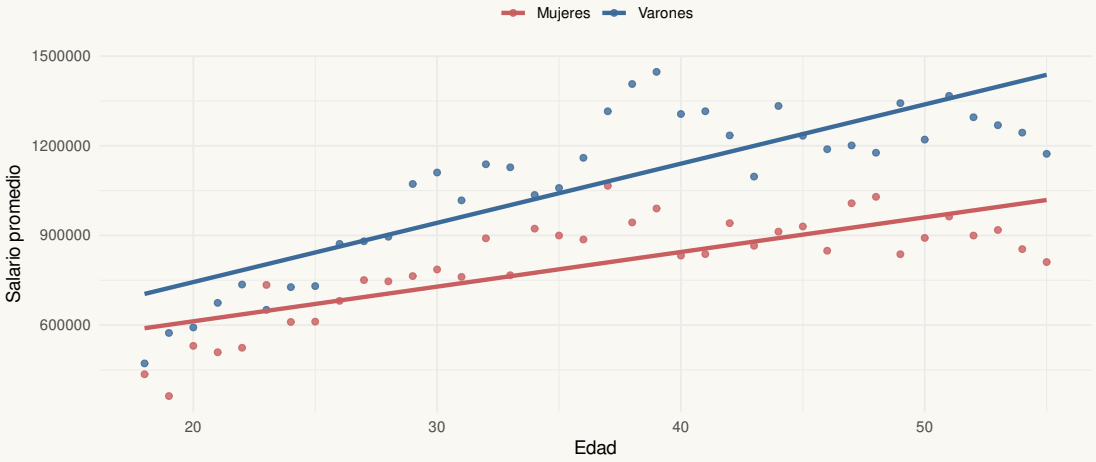
Varones y mujeres

```
salario_sexo <- df_filt %>%
  filter(between(edad, 18, 55)) %>%
  mutate(sexo = ifelse(CH04 == 1, "Varones", "Mujeres")) %>%
  group_by(sexo, edad) %>%
  summarise(salario_prom = mean(salario))

ggplot(salario_sexo, aes(x = edad, y = salario_prom, col = sexo)) +
  geom_point(alpha = 0.8) +
  geom_smooth(method = "lm", se = FALSE) +   # una recta por grupo
  theme_minimal()
```

EXTENDIENDO EL ANÁLISIS

Varones y mujeres



EXTENDIENDO EL ANÁLISIS

Varones y mujeres

```
salario_sexo %>%
  group_by(sexo) %>%
  group_modify(~ tidy(lm(salario_prom ~ edad, data = .x))) %>%
  select(sexo, term, estimate)
```

```
     sexo        term  estimate
1 Mujeres (Intercept) 380709.38
2 Mujeres        edad  11597.53
3 Varones (Intercept) 347509.67
4 Varones        edad  19818.74
```

EXTENDIENDO EL ANÁLISIS

Varones y mujeres

```
salario_sexo %>%
  group_by(sexo) %>%
  group_modify(~ tidy(lm(salario_prom ~ edad, data = .x))) %>%
  select(sexo, term, estimate)
```

```
     sexo        term  estimate
1 Mujeres (Intercept) 380709.38
2 Mujeres        edad  11597.53
3 Varones (Intercept) 347509.67
4 Varones        edad  19818.74
```

Arrancan casi igual, pero el sueldo de los varones crece **casi al doble de velocidad** por cada año de edad.

EXTENDIENDO EL ANÁLISIS

Varones y mujeres: salario por hora

```
df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
  mutate(sexo = ifelse(CH04 == 1, "Varones", "Mujeres"),
         por_hora = salario / (PP3E_TOT * 365/7/12)) %>%
  group_by(sexo) %>%
  summarise(salario = mean(salario),
            horas = mean(PP3E_TOT),
            por_hora = mean(por_hora))
```

```
     sexo salario  horas por_hora
1 Mujeres  818574 32.063   6540.1
2 Varones 1113068 40.142   6967.3
```

EXTENDIENDO EL ANÁLISIS

Varones y mujeres: salario por hora

```
df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
  mutate(sexo = ifelse(CH04 == 1, "Varones", "Mujeres"),
         por_hora = salario / (PP3E_TOT * 365/7/12)) %>%
  group_by(sexo) %>%
  summarise(salario = mean(salario),
            horas = mean(PP3E_TOT),
            por_hora = mean(por_hora))
```

```
     sexo salario  horas por_hora
1 Mujeres  818574 32.063   6540.1
2 Varones 1113068 40.142   6967.3
```

Por mes la brecha es de **26,5%** y por hora de trabajo **6,1%**.

EXTENDIENDO EL ANÁLISIS

Varones y mujeres: salario por hora

```
df_filt %>%
  filter(PP3E_TOT > 0, PP3E_TOT < 100) %>%
  mutate(sexo = ifelse(CH04 == 1, "Varones", "Mujeres"),
         por_hora = salario / (PP3E_TOT * 365/7/12)) %>%
  group_by(sexo) %>%
  summarise(salario = mean(salario),
            horas = mean(PP3E_TOT),
            por_hora = mean(por_hora))
```

```
     sexo salario  horas por_hora
1 Mujeres  818574 32.063   6540.1
2 Varones 1113068 40.142   6967.3
```

Por mes la brecha es de **26,5%** y por hora de trabajo **6,1%**.

La brecha mensual mezcla dos cosas distintas: cuánto se cobra por hora y cuántas horas se trabajan.

Elegir la variable de respuesta ya es elegir qué pregunta contestamos.

Esto es todo.