Analítica de datos

Regresión lineal simple

Pontificia Universidad JAVERIANA Bogotá

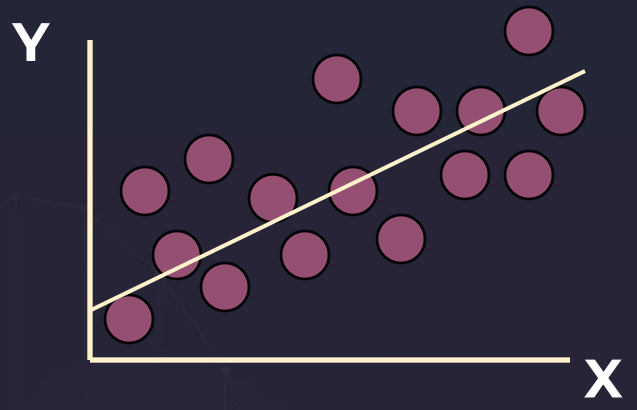
Profesor: Nicolás Velásquez

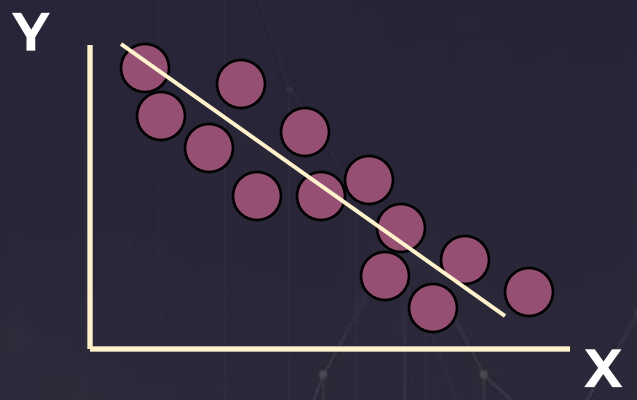
Correlación

- Podemos utilizar un diagrama de dispersión para realizar un primer análisis de la relación entre dos variables.
- El coeficiente de correlación (lineal) es utilizado para medir la fuerza de la asociación (lineal) entre dos variables.
- Solamente nos habla de la relación (lineal) entre dos variables.
- No implica relación causal.

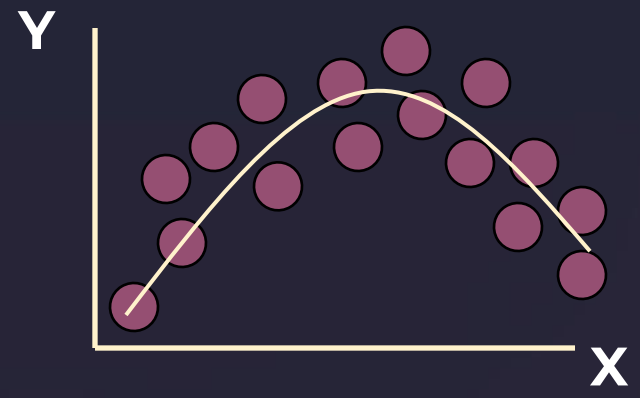
Tipos de relaciones entre dos variables

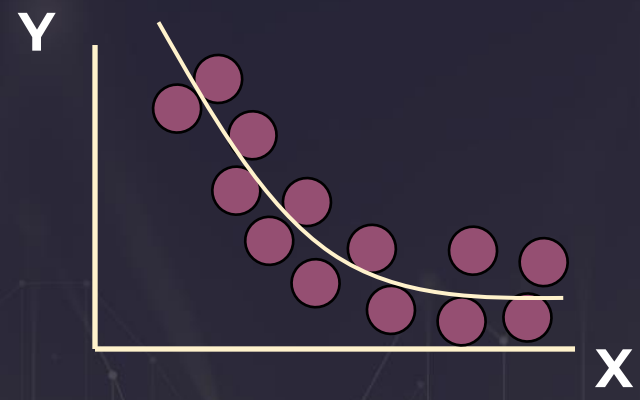
Relaciones lineales





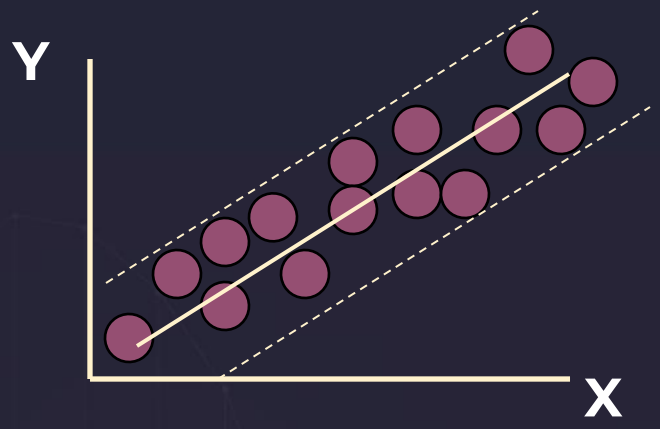
Otras relaciones

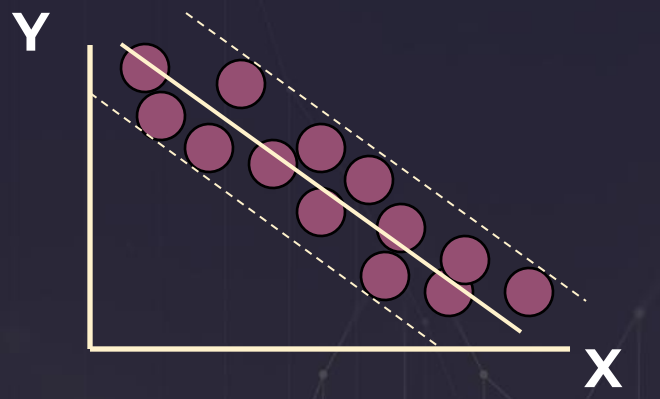




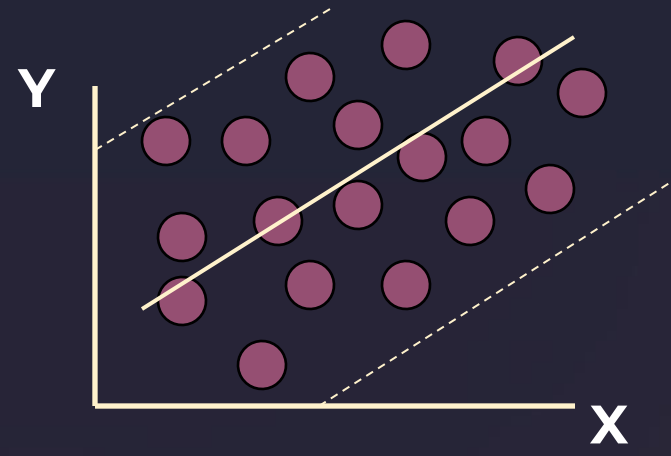
Tipos de relaciones entre dos variables

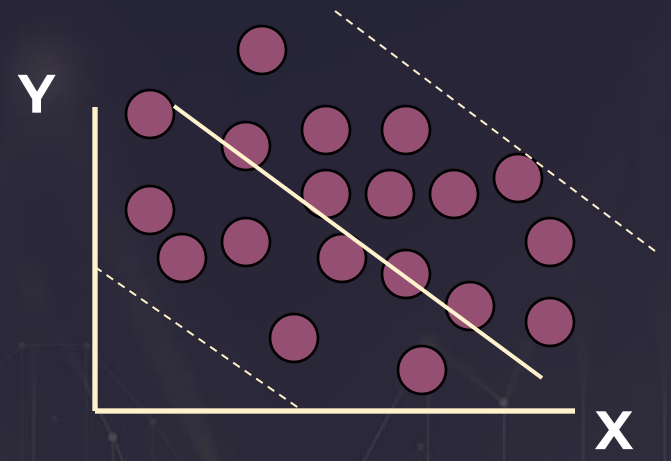
Relaciones fuertes



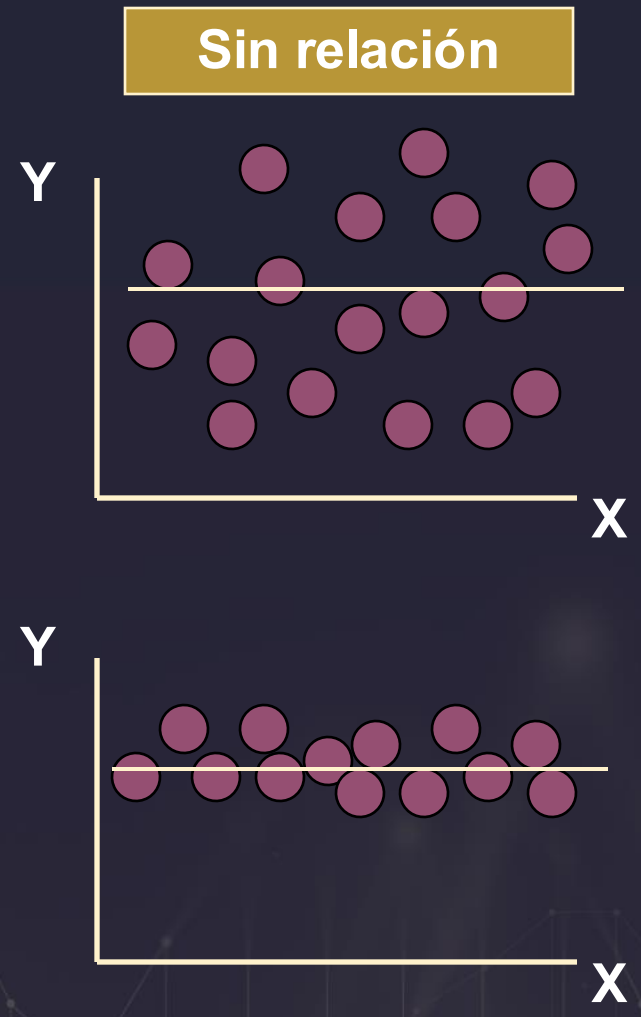


Relaciones débiles





Tipos de relaciones entre dos variables



Introducción al análisis de regresión

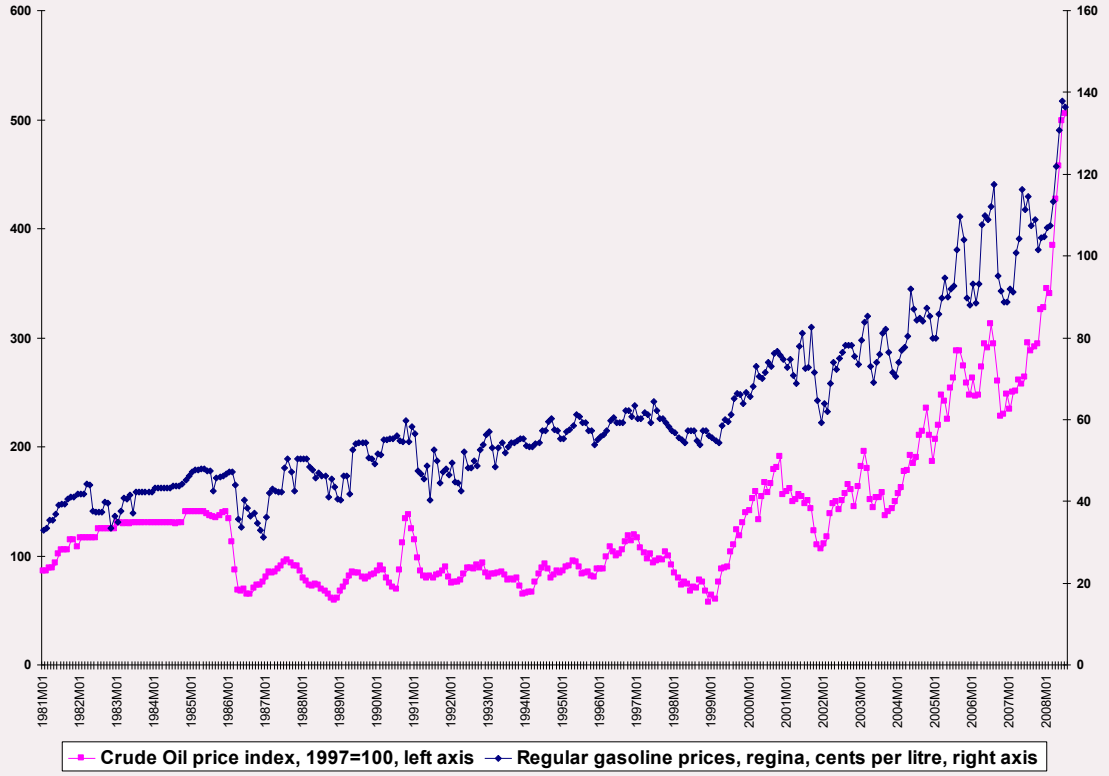
- El análisis de regresión se usa para:
  - Predecir el valor de una variable dependiente, basándonos en los valores de por lo menos una variable independiente.
  - Conocer el efecto de cambios en una variable independiente sobre la variable dependiente.

Variable dependiente: variable que queremos predecir o explicar.

Variable independiente: variable utilizada para predecir o explicar la dependiente.

Ejemplos

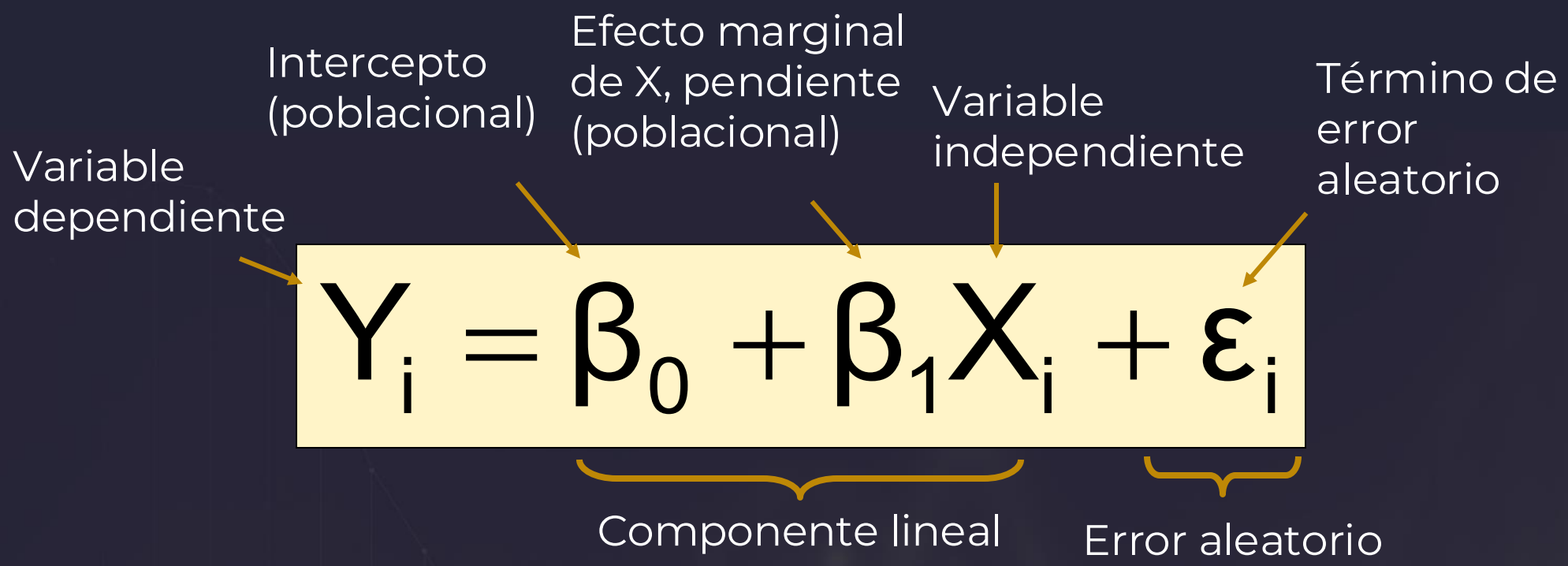
- Precio de la gasolina y el precio del petróleo:
  - **Variable dependiente**: precio retail (minorista) de la gasolina en Bogotá
  - **Variable independiente**: precio internacional del petróleo.



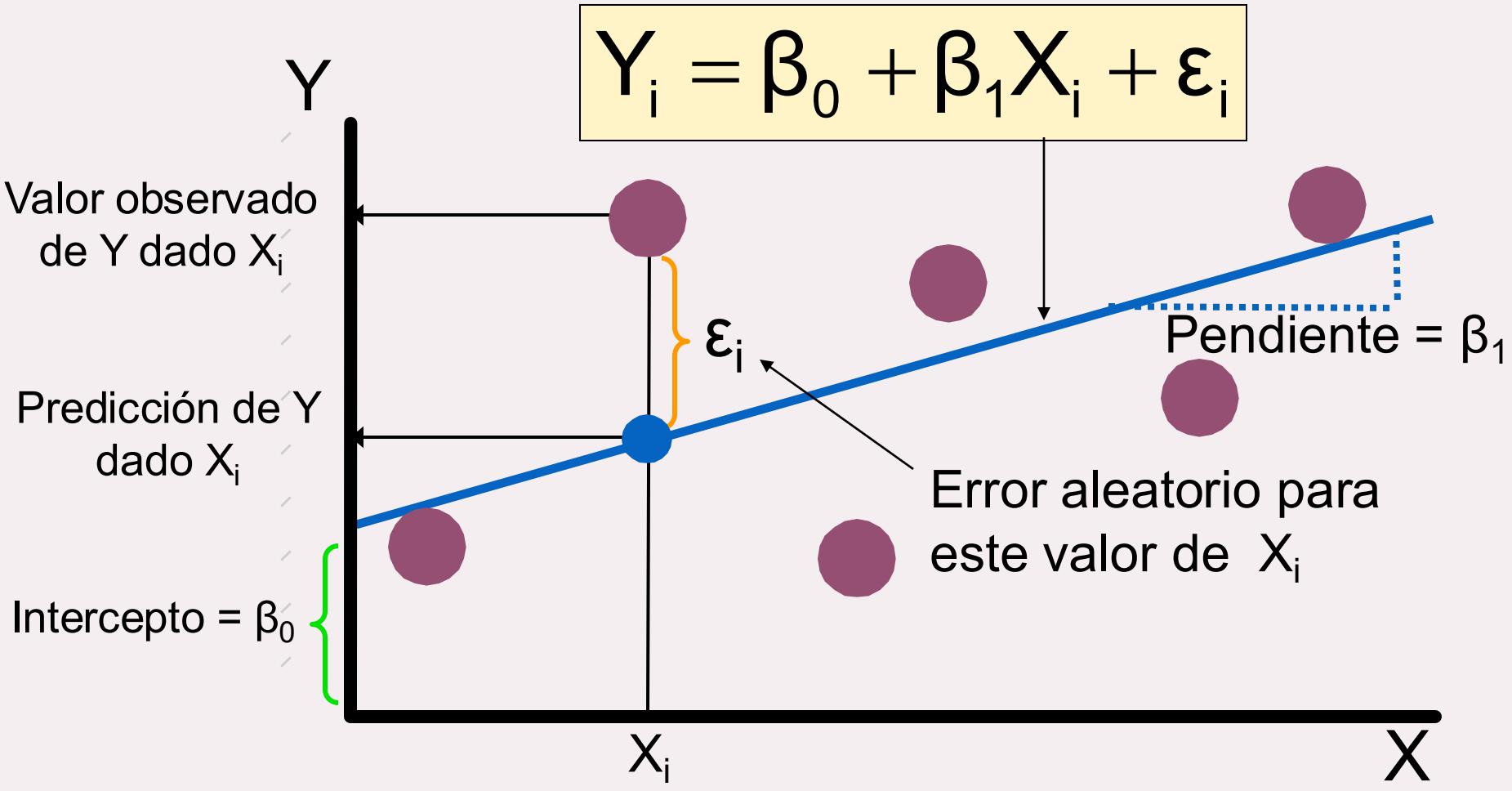
Modelo de regresión lineal simple

- Sólo **una** variable independiente, X.
- La relación entre X y Y es descrita por una función lineal.
- Se asume que cambios en Y están relacionados con cambios en X (causalidad).

Modelo de regresión lineal simple



Modelo de regresión lineal simple



Ecuación de regresión lineal simple (línea de predicción)

Esta ecuación provee un estimado de la línea regresión poblacional.

Valor estimado de Y para la observación i.

Estimado del intercepto

Estimado de la pendiente/efecto marginal

$$\hat{Y}_i = b_0 + b_1 X_i$$

Valor de X para la observación i

El método de mínimos cuadrados

b_0 y b_1 son los valores que minimizan la suma de las diferencias al cuadrado (una medida de distancia) entre Y y $\hat{Y}$

.

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Interpretación del intercepto y la pendiente (estimados)

- b_0 es el valor medio estimado de Y cuando el valor de X es cero.
- b_1 es el cambio estimado en valor medio de Y, como resultado de un incremento de una unidad en X.

Ejemplo

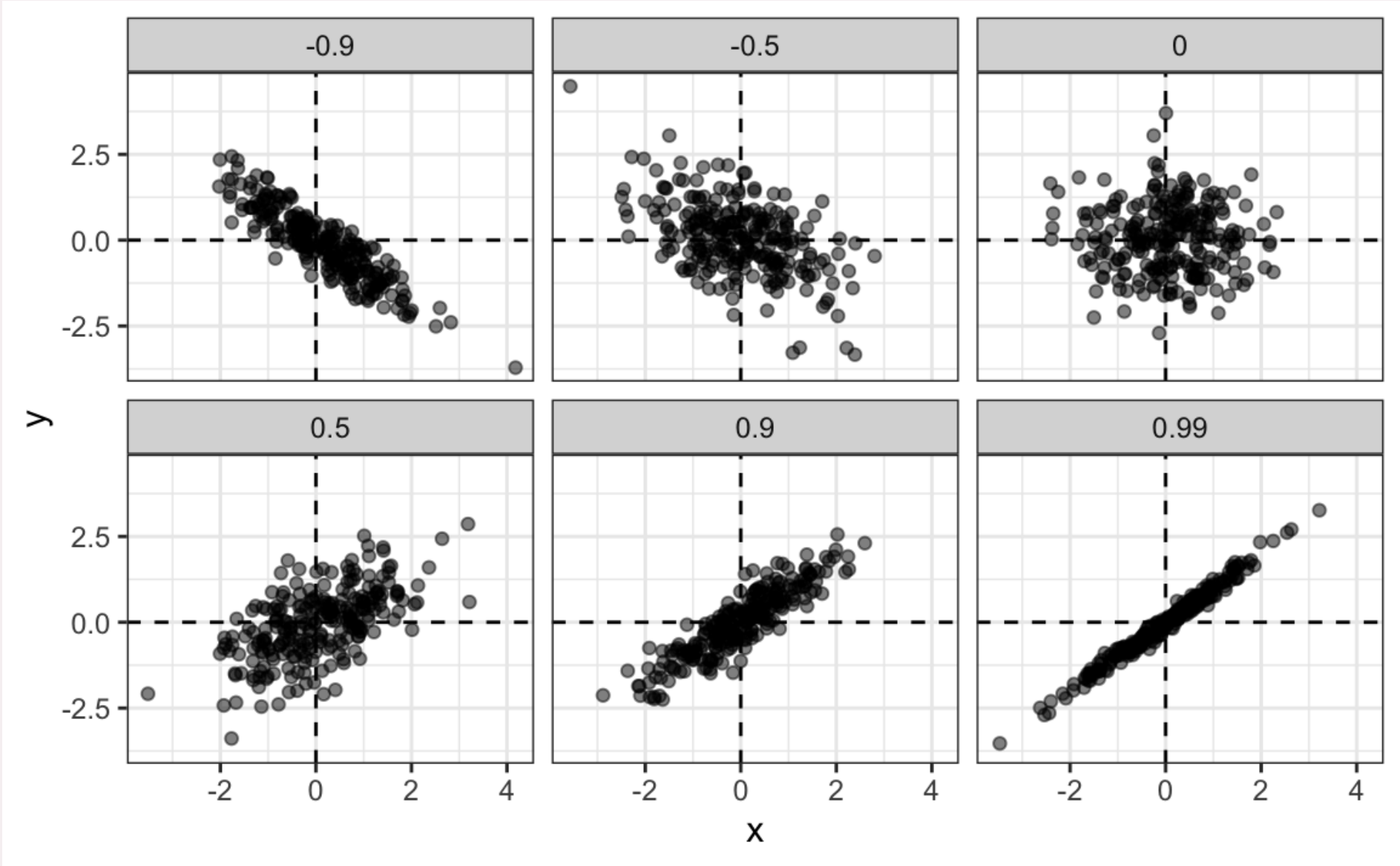
- Un agente de bienes raíces quiere examinar la relación entre el precio de venta de una casa y su tamaño (medido en pies cuadrados).

- Se selecciona muestral aleatoria de 29 casas.
  - Variable dependiente (Y) = precio en $1,000s.
  - Variable independiente (X) = pies cuadrados.

Ejemplo en R

```
## Importar archivo casas y
## cargar library

 library(tidyverse)
 library(ggthemes)

## Diagrama de dispersión con ggplot

Casas_ejemplo1slides |>
ggplot(aes(SQFT, PRICE)) +
geom_point(color = "#2E3B5F", size = 4,alpha = 0.8) +
theme_tufte()

## Coeficiente de correlación

Casas_ejemplo1slides |>
summarize(r = cor(SQFT, PRICE)) |>
pull(r)
```

Recordemos: coeficientes de correlación



Ejemplo en R

```
## Se puede agregar fácil la línea de regresión en ggplot2

Casas_ejemplo1slides |> ggplot(aes(SQFT, PRICE)) +
geom_point(color = "#2E3B5F", size = 4,alpha = 0.8)+
theme_tufte()+
geom_smooth(method = "lm")
```

Ejemplo en R

Podemos usar la función lm() en R para ajustar un modelo de regresión lineal simple.

```
library(jtools) #verificar si está instalado el paquete

# Ajusta el modelo de regresión lineal
lm_model <- lm(PRICE ~ SQFT, data = Casas_ejemplo1slides)

# Muestra los resultados del modelo
 summ(lm_model)

 summary(lm_model)
```

Para exportar tabla a Word dos opciones

```
library(broom)
library(flextable)
library(officer)
library(huxtable)

export_summs(lm_model,digits = 4,
to.file = "docx", file.name = "test.docx")

# Crear flextable
flextable(lm_summary) %>%
set_caption("Resultados de la regresión lineal simple") %>%
autofit() %>%
align(align = "center", part = "all") %>%
fontsize(size = 10) %>%
bold(part = "header") %>%
theme_vanilla  %>%
colformat_double(digits = 4)   %>%
add_footer_lines("Tabla de ejemplo clase analítica") # agregar nota al pie
%>%
save_as_docx(path="resultados_regression.docx")
```

Ejemplo: interpretación de b_o

$$\widehat{price} = 82.916 + 0.102 * SquareFeet$$

- b_0 es el valor medio estimado de Y cuando el valor de X es cero.
- Como una casa no puede tener 0 metros cuadrados, b_0 no tiene aplicación práctica.

Ejemplo: interpretación de b_1

$$\widehat{price} = 82.916 + 0.102 * SquareFeet$$

- b_1 estima el cambio en valor medio de Y como resultado de un incremento en una unidad de X.
- Aquí, $b_1 = 0.102$ nos dice que se estima que el valor medio de una casa se incrementa en 0.102*($1,000) = $102, por cada pie cuadrado adicional.

Ejemplo: predicción

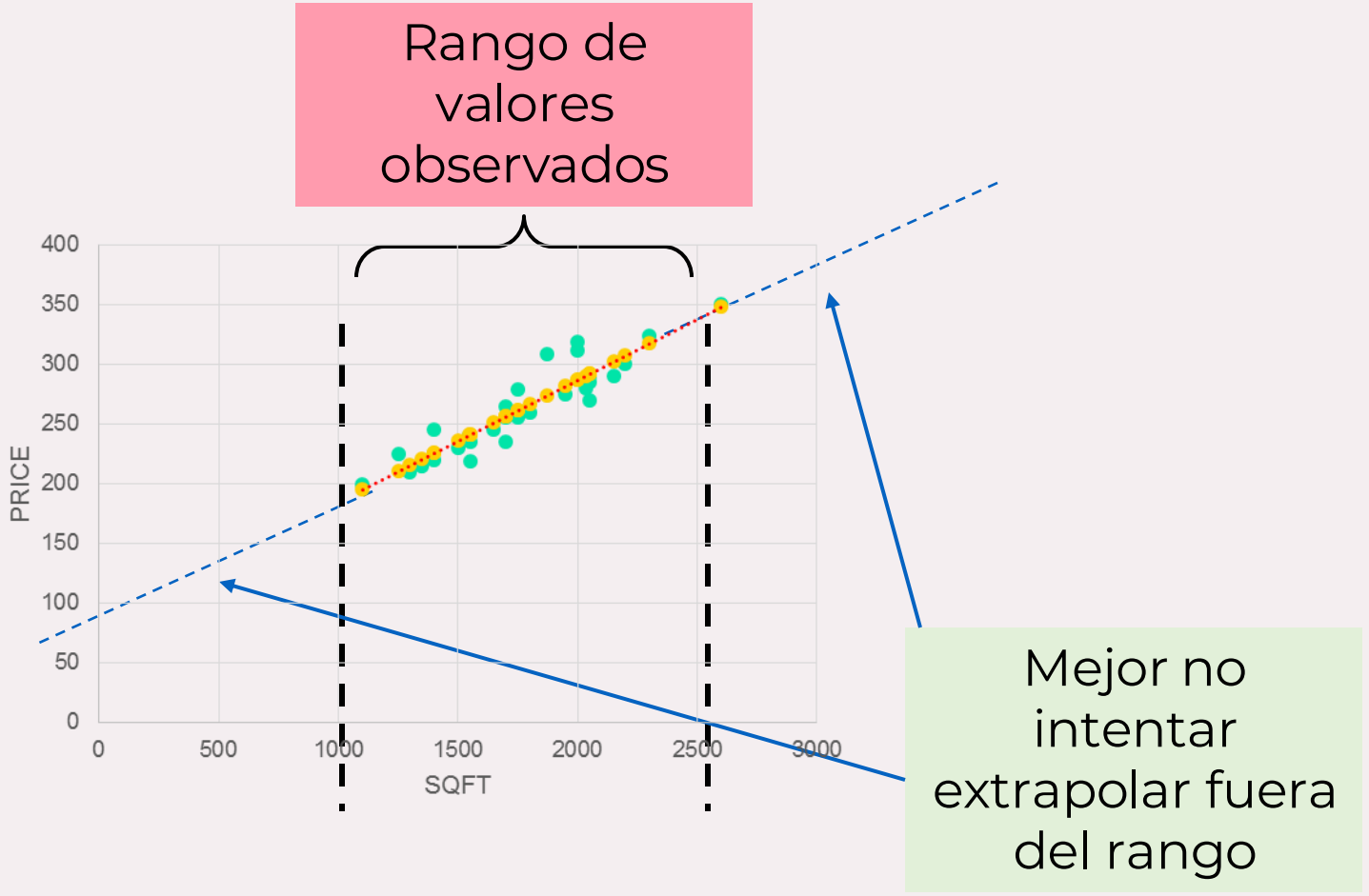
Prediciendo el precio de una casa de 2,000 pies cuadrados:

$$\widehat{price} = 82.916 + 0.102 * SquareFeet$$
$$= 82.916 + 0.102 * 2000 = 286.916$$

Predecimos que el precio de una casa de 2,000 pies cuadrados es 286.916($1,000s) = $286,916

Ejemplo: predicción

- Solamente hacer predicciones utilizando valores de X en el rango de valores observados.



Coeficiente de determinación: r^2

¿Qué tan bueno es nuestro modelo?

- El coeficiente de determinación nos dice qué tan bien capturamos cambios o variaciones en la variable dependiente con nuestra regresión lineal, comparado con simplemente usar el promedio de la variable dependiente para hacer predicción.

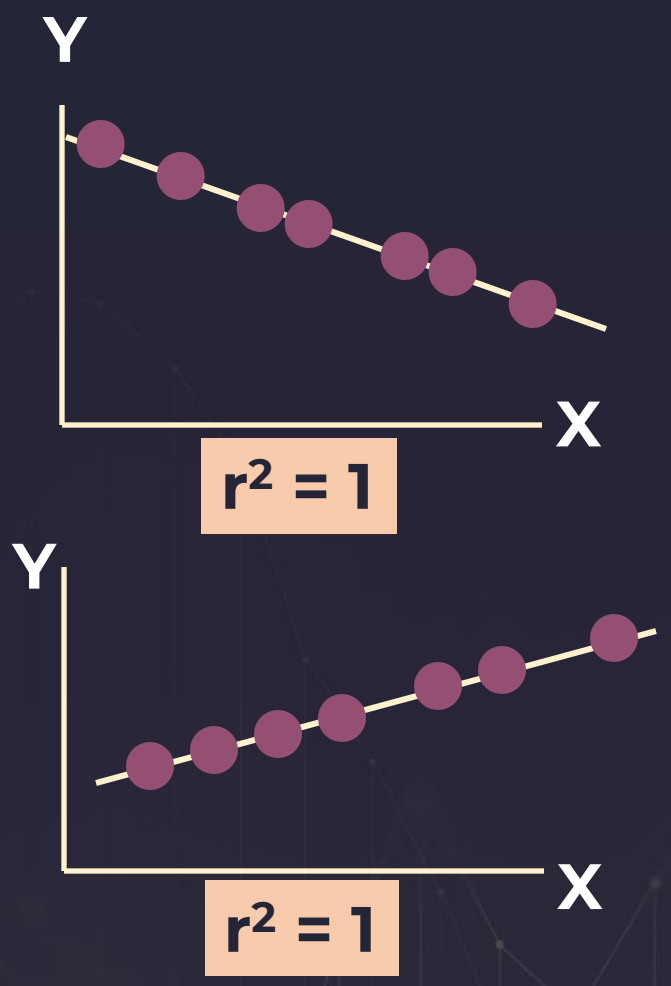
- En otras palabras, cuanto de la variación en Y es explicada por variaciones de X.

- También se le llama r-cuadrado y se le denota r^2.

- El r cuadrado nos dice que tan bien podemos predecir

$$0 \leq r^2 \leq 1$$

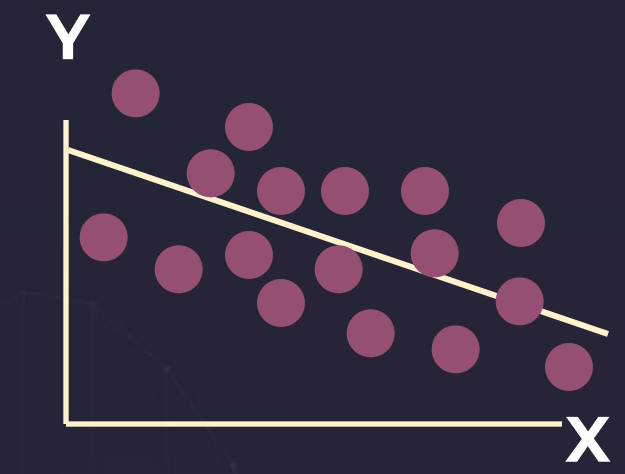
Ejemplos r^2

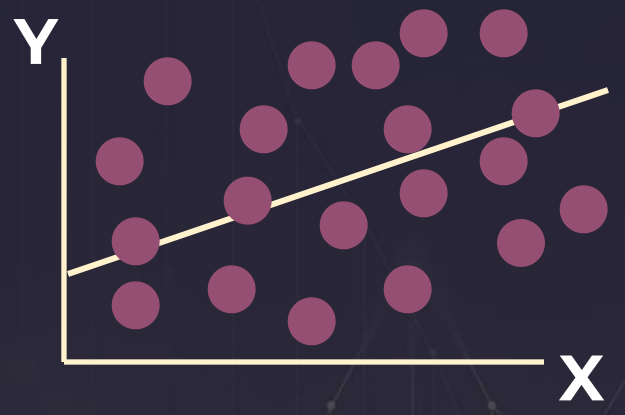


Relación lineal perfecta entre X y Y.

100% de la variación en Y explicada con variación en X.

Ejemplos r^2



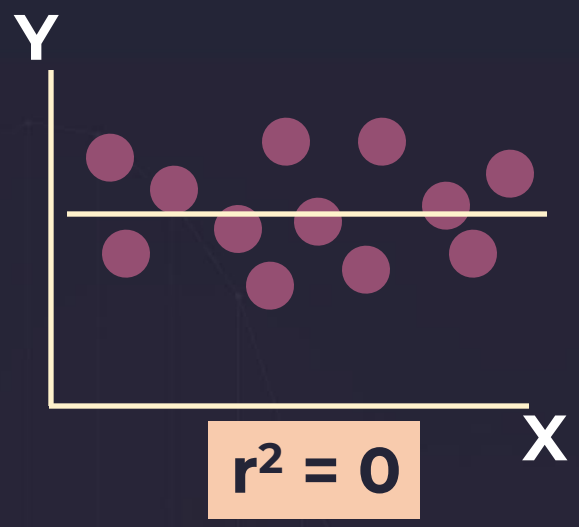


$0 < r^2 < 1$

Relaciones lineales más débiles entre X y Y.

Una parte, pero no toda la variación en Y es explicada por variación en X.

Ejemplos r^2



$r^2 = 0$

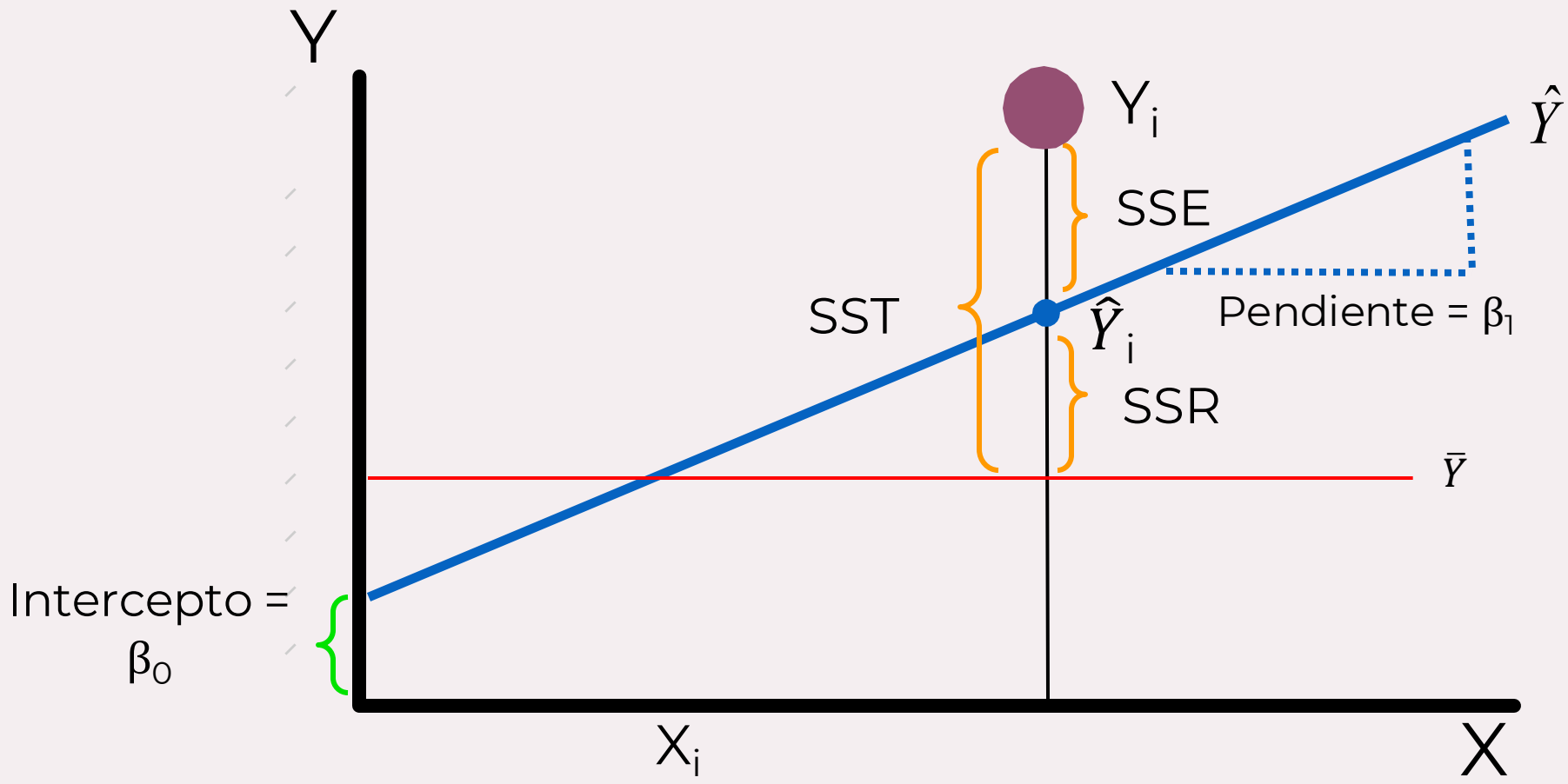
No hay relación lineal entre X y Y.

El valor de Y no depende de X. (Nada de la variación en Y se explica por variación en X.)

Coeficiente de determinación, r^2

Estimación:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{Error\ sum\ of\ squares}{Total\ sum\ of\ squares}$$



Coeficiente de determinación, r^2

Estimación:

$$R^2 = \frac{SSR}{SST} = \frac{Regression\ sum\ of\ squares}{Total\ sum\ of\ squares}$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{Error\ sum\ of\ squares}{Total\ sum\ of\ squares}$$

Donde,

$$SST = \sum_{i\bar{n}1}^{n} (Y_i - \bar{Y})^2$$

$$SSR = \sum_{i\bar{n}1} (\widehat{Y}_i - \bar{Y})^2$$

$$SSE = \sum_{i=1} (Y_i - \widehat{Y}_i)^2$$

Interpretando r^2 en nuestro ejemplo...

¿Qué porcentaje de la variación en los precios de las casas es explicado por la variación en los pies cuadrados?

Inferencia sobre la pendiente o el efecto marginal

¿Existe una relación lineal entre X y Y?

- Hipótesis nula y alternativa:
  - H_0: $\beta_1 = 0$ (no existe relación)
  - H_1: $\beta_1 \neq 0$ (sí existe relación)

Ejemplo: inferencia sobre la pendiente

Ecuación estimada:

$$\widehat{price} = 82.916 + 0.102 * SquareFeet$$

La pendiente/efecto marginal es 0.102

¿Existe una relación entre los pies cuadrados de una casa y su precio de venta?

Ejemplo: inferencia sobre la pendiente

$H_0: \beta_1 = 0$

De R: $H_1: \beta_1 \neq 0$

	Model 1
(Intercept)	82.9162 ***
	(14.5281)
SQFT	0.1020 ***
	(0.0081)
N	29
R2	0.8555

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
>   summ(lm_model)
MODEL INFO:
Observations: 29
Dependent variable: PRICE
Type: OLS linear regression

MODEL FIT:
F(1,27) = 159.89, p = 0.00
R² = 0.86
Adj. R² = 0.85

Standard errors: OLS
------------------------------------------------
                      Est.    S.E.   t val.      p
----------------- ------- ------- -------- ------
(Intercept)          82.92   14.53     5.71   0.00
SQFT                  0.10    0.01    12.64   0.00
------------------------------------------------
```

Decisión:
Rechazar H_0, ya que p-value < α.

Estadísticamente, hay suficiente evidencia para decir que el número de pies cuadrados afecta el precio de venta de las casas.

REGRESIÓN LINEAL MÚLTIPLE

Modelo de regresión múltiple con dos variables independientes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

donde:

β_0 = intercepto.

β_1 = "efecto" de un cambio en 1 unidad en X_1 sobre Y, manteniendo X_2 constante.

β_2 = "efecto" de un cambio en 1 unidad en X_1 sobre Y, manteniendo X_2 constante.

ε_i = error aleatorio en Y para la observación i.

Ejemplo: 2 variables independientes

- Un distribuidor de pies congelados quiere examinar los factores que influyen en la demanda.
  - Variable depend.: Ventas (unidades x sem.)
  - Variables independ.: Precio (en $), Gasto publicidad ($100's)
- Se recolectan datos por 15 semanas.

Ejemplo

Semana	Ventas	Precio ($)	Publicidad ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Ecuación de regresión múltiple:

$$\widehat{\text{Ventas}} = b_0 + b_1\,(\text{Precio}) + b_2\,(\text{Publicidad}).$$

Estimar en R.

Crear data frame y correr regresión

```
datos_ejemplo <- data.frame(
Semana = 1:15,
Ventas = c(350, 460, 350, 430, 350, 380, 430, 470, 450, 490, 340, 300, 440, 450, 300),
Precio = c(5.50, 7.50, 8.00, 8.00, 6.80, 7.50, 4.50, 6.40, 7.00, 5.00, 7.20, 7.90, 5.90, 5.00, 7.00),
Publicidad = c(3.3, 3.3, 3.0, 4.5, 3.0, 4.0, 3.0, 3.7, 3.5, 4.0, 3.5, 3.2, 4.0, 3.5, 2.7)
)
```

```
lm_model_multiple <-lm(Ventas ~ Precio + Publicidad,
data = datos_ejemplo)

summ(lm_model_multiple)
```

Ejemplo

$$\widehat{\text{Ventas}} = 306.526 - 24.975(\text{Precio}) + 74.131(\text{Publicidad})$$

Ventas en número de pies por semana
Precio en dólares
Publicidad en $100's.

b_1 = -24.975: se estima que las las ventas disminuyen, en promedio,

_____

_____

_____

b_2 = 74.131:

_____

_____

_____

_____

_____

Utilizando la ecuación para hacer predicciones

Predecimos las ventas en una semana cuando el precio es $5.50 y el gasto en publicidad $350:

$$\widehat{Ventas} = 306.526 - 24.975(\text{Precio}) + 74.131(\text{Publicidad})$$
$$= 306.526 - 24.975\,(5.50) + 74.131\,(3.5)$$
$$= 428.6216$$

Predecimos venta de 428.6216 pies.

Notar que publicidad está en $100s, entonces $350 sería $X_2 = 3.5$.

Variables Dummy

- Una variable categórica con dos niveles:
  - sí o no, masculino o femenino.
  - Codificado como 0 o 1.

- Si la variable categórica tiene más de dos niveles, el número de dummies necesarias es: número de niveles – 1.

Variable dummy con 2 niveles

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Sea:

Y = venta de pies

X_1 = precio

X_2 = festivo (X_2 = 1 si hubo día festivo durante la semana y X_2 = 0 si no hubo día festivo).

Variable dummy con 2 niveles

$\hat{Y} = b_0 + b_1X_1 + b_2(1) =$	$(b_0 + b_2) + b_1X_1$	**Festivo**
$\hat{Y} = b_0 + b_1X_1 + b_2(0) =$	$b_0 + b_1X_1$	**No festivo**

Diferentes interceptos — **Misma pendiente**

Si $H_0: \beta_2 = 0$ se rechaza, entonces "Festivo" tiene un efecto significativo sobre ventas.

Variable dummy con 2 niveles

Ejemplo: $\widehat{\text{Ventas}} = 300 - 30(\text{Precio}) + 15(\text{Festivo})$

Ventas: número de pies vendidos en semana

Precio: precio del pie en $

Festivo: $\begin{cases} 1 & \text{si festivo en la semana} \\ 0 & \text{si no festivo en la semana} \end{cases}$

$b_2 = 15$

Estimamos _____

_____, dado el mismo precio.

Variable dummy con más de 2 niveles

- Número de dummies que necesitaremos es **uno menos el número de niveles.**
- Ejemplo:

  Y = precio casa; X_1 = pies cuadrados.

- Si pensamos que el estilo de la casa importa:

  Estilo = ranch, split level, colonial.

3 niveles, entonces necesitamos 2 dummies

Variable dummy con más de 2 niveles

- Ejemplo: Podríamos seleccionar “Colonial” como categoría de “default”. X_2 y X_3 se usan para las otras dos categorías:

Y = precio casa

X_1 = pies cuadrados

X_2 = 1 si ranch, 0 si no

X_3 = 1 si split-level, 0 si no

La ecuación de regresión múltiple sería:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Variable dummy con más de 2 niveles

Supongamos que estimamos:

$$\hat{Y} = 20.43 + 0.045X_1 + 23.53X_2 + 18.84X_3$$

Si casa colonial: $X_2 = X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1$$

Si rancho: $X_2 = 1$; $X_3 = 0$

$$\hat{Y} = 20.43 + 0.045X_1 + 23.53$$

Si split-level: $X_2 = 0$; $X_3 = 1$

$$\hat{Y} = 20.43 + 0.045X_1 + 18.84$$

Manteniendo el número de pies cuadrados constante, estimamos que un rancho tendrá, en promedio, un precio 23.53 miles de dólares mayor que una casa colonial.

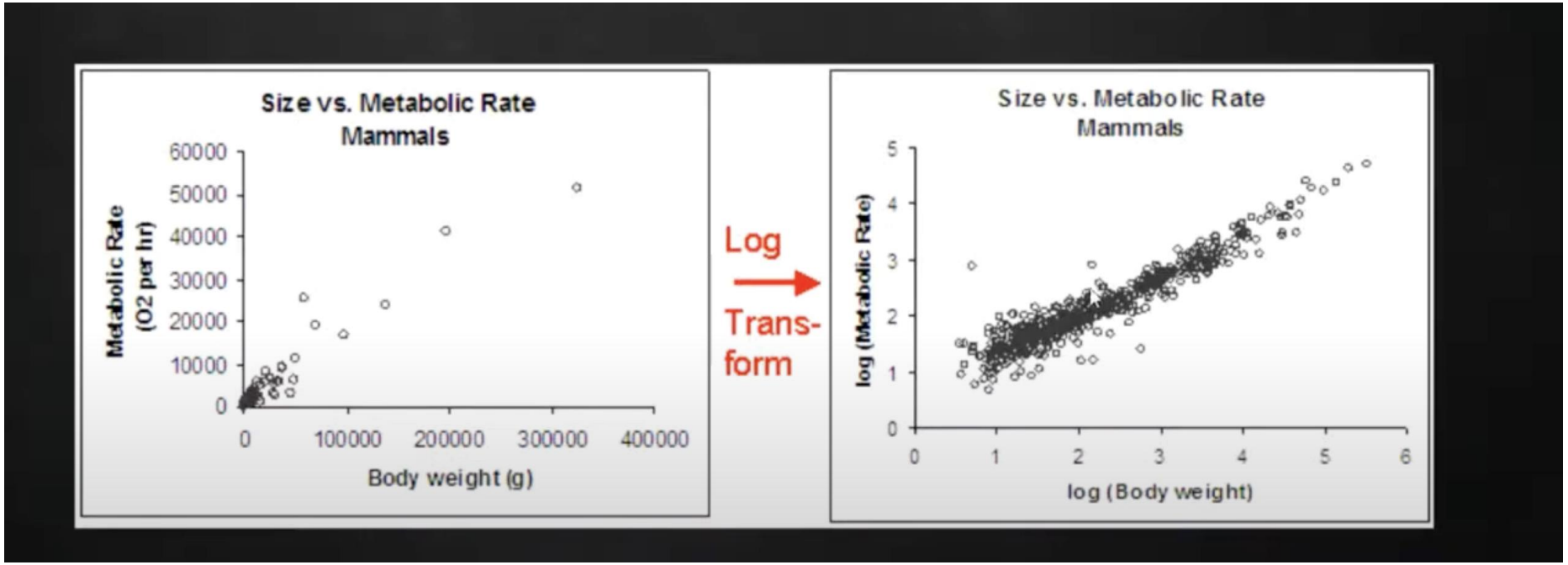
Manteniendo el número de pies cuadrados constante_____

Interpretación de los logaritmos en econometría

A continuación, mostramos una tabla resumen de cómo se calculan e interpretan los logaritmos en un modelo econométrico de regresión.

Modelo	Regresión	Variable Dep. (Y)	Variable Indep. (X)	Interpretación del regresor (β_1)
Nivel - Nivel	$Y_i = \beta_0 + \beta_1 X_i + u_i$	Y	X	$\Delta Y = \beta_1 \Delta X$
Nivel – Log	$Y_i = \beta_0 + \beta_1 \log(X_i) + u_i$	Y	log(X)	$\Delta Y = \left(\frac{\beta_1}{100}\right) \%\Delta X$
Log - Nivel	$\log(Y_i) = \beta_0 + \beta_1 X_i + u_i$	log(Y)	X	$\%\Delta Y = (100\beta_1)\Delta X$
Log-Log	$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + u_i$	log(Y)	Log(X)	$\%\Delta Y = \beta_1 \%\Delta X$

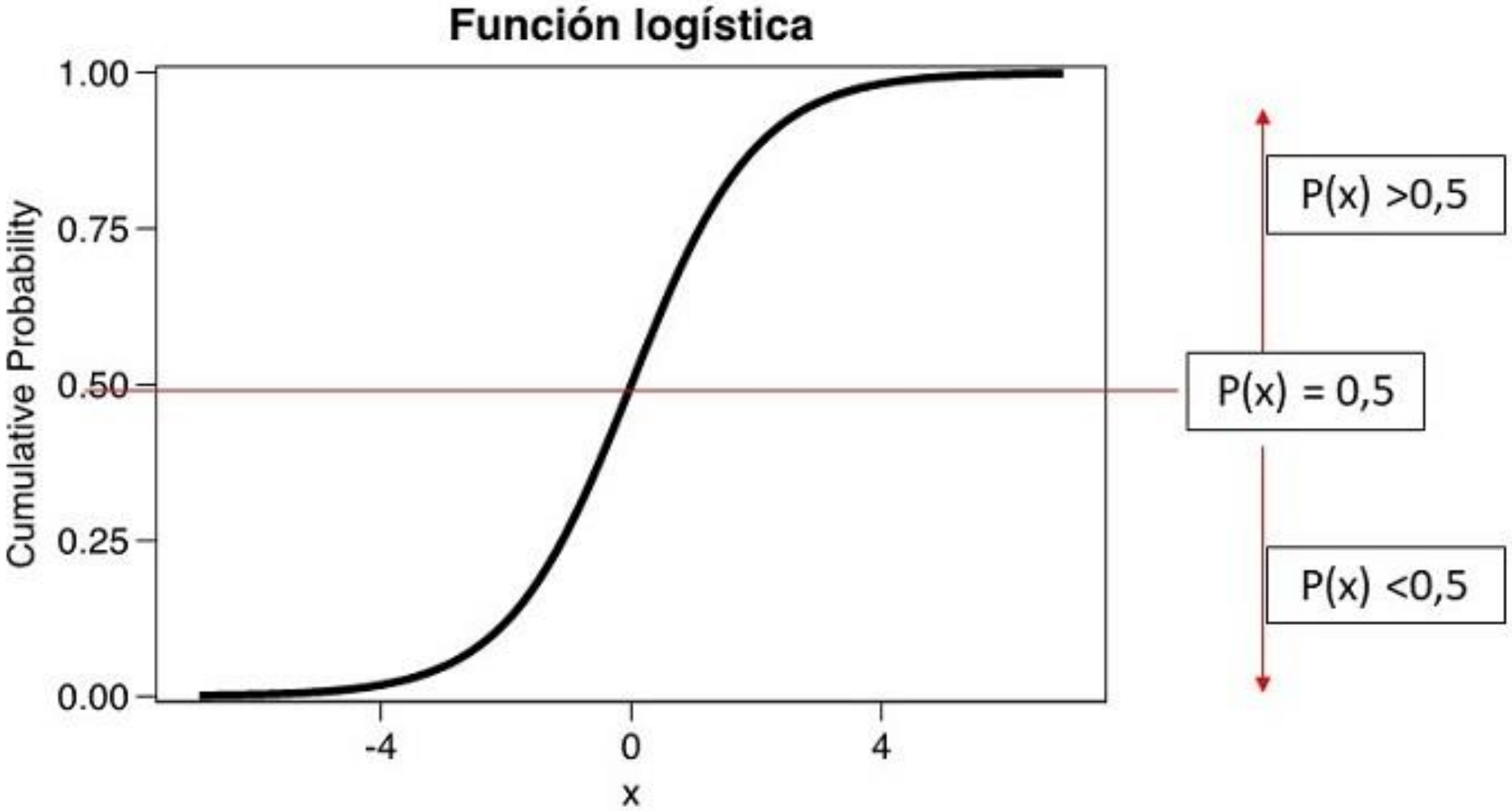
Para qué sirven las transformaciones ?



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

Esto “abre” el rango de 0 a 1 y lo convierte en $(-\infty, +\infty)$.

Para qué sirven las transformaciones ?



Sesgo por variable omitida

Supongamos que tu modelo “verdadero” es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

pero tú estimas solo:

$$Y = \beta_0 + \tilde{\beta}_1 X_1 + u'$$

El sesgo de $\tilde{\beta}_1$ (el estimador de X_1) es:

$$Bias(\tilde{\beta}_1) = \beta_2 \cdot \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)}$$

Sesgo por variable omitida

1. $\beta_2 > 0$ y X_1, X_2 **están positivamente correlacionados** →
   👉 El coeficiente de X_1 queda **sobreestimado**.
   Ejemplo: Publicidad y calidad de producto, ambos suben ventas.
2. $\beta_2 > 0$ y X_1, X_2 **están negativamente correlacionados** →
   👉 El coeficiente de X_1 queda **subestimado**.
   Ejemplo: Publicidad digital y promociones físicas (sustitutos).
3. $\beta_2 < 0$ y X_1, X_2 **están positivamente correlacionados** →
   👉 El coeficiente de X_1 queda **subestimado**.
   Ejemplo: Horas de estudio y horas de trabajo (trabajar baja notas, y trabajar se asocia con estudiar más).
4. $\beta_2 < 0$ y X_1, X_2 **están negativamente correlacionados** →
   👉 El coeficiente de X_1 queda **sobreestimado**.
   Ejemplo: Ejercicio y comida chatarra (menos chatarra al hacer más ejercicio → el efecto del ejercicio parece más grande).

$$Y = \beta_0 + \tilde{\beta}_1 X_1 + u'$$

$$\text{Signo del sesgo} = \text{Signo}(\beta_2) \times \text{Signo}(\text{Cov}(X_1, X_2))$$

Man's search for meaning: The case of Legos

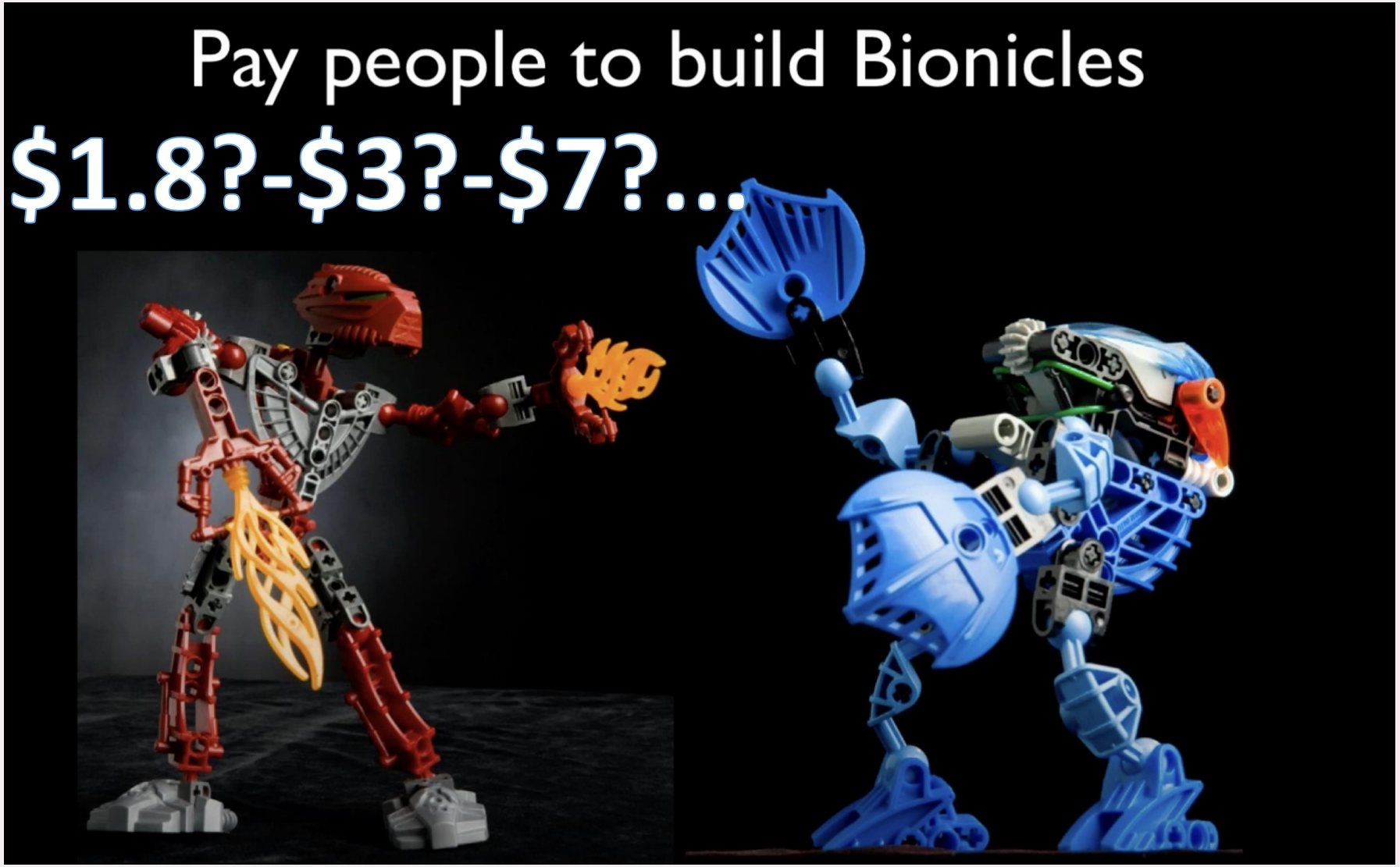
Dan Ariely [a], Emir Kamenica [b,*], Dražen Prelec [a]

[a] *MIT Sloan School of Management, 38 Memorial Drive, Cambridge, MA 02142, United States*
[b] *University of Chicago Graduate School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, United States*

Received 26 December 2005; received in revised form 17 January 2008; accepted 17 January 2008
Available online 24 January 2008

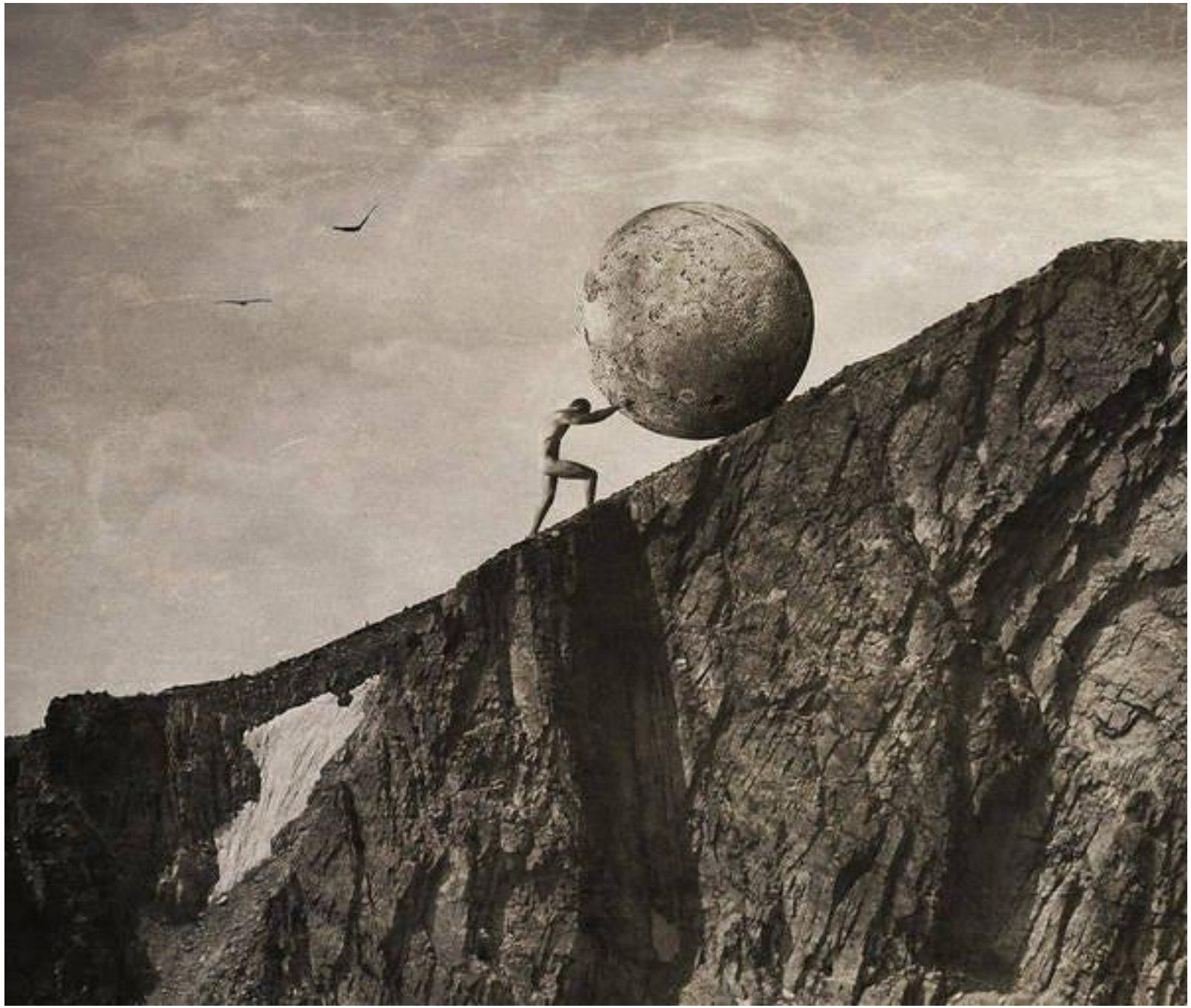
Pay people to build Bionicles

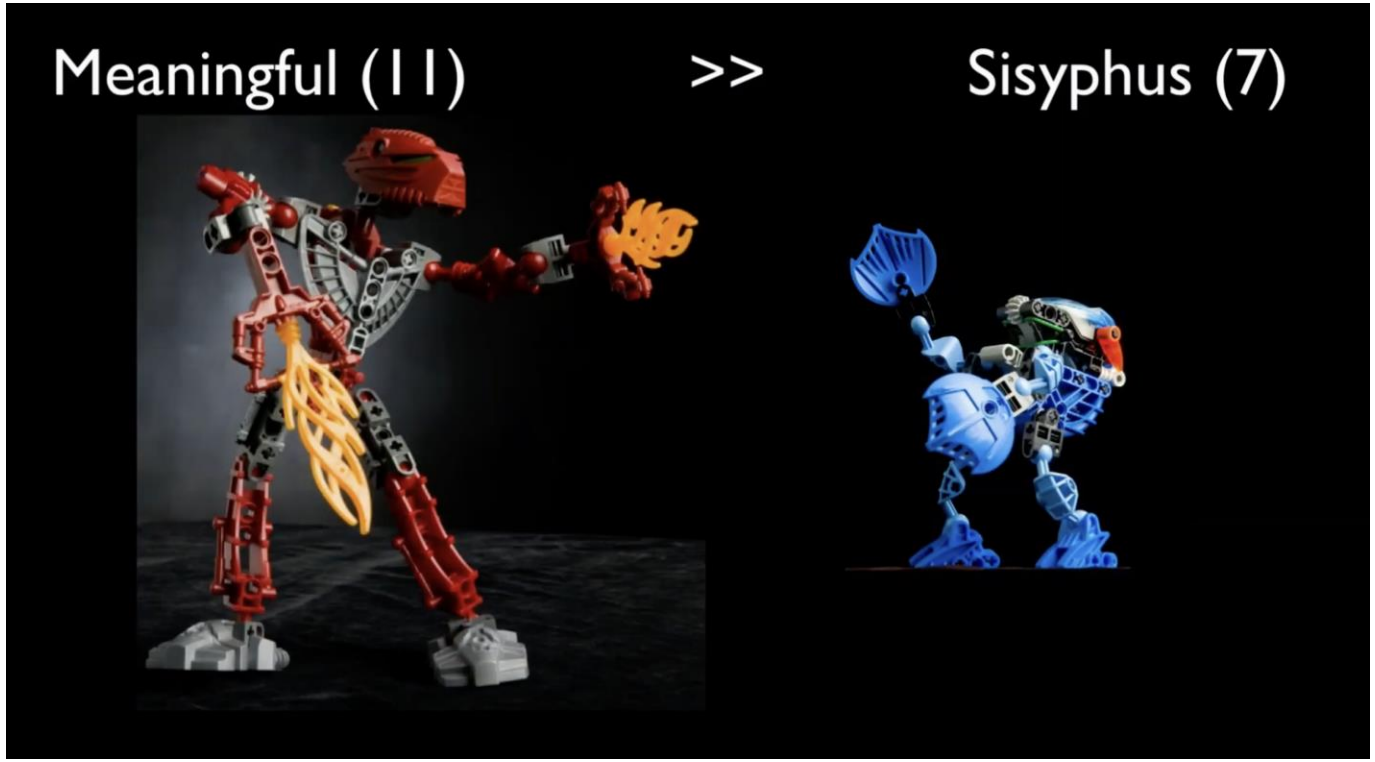
$1.8?-$3?-$7?...



Meaningful

Sisyphic





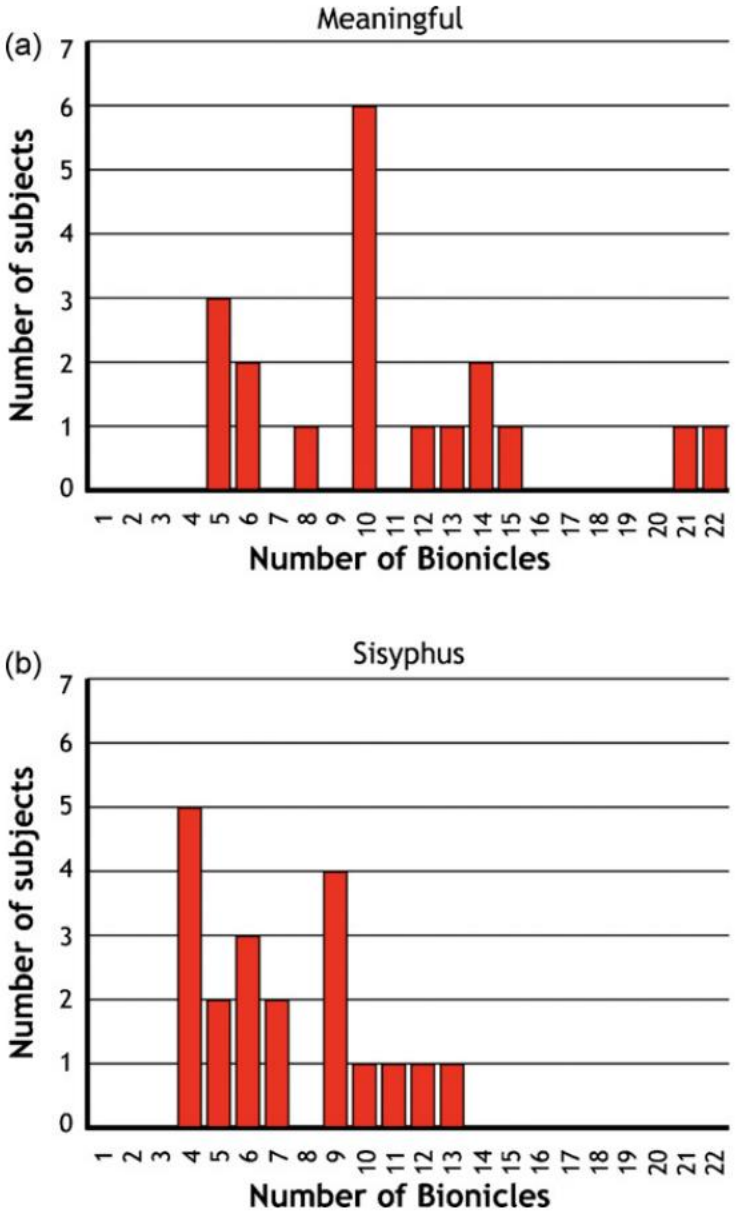
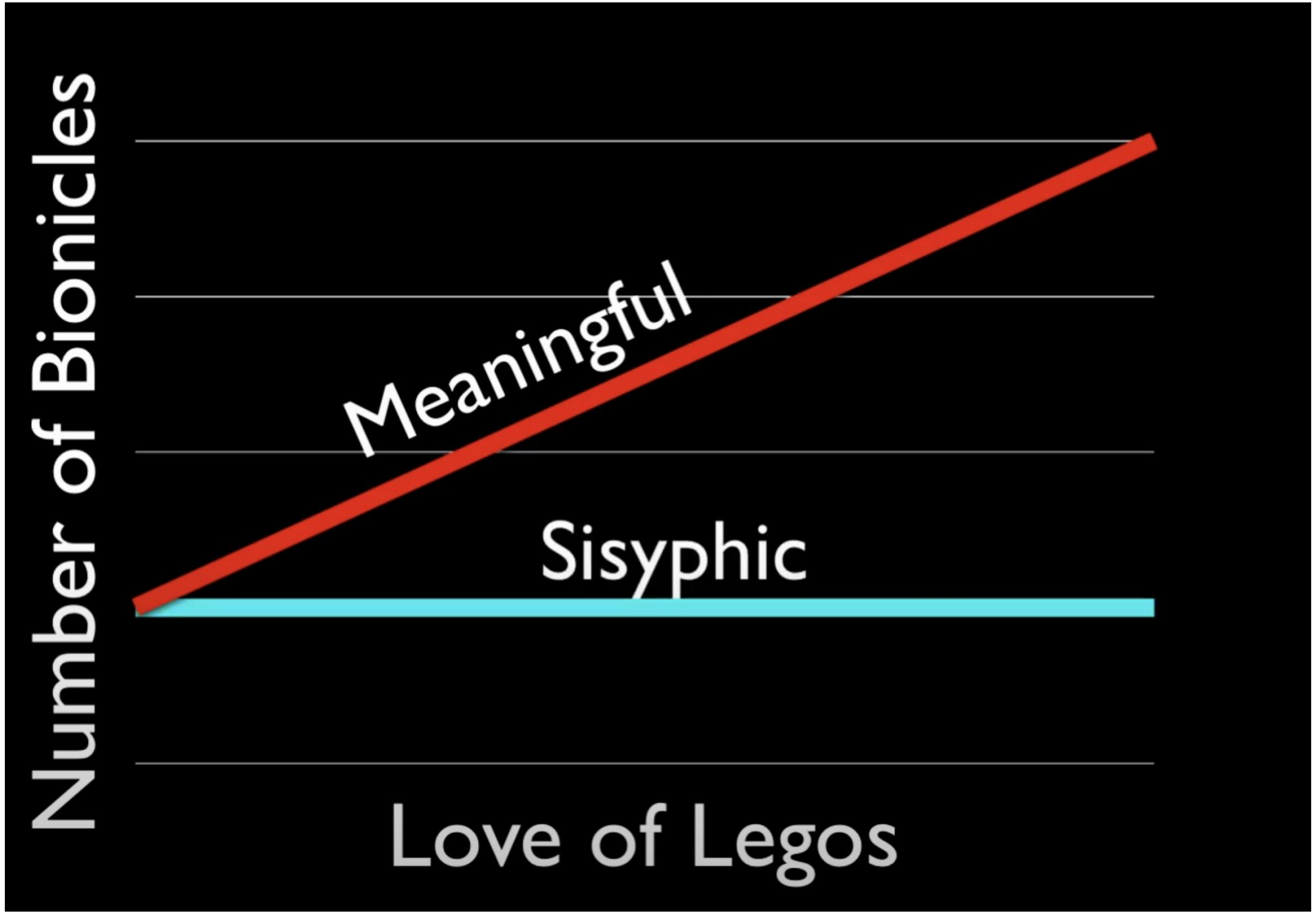


Fig. 2. Number of Bionicles completed in the Meaningful and Sisyphus conditions.


Number of Bionicles
Meaningful
Love of Legos

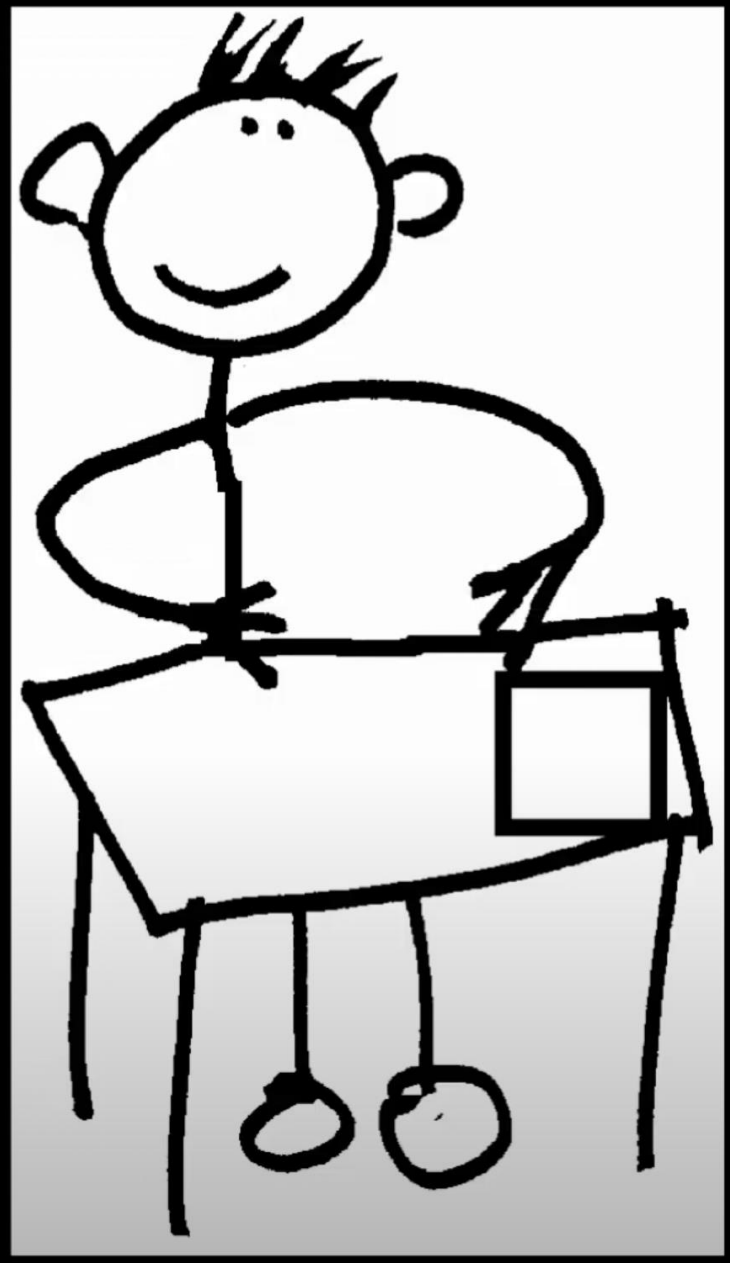

Number of Bionicles
Meaningful
Sisyphic
Love of Legos

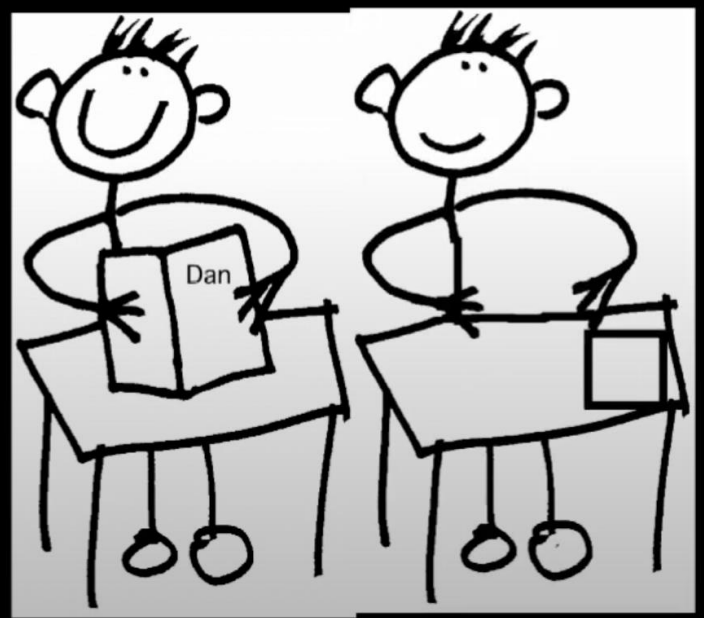
Paperwork

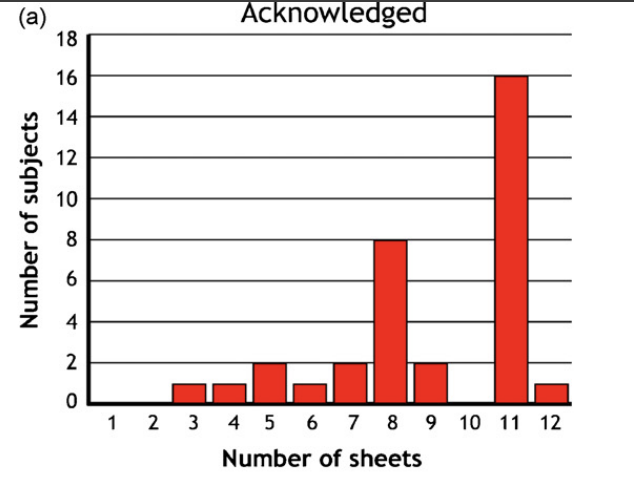
kassfadkaakssakska
asasadasdasasassdda
adaadadasdadassdas
lawekwelwexeqissxq
lekxekqwsddasksskd
aekakdssfowekqwek
qlwkqwssweokmghsa
qlwssqwqkwkwwew
wsswewqweqwqwe

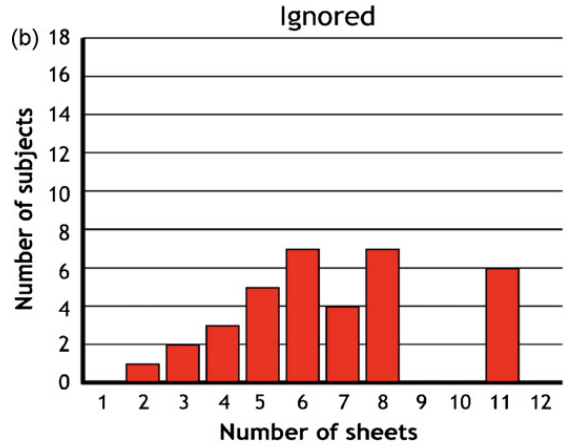

Dan

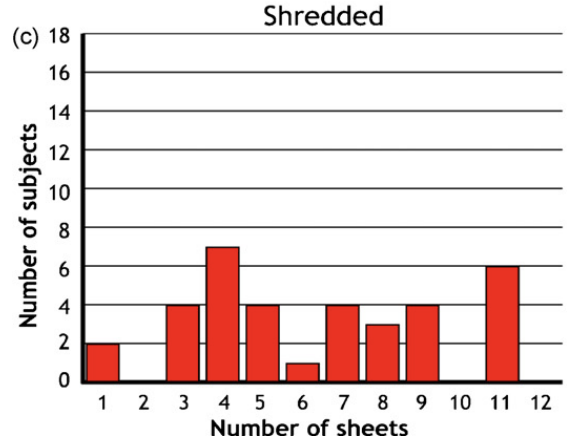

Dan




Dan


(a)
Acknowledged
Number of subjects
0
2
4
6
8
10
12
14
16
18
1 2 3 4 5 6 7 8 9 10 11 12
Number of sheets


(b)
Ignored
Number of subjects
0
2
4
6
8
10
12
14
16
18
1 2 3 4 5 6 7 8 9 10 11 12
Number of sheets


(c)
Shredded
Number of subjects
0
2
4
6
8
10
12
14
16
18
1 2 3 4 5 6 7 8 9 10 11 12
Number of sheets


Reservation wage in ¢
30
22.5
15
7.5
0
Acknowledged
Ignored
Shredded


Reservation wage in ¢
30
22.5
15
7.5
0
Acknowledged
Ignored
Shredded

Manos a la obra!

📘 Tarea– Caso Tour Marketing (Opry Dataset)

Objetivo

Explorar la relación entre la **inversión en publicidad (Gasto_Publicidad)** y las **ventas (Ventas)** de Opry, estimar modelos de regresión, interpretar los resultados y reflexionar sobre las limitaciones del análisis.

Ir al repo:
https://github.com/Nicolas-Velasquez-Oficial/business-analytics-2025-3/tree/main/Casos/Tours_Marketing