Unidad 4

Modelamiento estadístico

Iris Ashimine

Relaciones no lineales

Las funciones de regresión no lineales son aquellas en las que la relación entre la variable dependiente y y las variables independientes $x_1, x_2, ..x_k$ no es una combinación lineal, es decir, el efecto de un cambio en la variable x_1 no es una constante, sino que depende de otra variable.

La relación puede involucrar **polinomios, logaritmos e interacciones**. Estas regresiones se utilizan cuando se sabe que el comportamiento de la variable dependiente sigue una forma **no lineal** en lugar de una línea recta.

Logarítmica

Las regresiones con logarítmos tienen dos funciones: - transformar variables en modelos de regresión cuando su relación no es lineal

$$y_i = e^{\beta_0+\beta_1 x_{1i}+u_i} log(y_i) = log(e^{\beta_0+\beta_1 x_{1i}+u_i}) log(y_i) = \beta_0 + \beta_1 x_{1i} + u_i$$

- Ajustar los datos cuando tienen **distribuciones sesgadas**

Transformar datos a logaritmos

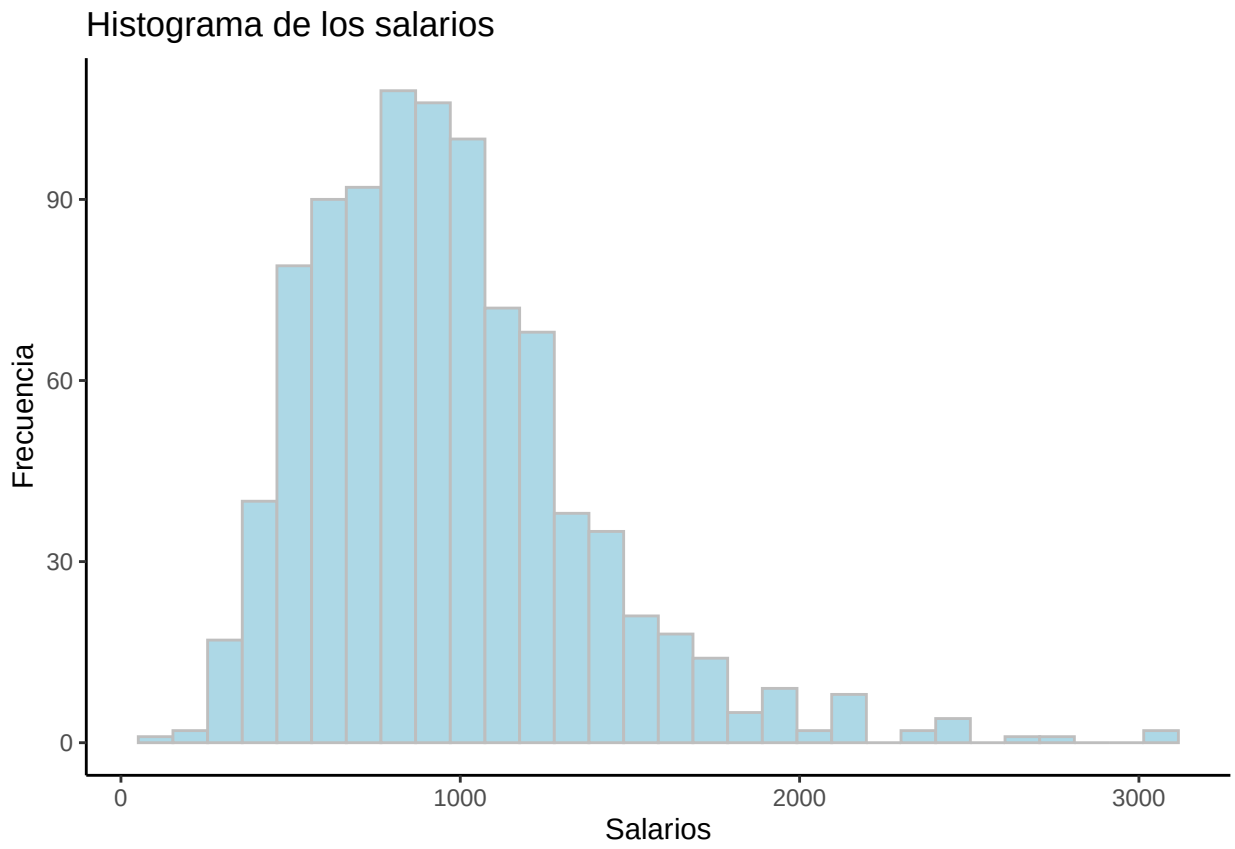
Cuando las variables tiene distribuciones con colas largas o asimétricas, se los transforma en logaritmos, debido a que estos **comprimen los valores grandes y acentúan los pequeños**, lo que ayuda a normalizar la distribución de los datos.

- Las variables económicas, como ingresos, precios o tamaños de empresas, suelen tener distribuciones sesgadas a la derecah (colas largas). Aplicar una transformación logarítmica hace que su distribución se aproxime a una **distribución normal**.
- La transformación logarítmica también ayuda a **estabilizar la varianza** por tanto, *puede* contribuir a reducir la heterocedasticidad.
- Los coeficientes de regresiones logarítmicas también facilitan la interpretación porque se pueden leer como **elasticidades** o **cambios porcentuales**.

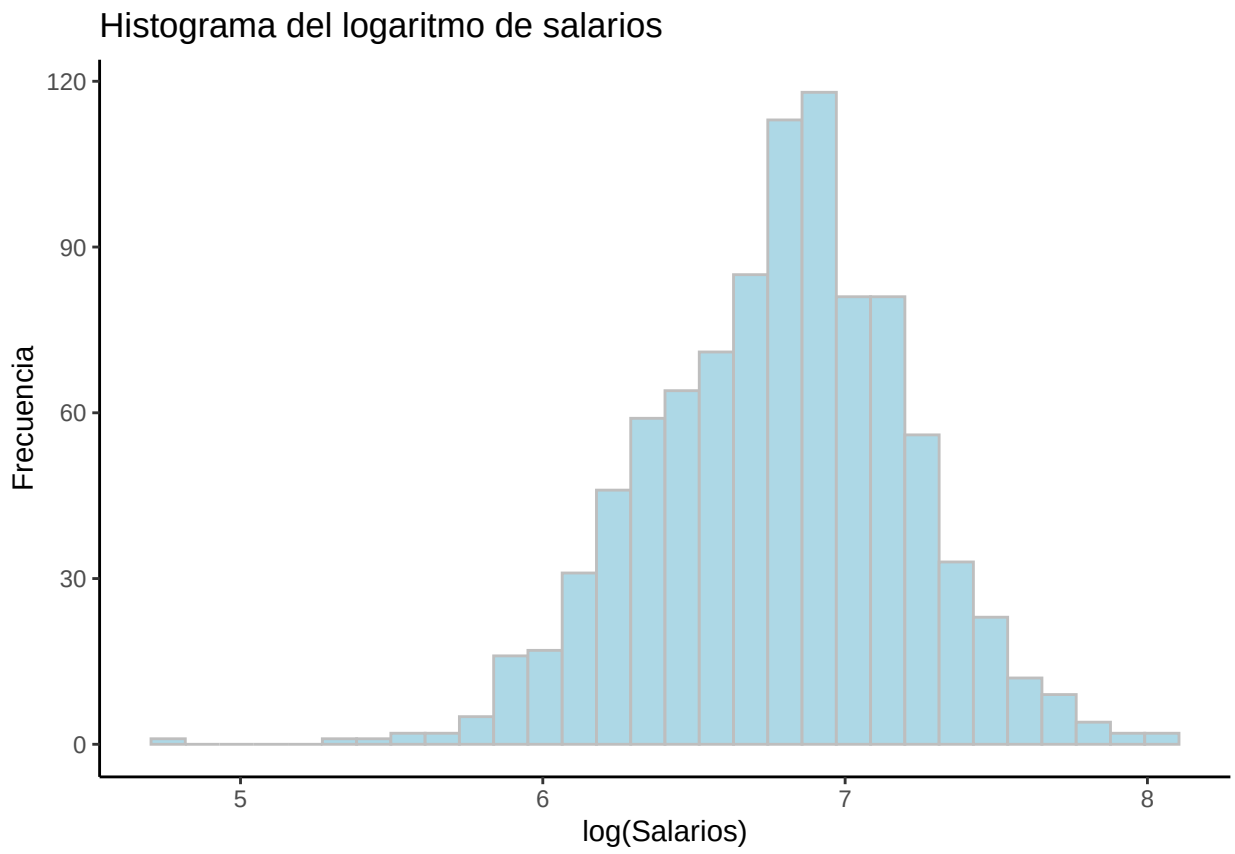
Ejemplos reales

1. **Los ingresos** suelen tener una distribución muy sesgada, ya que la mayoría de las personas ganan un salario relativamente bajo, mientras un pequeño número de personas gana salarios extremadamente altos.

```
ggplot(data = wage2, aes(x = wage)) +
  geom_histogram(bins=30, color = "gray", fill = "lightblue") +
  theme_classic() +
  labs(title = "Histograma de los salarios", x = "Salarios", y = "Frecuencia")
```

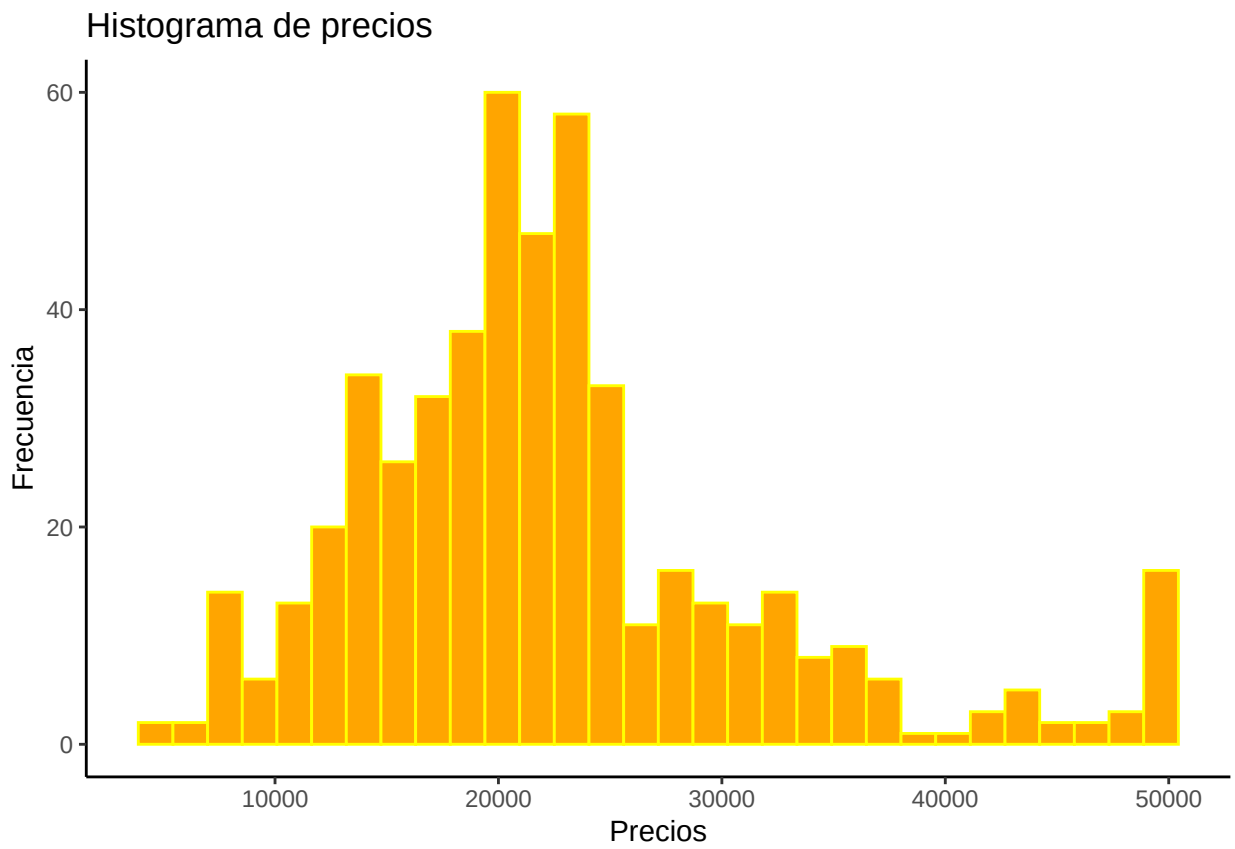



```
ggplot(data = wage2, aes(x = lwage)) +
  geom_histogram(bins=30, color = "gray", fill = "lightblue") +
  theme_classic() +
  labs(title = "Histograma del logaritmo de salarios", x = "log(Salarios)", y = "Frecuencia")
```

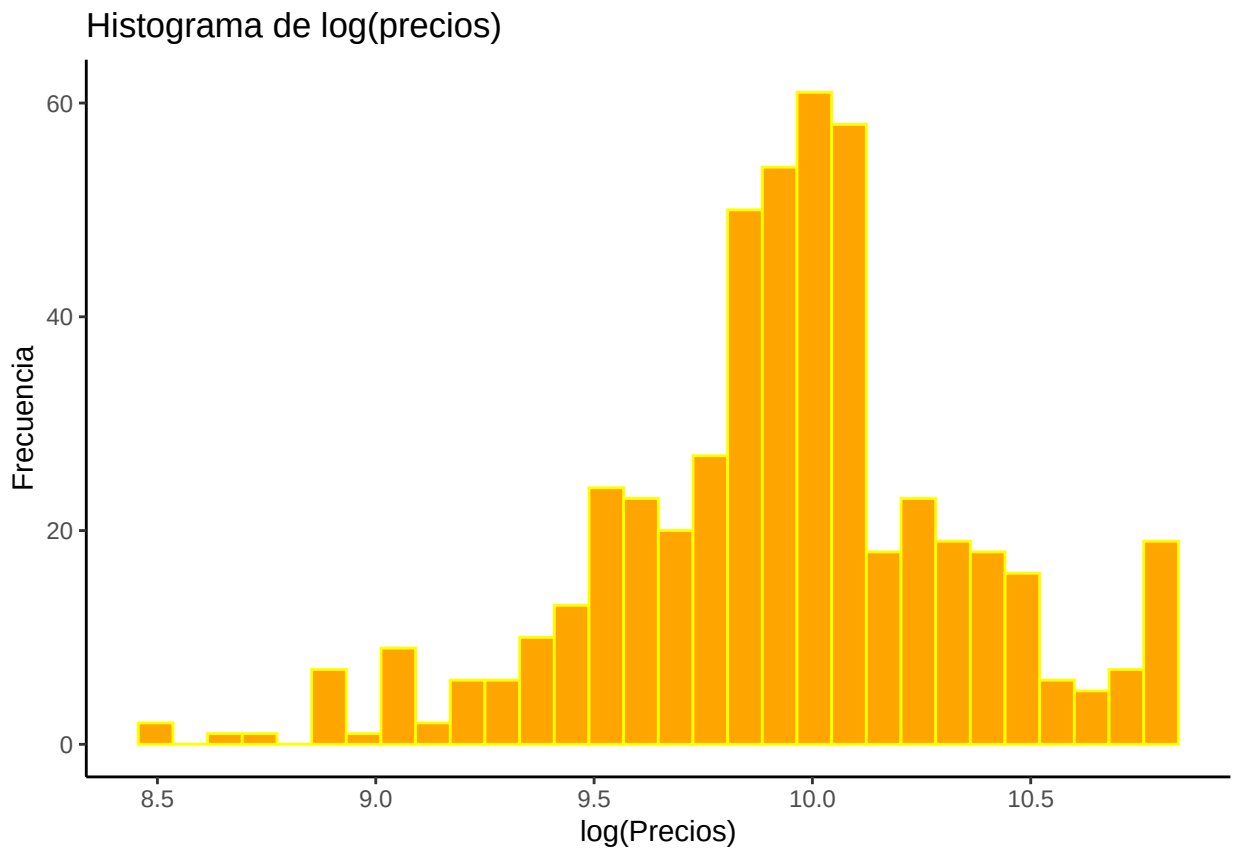


2. **Los precios de viviendas** también siguen una distribución sesgada a la derecha, ya que son muy pocas las casas de "lujo" que tienen precios muy por encima de la media.

```
data("hprice2")
ggplot(data = hprice2, aes(x =price)) +
  geom_histogram(bins=30, color = "yellow", fill = "orange") +
  theme_classic() +
  labs(title = "Histograma de precios", x = "Precios", y = "Frecuencia")
```




```
ggplot(data = hprice2, aes(x =lprice)) +
  geom_histogram(bins=30, color = "yellow", fill = "orange") +
  theme_classic() +
  labs(title = "Histograma de log(precios)", x = "log(Precios)", y = "Frecuencia")
```



3. **El tamaño de las empresas** si se quieren comparar o se tiene en una misma muestra empresas grandes y empresas pequeñas, por la diferencia del tamaño de estas, aplicar una transformación logarítmica ayuda a comparar sus variables en una escala razonable.

Log-nivel

En este caso, la **variable dependiente** está en logaritmo, mientras las **variables independientes** matienen su forma original (nivel).

$$log(y_i) = \beta_0 + \beta_1 x_{1i} + u_i$$

Interpretación: - Los coeficientes β_i representan cómo un *cambio unitario* en x provoca un *cambio porcentual* en y - Si $\beta_1 = 0.03$, quiere decir que un aumento de **1 unidad de** x_1, llevará a un aumento de 3% en y.

Log-Log

En este tipo de regresión, tanto la **variable dependiente** como las **variables independientes** son transformadas por logaritmos.

$$log(y_i) = \beta_0 + \beta_1 log(x_{1i}) + \beta_2 log(x_{2i}) + u_i$$

Interpretación

Los coeficientes β_i en una regresión *log-log* representan **elasticidades**. Es decir, cómo cambia la variable dependiente y en **porcentaje** por cada **1%** de cambio en x_i. - Si $\beta_1 = 0.5$, un aumento del 1% en x_1 llevará a un aumento de 0.5% en y.

Derivación matemática

Derivamos con respecto a $log(x_1)$

$$\frac{\Delta log(y)}{\Delta log(x)} = \beta_1$$

Nivel-Log

En este caso, la **variable dependiente** se mantiene en su forma original (nivel), pero las **variables independientes** se transforman utilizando logaritmos.

$$y_i = \beta_0 + \beta_1 log(x_{i1}) + u_i$$

Interpretación:

- Los coeficientes β_i representan cómo un **cambio porcentual** en x afecta el **nivel** de la variable dependiente y.
- Si $\beta_1 = 2$, un aumento del **1%** en x_1 provocaría un aumento de **2 unidades** en y.

Tabla de resumen

Tipo de regresión	Interpretación de los coeficientes
Log-Log	β_i representa la **elasticidad**: un cambio del 1% en X_i produce un cambio del β_i% en Y.
Nivel-Log	β_i indica el cambio en el nivel de Y por un **1% de cambio** en X_i.
Log-Nivel	β_i indica el cambio porcentual en Y por un **cambio unitario** en X_i.
Nivel-Nivel	β_i indica el **cambio absoluto** en Y por un **cambio unitario** en X_i.

Polinomios

Las **regresiones con polinomios** son un tipo de regresión **no lineal** que permite modelar relaciones más *flexibles*, es decir, permite un mejor ajuste cuando los datos presentan **curvaturas** o **patrones no lineales**.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + \beta_3 x_{1i}^2 + u_i$$

Este tipo de regresiones se usa cuando: - conocemos la posibilidad de que la relación tenga una curvatura, por ejemplo, cuando el efecto de x en y cambia a medida que x crece.

Sin embargo, utilizar polinomios sin un motivo claro (sólo para ajustar mejor los datos), tiene consecuencias:

- **Interpretación**: Cuando utilizamos polinomios de orden superior al cuadrado, es difícil dar una interpretación "económica" al coeficiente.
- **Inestabilidad de los coeficientes**: Los coeficientes se vuelen altamente sensibles a pequeños cambios o a la especificación del modelo.
- **Sobre ajuste**: u *overfitting*, es cuando se utiliza polinomios de orden superior, sin embargo, el modelo se ajusta al **ruido** de los datos y no al patrón en el que estamos interesados, esto lleva a que el modelo tenga bajo **poder predictivo** en otros datos.

Interpretación de una regresión polinómica

La interpretación de los coeficientes de una regresión polinómica depende del grado del polinomio:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + \beta_3 x_{1i}^2 + u_i$$

β_1 es el coeficiente *lineal* de x_1, β_2 es el coeficiente *cuadrático* de x_2.

Derivación

$$\frac{\delta(y|x_1)}{\delta x_1} = \beta_1 + 2\beta_3$$

- El signo de β_2 (el término cuadrático) indica la curvatura de la parábola:
  - Si β_2 es positivo, significa que es **convexa** (forma de U)
  - Si β_2 es negativa, significa que es **cóncava** (forma de U invertida).
- Dado que el efecto de x_1 en y **no es constante**, es decir, **depende del valor de** x_1 que se esté considerando, se interpreta el ***efecto marginal*** o la pendiente en un valor específico de x_1.
- Es posible calcular el *vértice* o el punto de inflexión, es decir, el valor de x_1 a partir del cual, la relación con y cambia de positivo a negativo o viceversa:

$$x_v = -\frac{\beta_1}{2\beta_3}$$

Ejemplo

Supongamos que tenemos el modelo del precio de las casas, donde estamos interesados en el *efecto* de la concentración de *óxido nitroso* y el *número de habitaciones* sobre el precio de estas:

$$lprice = \beta_0 + \beta_1 lnox + \beta_2 rooms + \beta_3 rooms^2 + \beta_4 dist + \beta_5 stratio + \beta_6 crime + u$$

```
library(fixest)
mco_hprice2<-feols(lprice ~ lnox + rooms + I(rooms^2)+ dist + stratio + crime, data = hprice2, vcov = "
summary(mco_hprice2)
```

```
## OLS estimation, Dep. Var.: lprice
## Observations: 506
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 13.698519   0.779575 17.57179  < 2.2e-16 ***
## lnox        -0.781222   0.089339 -8.74447  < 2.2e-16 ***
## rooms       -0.731679   0.237408 -3.08194 2.1702e-03 **
## I(rooms^2)   0.075816   0.018046  4.20117 3.1459e-05 ***
## dist        -0.035355   0.007537 -4.69110 3.5136e-06 ***
## stratio     -0.036889   0.004742 -7.77979 4.2057e-14 ***
## crime       -0.014673   0.002399 -6.11651 1.9334e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.230802   Adj. R2: 0.67749
```

- Para interpretar el efecto de *lnox* sobre *lprice* primero tenemos que reconocer que se trata de un modelo *log-log*, por tanto, se debe interpretar β_1 como una **elasticidad**.

- **Interpretación** β_1: Un incremento en **1%** en la concentración de óxido nitroso está asociado a una disminución de **0.78%** en los precios en las casas.

Ahora queremos saber el efecto de *rooms* sobre el precio de las casas, para la media de *rooms*

El efecto de una casa adicional esta dado por:

$$-0.731679 + 2 \times 0.075816 \times rooms$$

```
mean(hprice2$rooms)
```

```
## [1] 6.284051
```

Por tanto el efecto del número de habitaciones promedio sobre el precio de las casas es:

$$-0.731679 + 2 \times 0.075816 \times 6.284051 = 0.2211765$$

Es decir, al número promedio de habitaciones 6.24, añadir una habitación adicional se asocia con un incremento de aproximadamente **22.12%** en el precio de las casas.

El signo de $\beta_3 > 0$, el coeficiente de $rooms^2$, indica que la relación es **convexa** es decir, hasta un cierto número de habitaciones, el efecto del número de habitaciones es decreciente, y luego de este punto, el efecto se vuelve creciente.

Variables binarias

En algunos casos, tenemos variables independientes x no numéricas, en específico binarias. Estas variables son las que solamente toman dos valores, generalmente 0 o 1, y representan la **presencia o ausencia** de un atributo o característica. Un ejemplo común es la variable que indica si una persona tiene o no empleo: empleado=1, desempleado=0.

Las **variables binarias** se utilizan cuando se tiene una variable **cualitativa** o **categórica** que puede describirse en términos de dos grupos ***mutuamente excluyentes*** (ej, género). Son útiles en situaciónes donde se quiere evaluar el *efecto de una condición específica* en una variable dependiente.

Estimación con variables binarias o dummies

Cuando se incluyen las variables binarias en un modelo de regresión, el proceso depende de la naturaleza de la *variable dependiente* si esta es continua, mientras que alguna de las variables independientes es binaria, se utiliza el modelo de **regresión lineal**, en caso contrario, (la variable *dependiente* es binaria) se utilizan modelos de **regresión logística**. En este segmento solamente abordaremos el primer caso.

Regresión lineal con variable independiente binaria

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

En esta ecuación, x_{1i} es una variable binaria.

Interpretación

El coeficiente β_1 indica el cambio esperado en la variable dependiente y en presencia del atributo, es decir, cuando x cambia de 0 a 1.

- Por ejemplo,si se está modelando el *salario* en función del **género**, donde *género* = 1 para mujeres y *género* = 0 para hombres), el coeficiente β_1 nos diría **la diferencia en salario** entre hombres y mujeres.

Retomemos el ejemplo de la primera clase:

$$\begin{aligned} E[salario|género] &= \beta_0 + \beta_1 género \\ E[salario|género = 0] &= \beta_0 \text{ (salario de hombres)} \\ E[salario|género = 1] &= \beta_0 + \beta_1 \text{ (salario mujeres)} \\ E[salario|género = 1] - E[salario|género = 0] &= \beta_1 \text{ (diferencia salarial)} \end{aligned}$$

Interacciones

Los modelos de regresión con **interacciones** se utilizan cuando sabemos que las variables no actúan de forma independiente, sino que el impacto de una variable sobre la variable dependiente se ve modificado por la presencia o el valor de otra.

Los modelos con interacciones permiten modelar estas relaciones **no aditivas** entre variables. En lugar de suponer que cada variable tiene un **efecto constante** sobre la variable dependiente, un modelo con interacciones, asume que los efectos de las variables **pueden cambiar** dependiendo de los valores de las otras.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + u_i$$

β_3 representa el efecto de las variables x_1 y x_2 sobre y.

Dos variables continuas

Cuando x_1 y x_2 son variables continuas, el coeficiente de interacción β_3 representa el cambio en el **efecto de una variable** sobre la variable dependiente **debido** al **cambio en la otra variable continua**.

Ejemplo

Supongamos que se está analizando el salario en función de los años de experiencia y el nivel de educación. La relación entre experiencia y salario puede depender de **qué tan alta sea la educación** de la persona. Es decir, el impacto de la experiencia en el salario podría ser mayor para personas con **más altos niveles de educación**. Este sería un efecto combinado:

$$Salario = \beta_0 + \beta_1 Exp + \beta_2 Educ + \beta_3 (Exp \times Educ) + u$$

- **Interpretación**:

$$Salario = 450 + 210 Exp + 108 Educ + 190 Exp \times Educ + u$$

El efecto de un año adicional de educación para una persona con 5 años de experiencia es entonces:

$$\frac{\delta Salario}{\delta Educ} = 108 + 190 \times Exp(5\,años) = 1058$$

Dos variables binarias

Cuando las variables independientes que interactúan son binarias o dummies, el coeficiente de la interacción mide **el cambio en el efecto** de una de las variables binarias sobre y, **condicionado** a la presencia de la otra variable binaria.

Ejemplo

Supongamos ahora que se está analizando el efecto que tiene el género y el estado civil sobre el salario por hora de los individuos. Existe evidencia previa que muestra que cuando la persona es mujer, el efecto del estado civil sobre los ingresos es distinto que cuando el individuo es hombre.

$$ingreso_i = \beta_0 + \beta_1 mujer_i + \beta_2 casado_i + \beta_3 mujer_i \times casado_i$$

$mujer = 1$ cuando el individuo es mujer e igual a 0 en caso contrario, $casado = 1$ cuando el estado civil de la persona es casado e igual a 0 en caso contrario.

```
library(wooldridge)
data("wage1")
mco_int<-feols(wage~female+ married+female:married, data=wage1, vcov="hc1")
summary(mco_int)
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 526
## Standard-errors: Heteroskedasticity-robust
##                 Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)     5.168023   0.293016 17.63733  < 2.2e-16 ***
## female         -0.556440   0.400727 -1.38858 1.6555e-01
## married         2.815009   0.435098  6.46982 2.2614e-10 ***
## female:married -2.860683   0.543267 -5.26570 2.0445e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.33904   Adj. R2: 0.17628
```

En esta ecuación podemos distinguir los siguientes efectos:

- **efecto para mujer casada**:

$ingreso = \beta_0 + \beta_1 + \beta_2 + \beta_3 = 4.57$

- **efecto para mujer soltera**:

$ingreso = \beta_0 + \beta_1 = 4.61$

- **efecto para hombre casado**:

$ingreso = \beta_0 + \beta_2 = 7.98$

- **efecto para hombre soltero**:

$ingreso = \beta_0 = 5.17$

El efecto de estar casado es diferente para hombres y mujeres, mientras los hombres casados ganan 7.98USD por hora, las mujeres casadas ganan 4.57USD.

Variables categóricas

Las variables categóricas son aquellas que toman un **número limitado de valores distintos** (categorías) y representan *grupos* o *características cualitativas*.

Cuando se incluyen **variables categóricas** en una regresión MCO, estas deben ser ***transformadas en variables dummy*** (o binarias). Esto se hace asignando un valor de 1 para cada categoría específica y 0 para todas las demás.

Sin embargo, para evitar problemas de multicolinealidad perfecta, si tenemos m categorías, en la regresión sólo debemos incluir $m-1$ dummies. Cuando se utiliza una variable dummy por categoría, se cae en la *"trampa de la variable ficticia"*(dummy trap). Esto se debe a que la última variable puede ser predicha perfectamente por las otras.

Por lo tanto, al estimar colocamos $m-1$ dummies por las m categorías, y la categoría omitida será tomada como **referencia** o **base**, contra la que compararemos el coeficiente de todas las demás al momento de la interpretación.

Ejemplo

Suponga que estamos interesados en estimar un modelo que mida el efecto sobre el ingreso salarial que genera el vivir en diferentes zonas de la ciudad.

Las categorías son las siguientes:

```
wage1$region <- as.factor(wage1$region)
levels(wage1$region)
```

```
## [1] "east"     "northcen" "south"    "west"
```

Utilizando la paquetería `fastDummies` creamos una dummie por cada categoría

```
library("fastDummies")
wage1<-dummy_cols(wage1, select_columns = "region")
```

Manualmente se crea de la siguiente manera:

```
wage1$northcen <- ifelse(wage1$region == "northcen", 1, 0)
wage1$south <- ifelse(wage1$region == "south", 1, 0)
wage1$west <- ifelse(wage1$region == "west", 1, 0)
wage1$east <- ifelse(wage1$region == "east", 1, 0)
```

Ahora estimamos el siguiente modelo:

$$wage_i = \beta_0 + \beta_1 northcen_i + \beta_2 south_i + \beta_3 west_i + u_i$$

```
mco_region<-feols(wage~region_northcen+region_south+region_west, data=wage1, vcov="hc1")
summary(mco_region)
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 526
## Standard-errors: Heteroskedasticity-robust
##                  Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)      6.369746   0.402247 15.835399 < 2.2e-16 ***
## region_northcen -0.659291   0.495539 -1.330452  0.183950
## region_south    -0.982847   0.461809 -2.128257  0.033784 *
## region_west      0.243625   0.600751  0.405534  0.685251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.65715   Adj. R2: 0.011852
```

Interpretación

- Los trabajadores que viven en la región "northcen" perciben un salario por hora 0.66USD **menor** que aquellos que viven en la zona "east"
- Los que trabajan en la zona "west" por otro lado, perciben un salario 0.24USD más alto que los de la zona "east"
- Los trabajadores de la zona "south" son los que perciben menor salario por hora, recibiendo 0.98USD **menos** que los de la zona "east".