Analítica de los Negocios

Intro a la Regresión Lineal

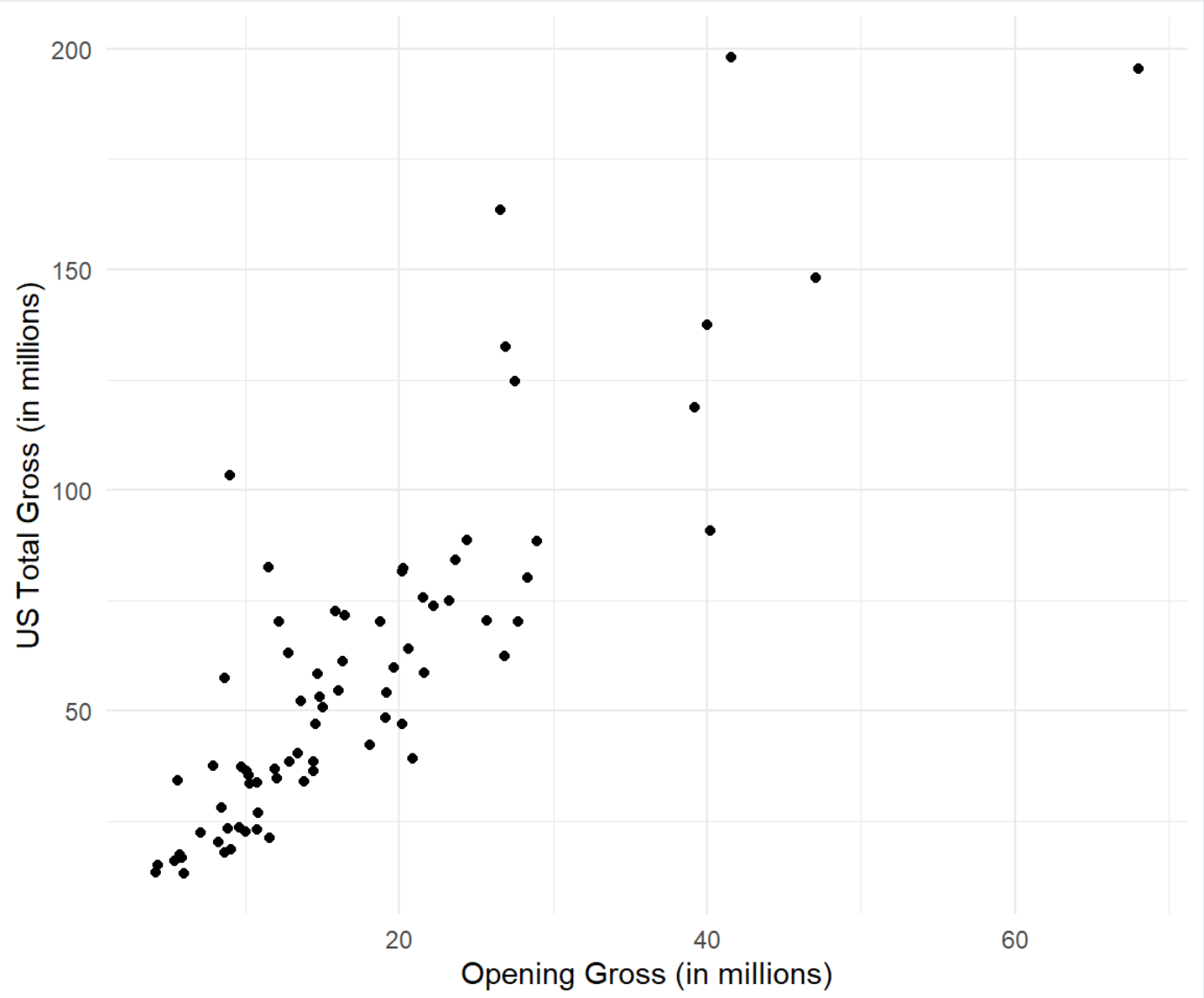
Carlos Cardona Andrade

Plan para hoy

1. Visualizar los Datos
2. Regresión Lineal Simple
3. Regresión Lineal Múltiple

Visualizar los Datos

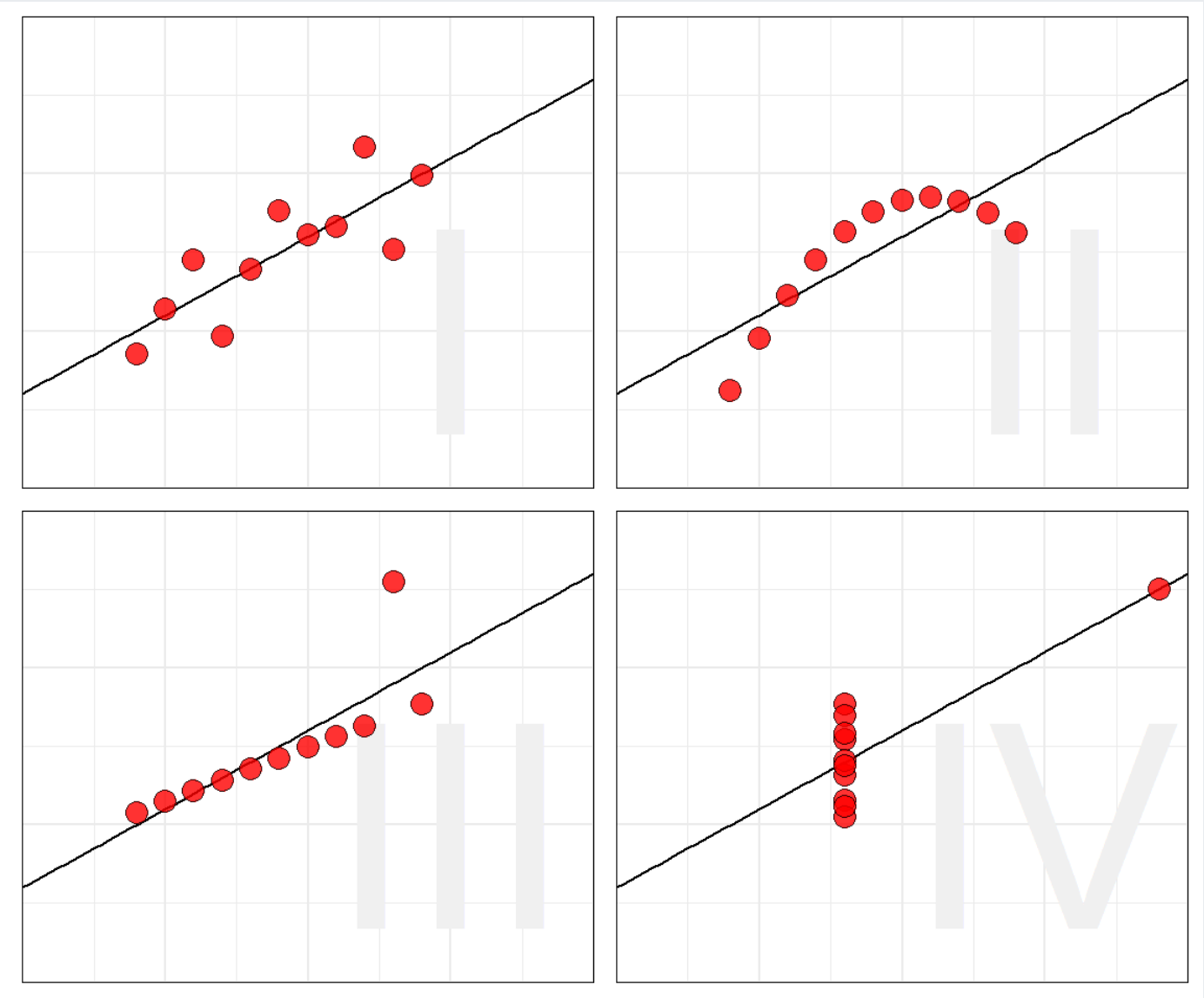
¿Por qué es importante graficar los datos?



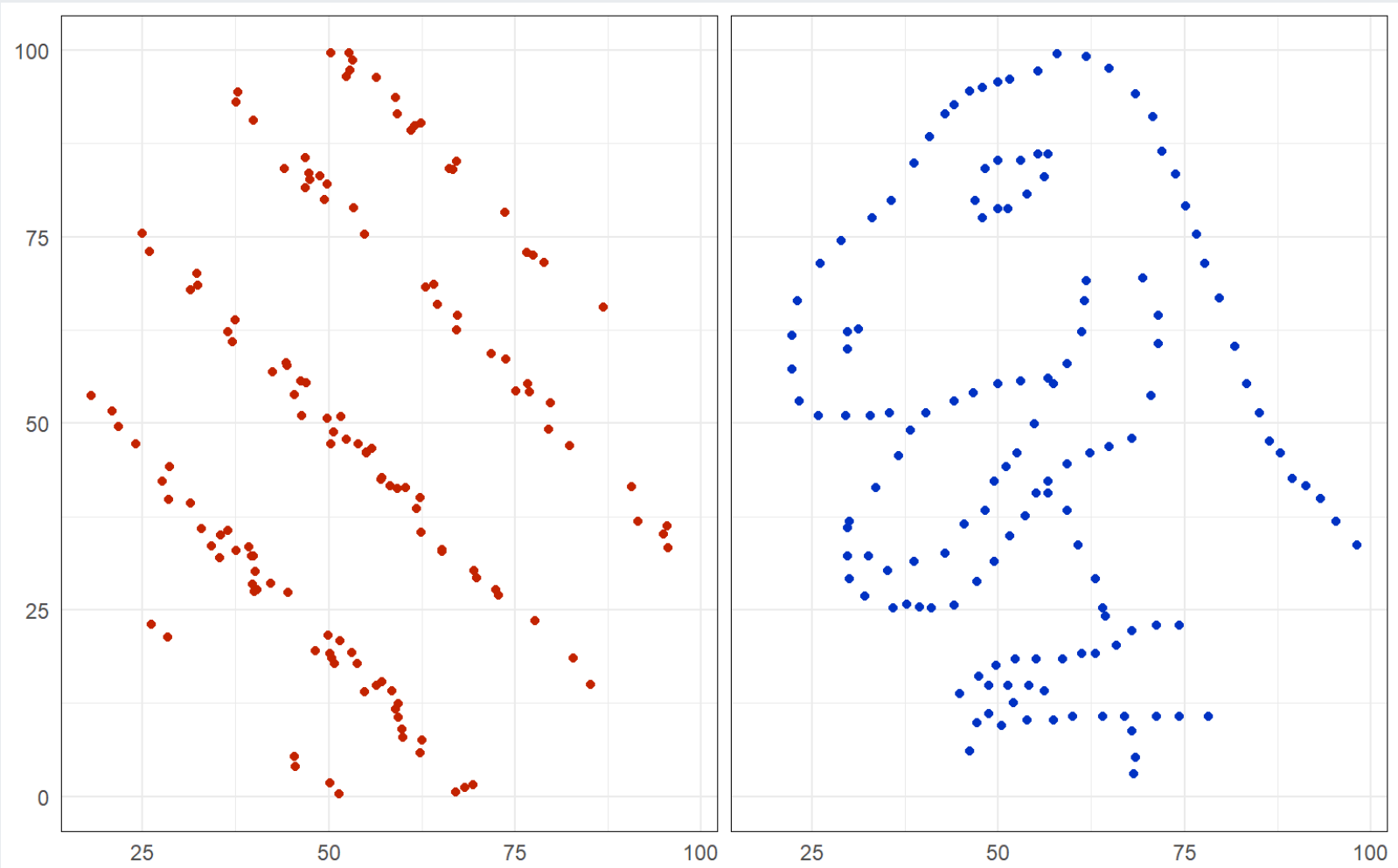
Cuarteto de Anscombe

	I		II		III		IV	
Obs.	**X**	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**
1	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
2	8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
3	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
4	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
5	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
6	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
7	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
8	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
9	12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
10	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
11	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
SD	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
Corr	0.82	NA	0.82	NA	0.82	NA	0.82	NA

¿Qué aprendemos cuando graficamos los datos?



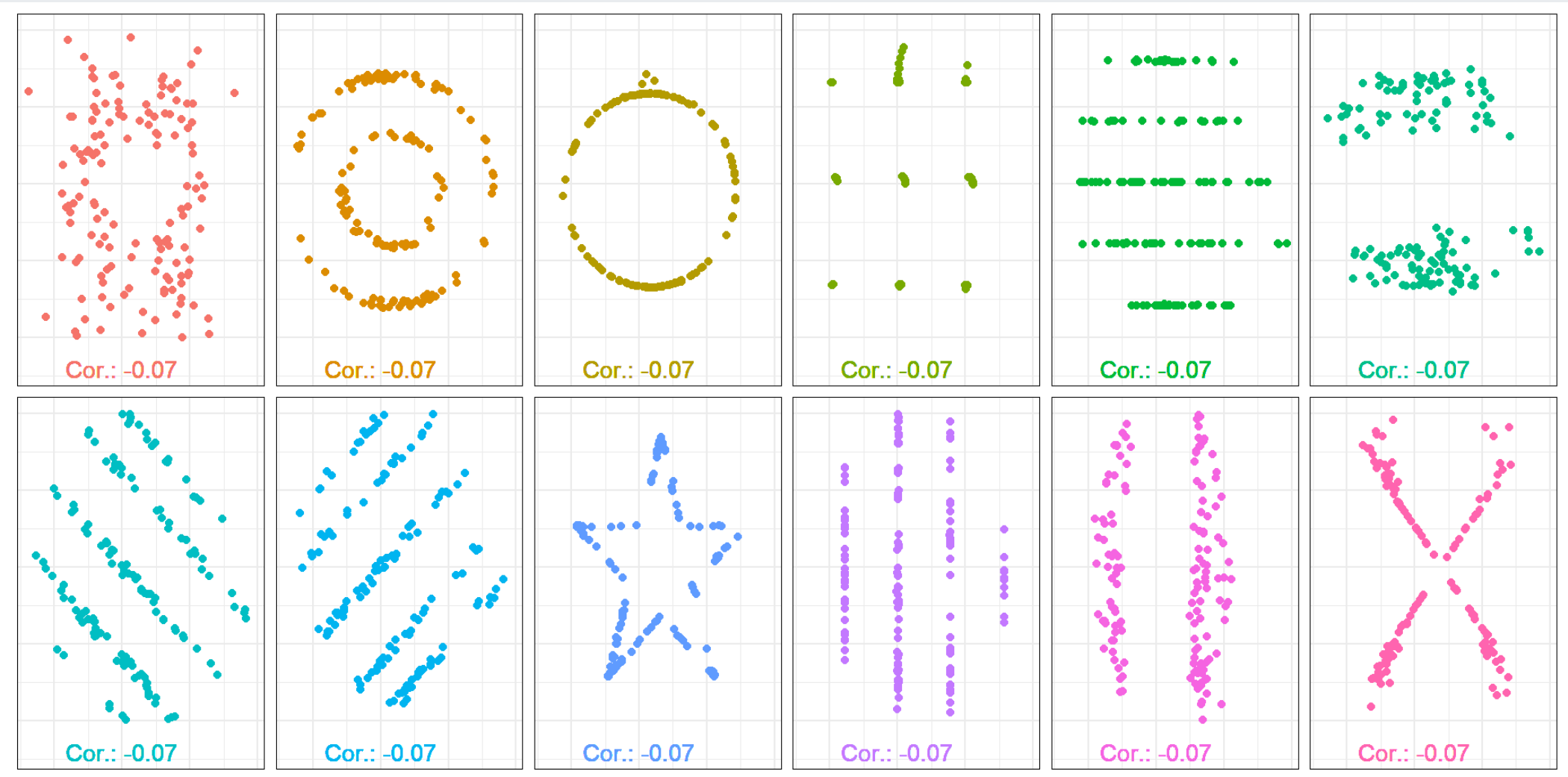
¿Ven correlación acá?



Correlación: -0.07

Correlación: -0.07

Iguales pero diferentes



¿Por qué es importante graficar los datos?

- Visualización de patrones: Los gráficos permiten identificar tendencias, patrones y relaciones entre variables que pueden no ser evidentes solo con los números

- Análisis visual vs. numérico: Aunque las estadísticas (media, desviación estándar, correlación) pueden ser iguales entre los conjuntos de datos, los gráficos muestran que los datos pueden comportarse de manera muy diferente
  - Ejemplo: Cuarteto de Ascombe

- En conclusión, la visualización de datos no solo facilita la comprensión de la información, sino que también ayuda a evitar interpretaciones erróneas basadas únicamente en estadísticas numéricas

Ejercicio 1

1. Carguen el paquete `tidyverse`, `janitor` y `corrplot` a R.
2. Ejecuten el siguiente código para importar los datos del caso *Hollywood Rules*:

```
hollywood <- read_excel("KEL702-XLS-ENG.xls", sheet = "Exhibit 1")
```

3. Limpiemos los nombres con el paquete `janitor` para que sean legibles:

```
hollywood <- hollywood |>
  clean_names()
hollywood <- hollywood |>
  rename(us_gross = total_u_s_gross,
         non_us_gross = total_non_u_s_gross)
```

Ejercicio 1

4. Creen un nuevo conjunto de datos que se llame `hollywood_sub` que *sólo* contenga las variables de recaudo en USA, el recaudo en el estreno, el recaudo en el mercado internacional y el presupuesto en **ese orden**
5. Grafiquemos la matriz de correlación con el siguiente código:

```
# Matriz de correlaciones
cor_matrix <- cor(hollywood_sub)

# Renombrar columnas y filas en la matriz
colnames(cor_matrix) <- c("US Gross", "Opening Gross", "Non-US Gross", "Bud
rownames(cor_matrix) <- c("US Gross", "Opening Gross", "Non-US Gross", "Bud

# Crear el gráfico de correlaciones
corrplot(cor_matrix, method = "number", type = "upper")
```

Ejercicio 1

6. Analicen las dos variables que tengan la correlación más alta y las que tengan la correlación más baja.
7. Creen el gráfico de dispersión entre el recaudo total y el recaudo en el día del estreno como en la diapositiva 4. ¿Es consistente el gráfico con el valor de la correlación?

Regresión Lineal Simple

¿Por qué una regresión?

Ejemplo

El promedio en la universidad es resultado de la habilidad y las horas estudiadas. Por lo tanto, uno podría pensar en el siguiente modelo:

$$Promedio = f(H, Saber11, PCA)$$

Donde H es la habilidad, $Saber11$ es el puntaje en la prueba Saber 11 y PCA es el porcentaje de clases a las que el estudiante fue. Esperamos que el $Promedio$ aumente cada que alguna de esas 3 variables aumente.

Sin embargo, no necesitamos esperar. Podemos evaluar esta hipótesis usando un **modelo de regresión!**

¿Por qué una regresión?

Modelo de Regresión:

$$Promedio_i = \beta_0 + \beta_1 H_i + \beta_2 Saber11_i + \beta_3 PCT_i + e_i$$

Queremos estimar/evaluar la relación

$$Promedio = f(H, Saber11, PCA)$$

Preguntas a responder:

- ¿Cómo se interpretan β_0, β_1, β_2 y β_3?
- ¿Son los términos β_k parámetros o son estimaciones muestrales?
- ¿Qué es e_i?

Partes esenciales de una regresión

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

y

- Variable dependiente
- Variable resultado
- Básicamente lo que queremos predecir o explicar

x

- Variable explicativa
- Predictor
- Variable independiente
- Lo que usamos para predecir o explicar y

Objetivos de una regresión

Usualmente ajustamos a una línea por dos razones:

Predicción

- Predecir el futuro
- Nos enfocamos en y
- Netflix tratando de predecir la siguiente serie que veremos

Explicación

- Explicar el efecto de x en y
- Nos enfocamos en x
- Netflix evaluando el efecto de la hora del día en la selección de una serie

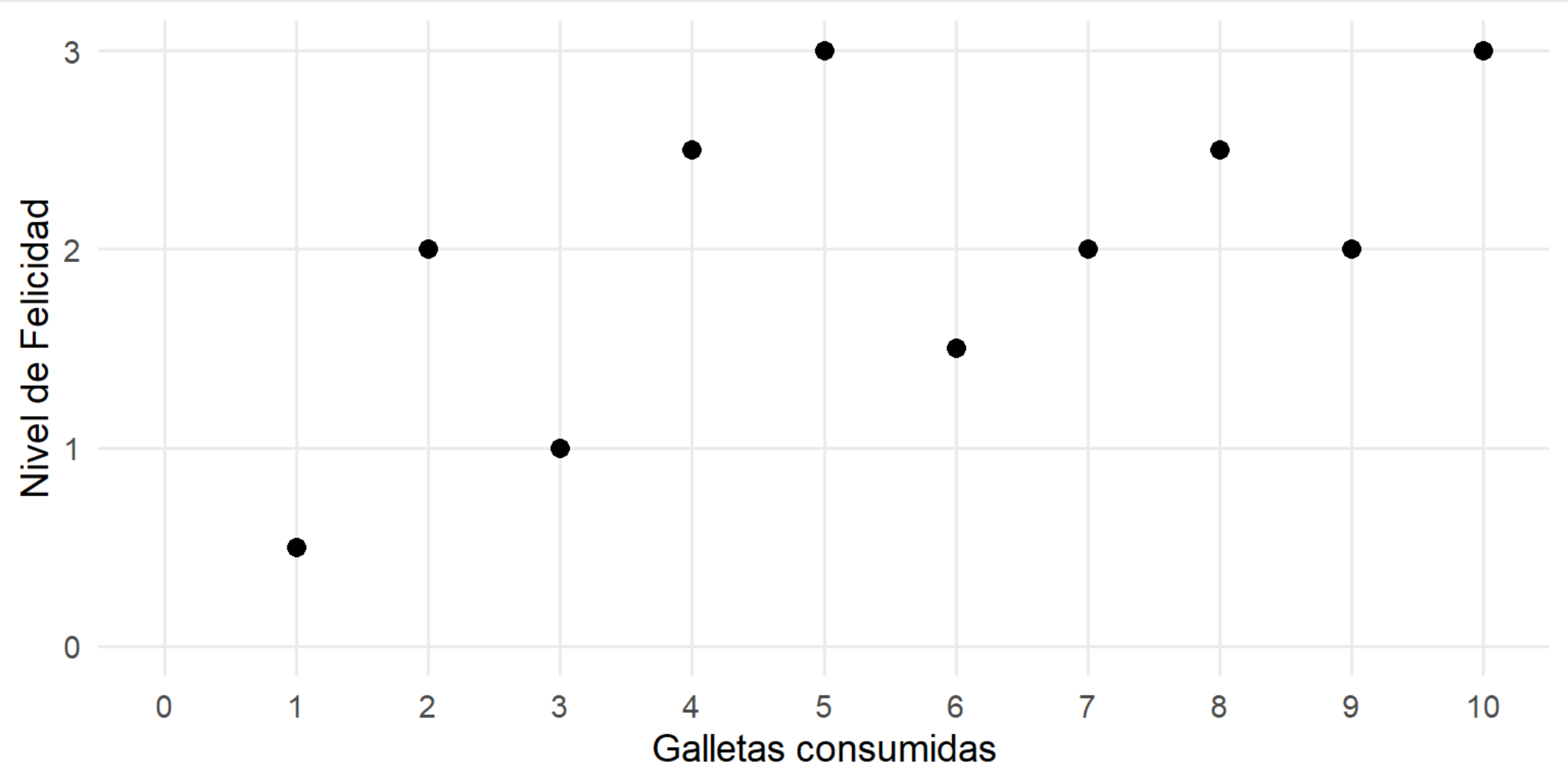
¿Cómo se estima la línea de regresión?

1. Graficar ambas variables y y x
2. Dibujar una recta que se aproxime a la relación observada (ojalá funcione para datos que no están en la muestra)
3. Estimar los números que componen esa recta
4. Interpretar esos números

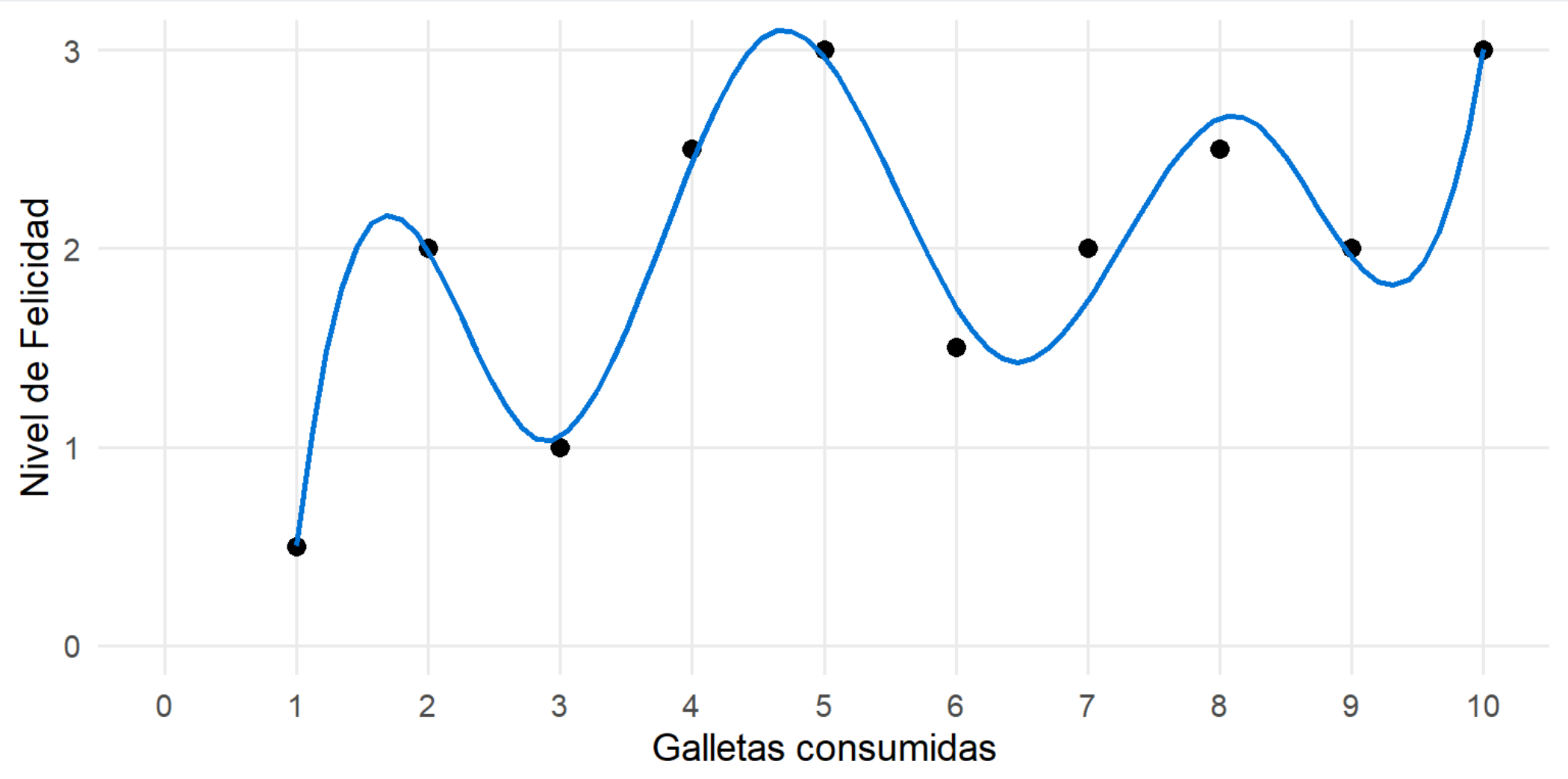
Galletas y Felicidad

felicidad	galletas
0.5	1
2.0	2
1.0	3
2.5	4
3.0	5
1.5	6
2.0	7
2.5	8
2.0	9
3.0	10

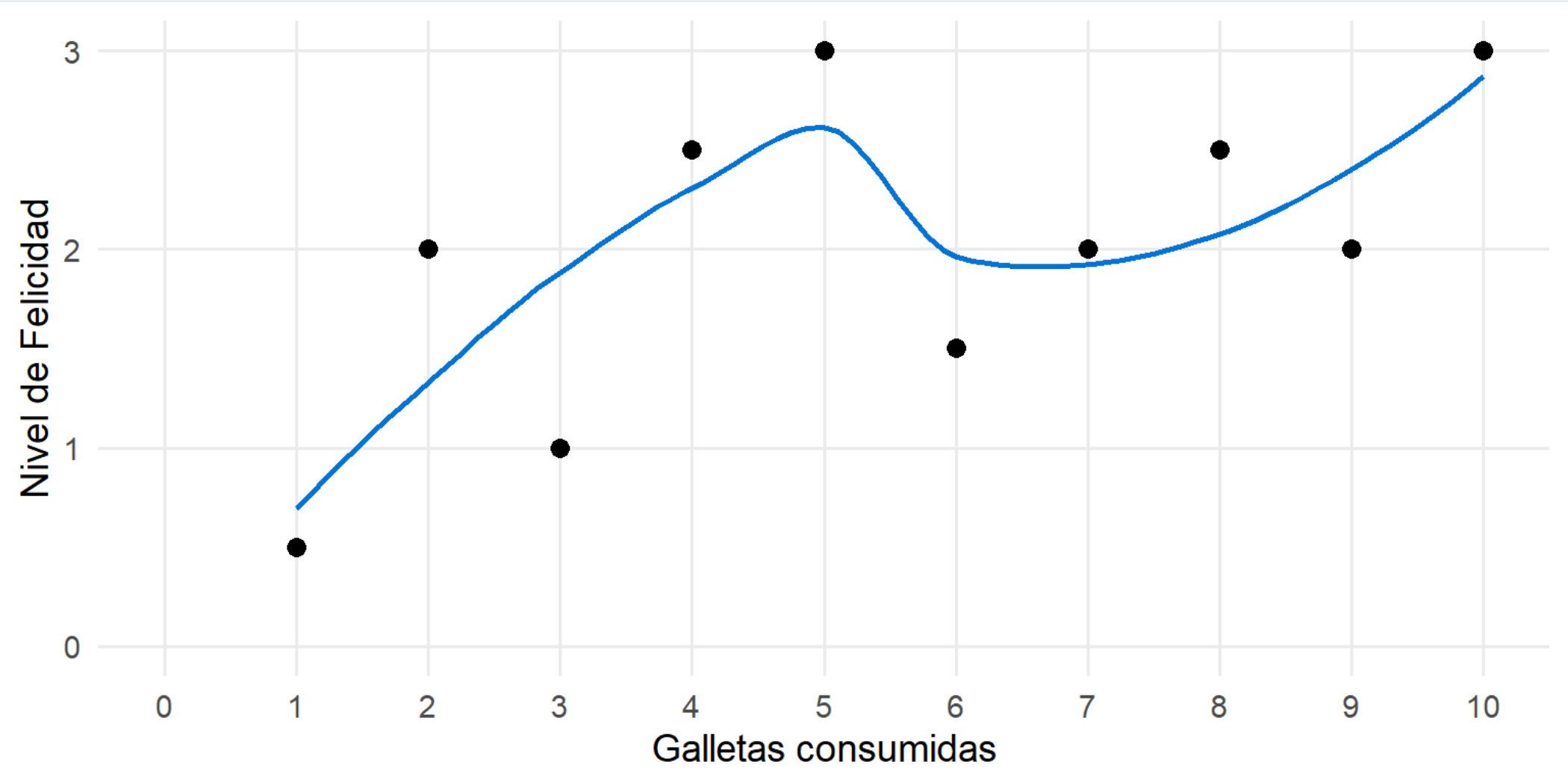
¿Cómo se relacionan las galletas con la felicidad?



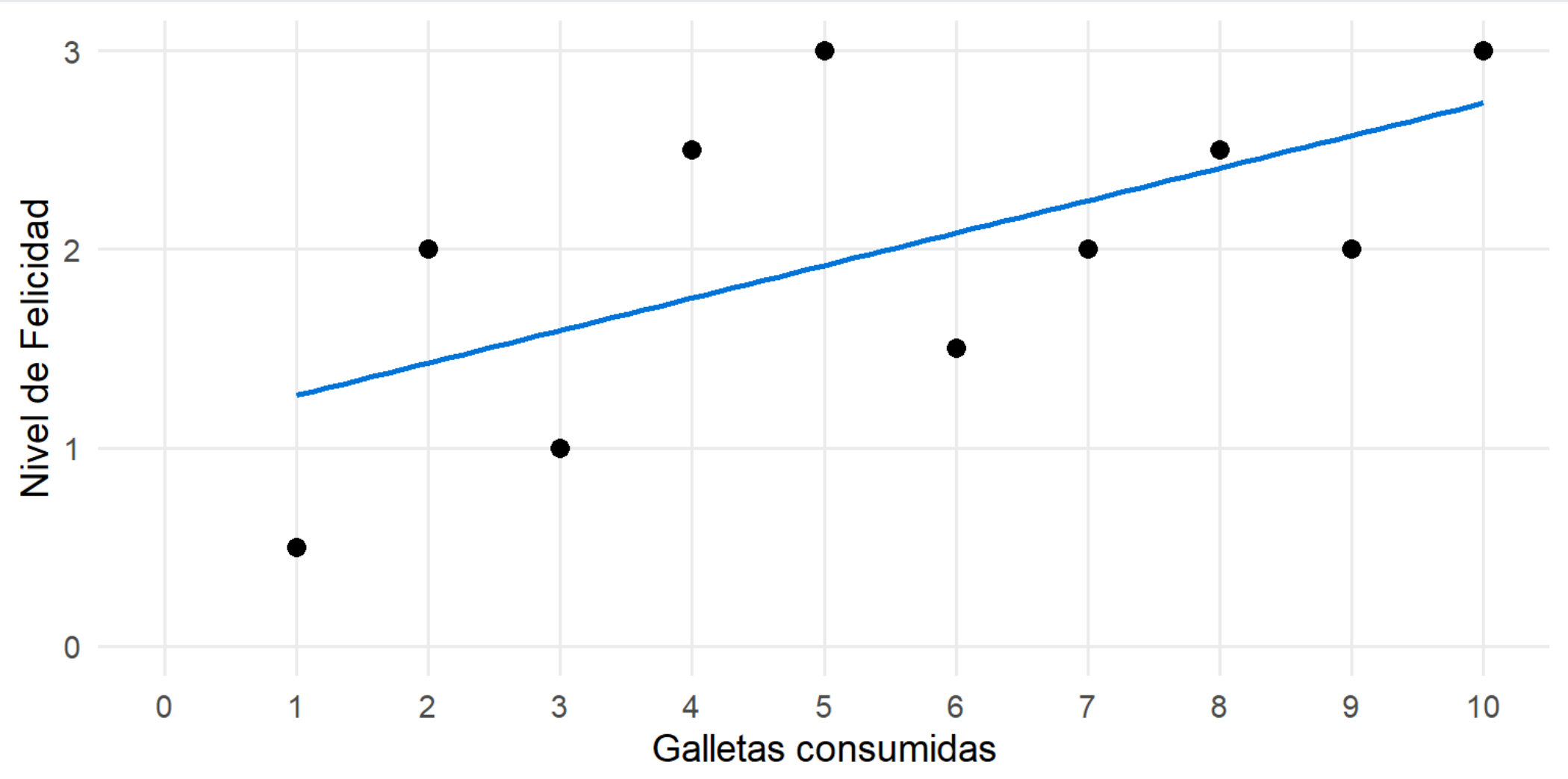
Trazando una curva que sigue los datos



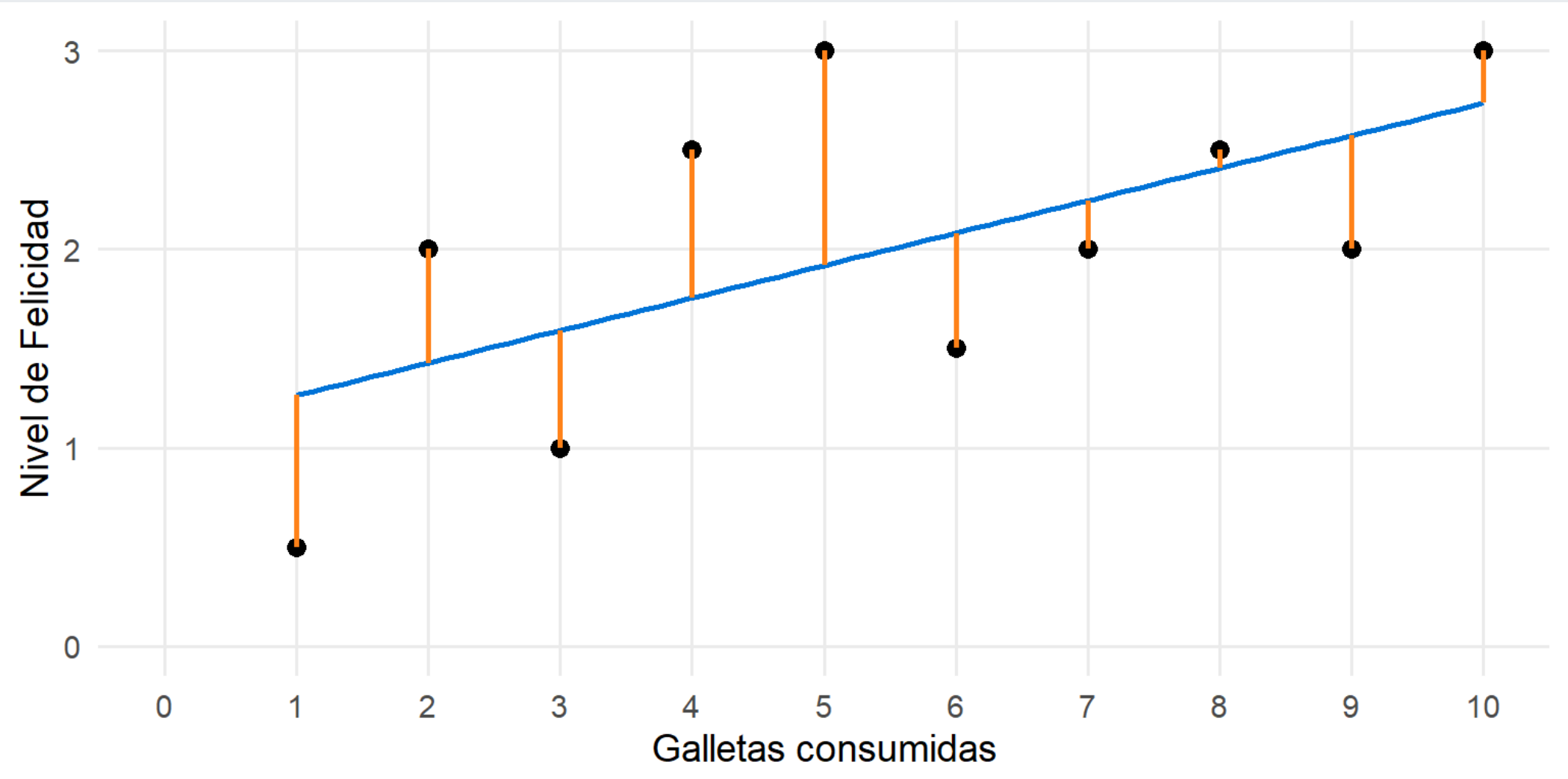
Otra línea que se adapta a los datos



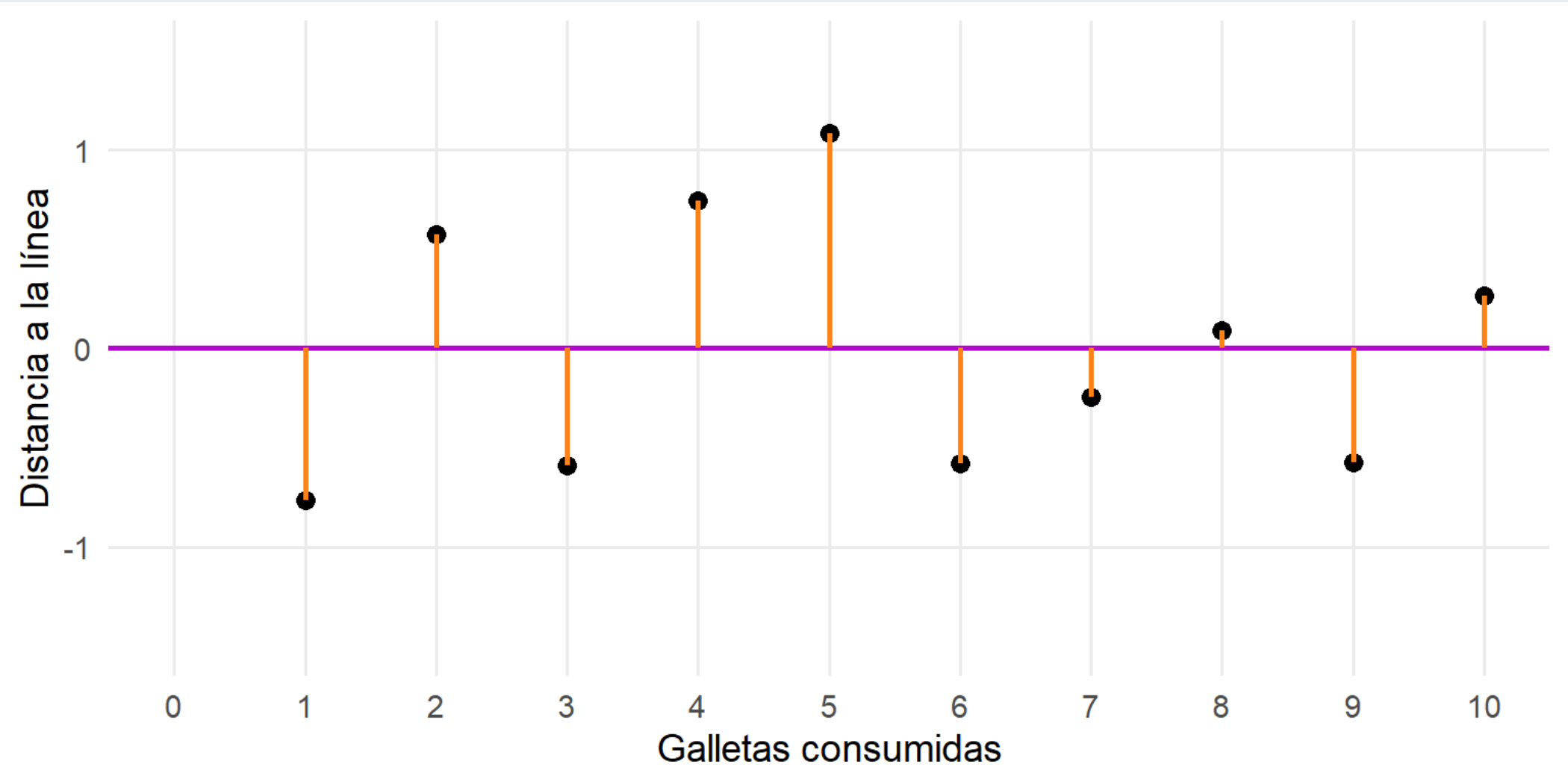
Ajuste con una línea recta (Regresión lineal)

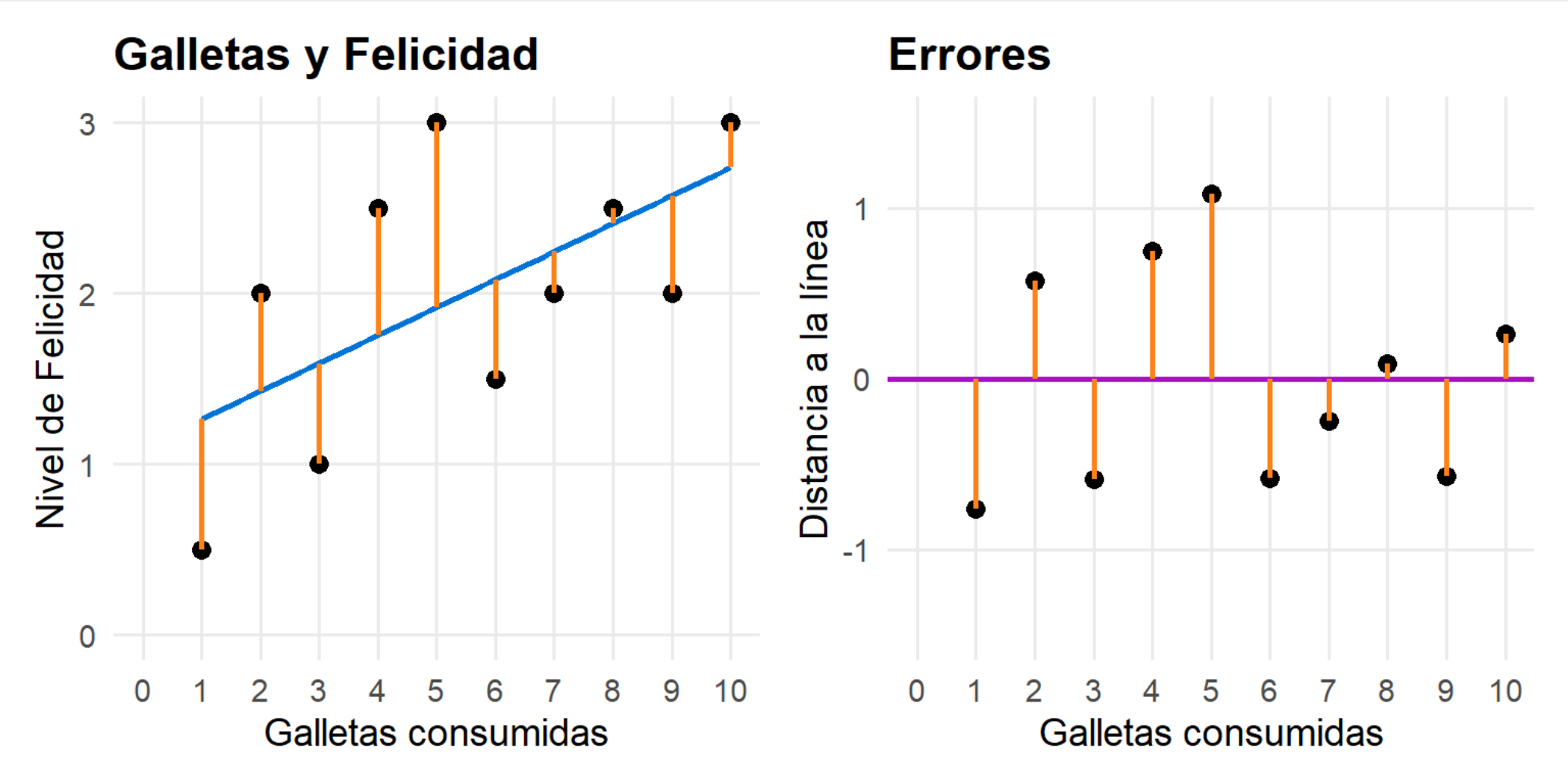


Errores de predicción



Visualización de los errores




Galletas y Felicidad
Nivel de Felicidad
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
Galletas consumidas
Errores
Distancia a la línea
1
0
-1
0 1 2 3 4 5 6 7 8 9 10
Galletas consumidas

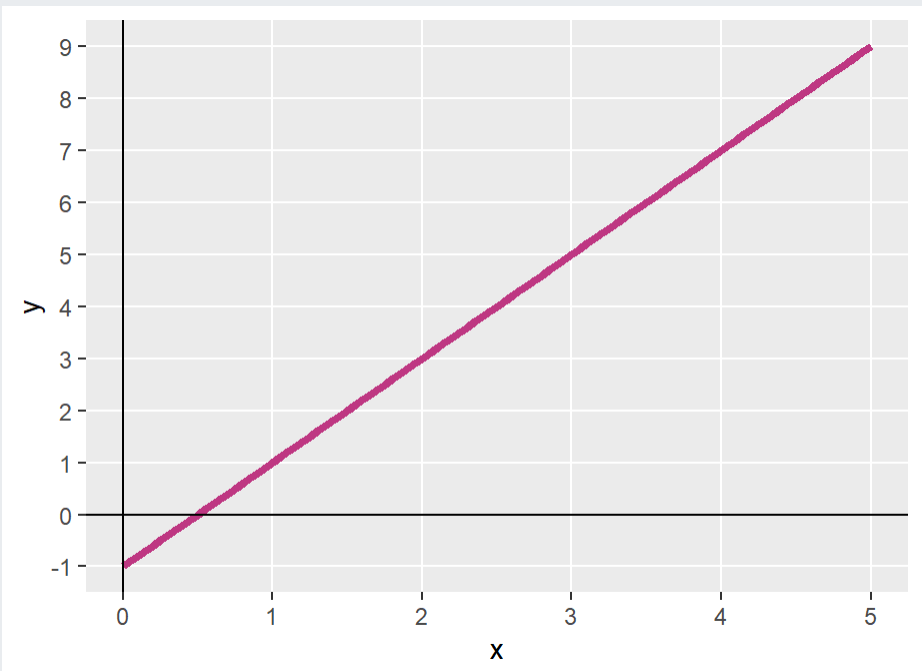
Pendiente de una recta

$$y = mx + b$$

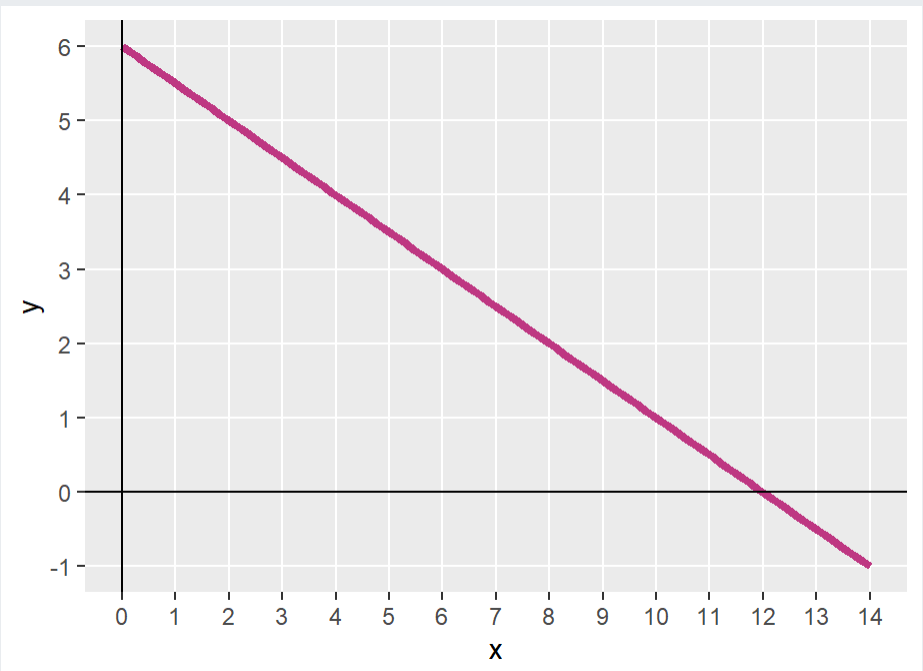
- y es un número
- x es un número
- m es la pendiente $\frac{\Delta y}{\Delta x}$
- b es el intercepto con y

Pendiente de una recta: Ejemplos

$$y = 2x - 1$$



$$y = -0.5x + 6$$



Partes esenciales de una regresión II

$$y_i = \beta_0 + \beta_1 x_i + e$$

- β_1: la pendiente verdadera de la relación entre x y y
- β_0: el intercepto verdadero de la relación entre x y y
- e: el error

Partes esenciales de una regresión II

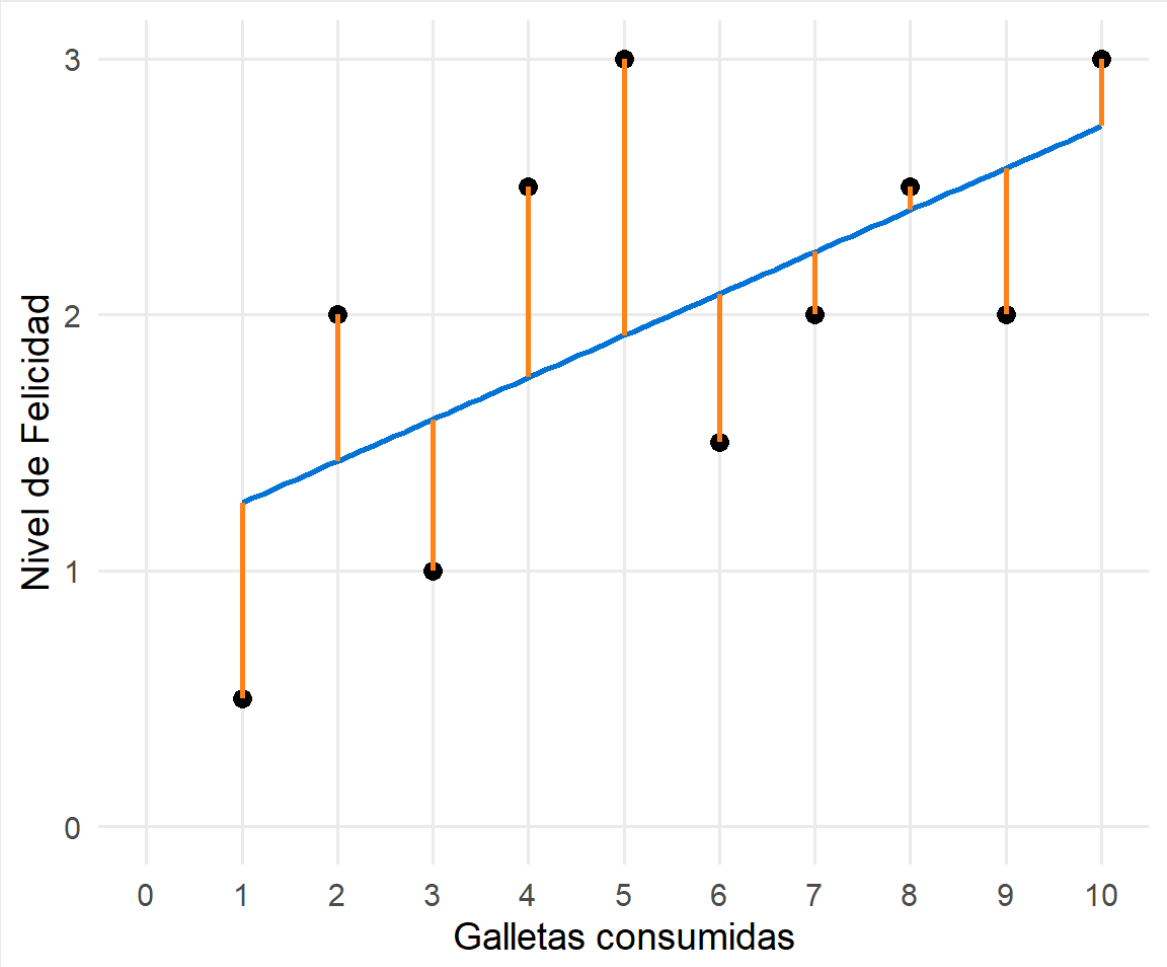
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{y}$: es la estimación de la variable dependiente dado nuestro modelo
- $\hat{\beta}_1$: la pendiente estimada de la relación entre x y y
- $\hat{\beta}_0$: el intercepto estimada de la relación entre x y y
- No hay error!!!

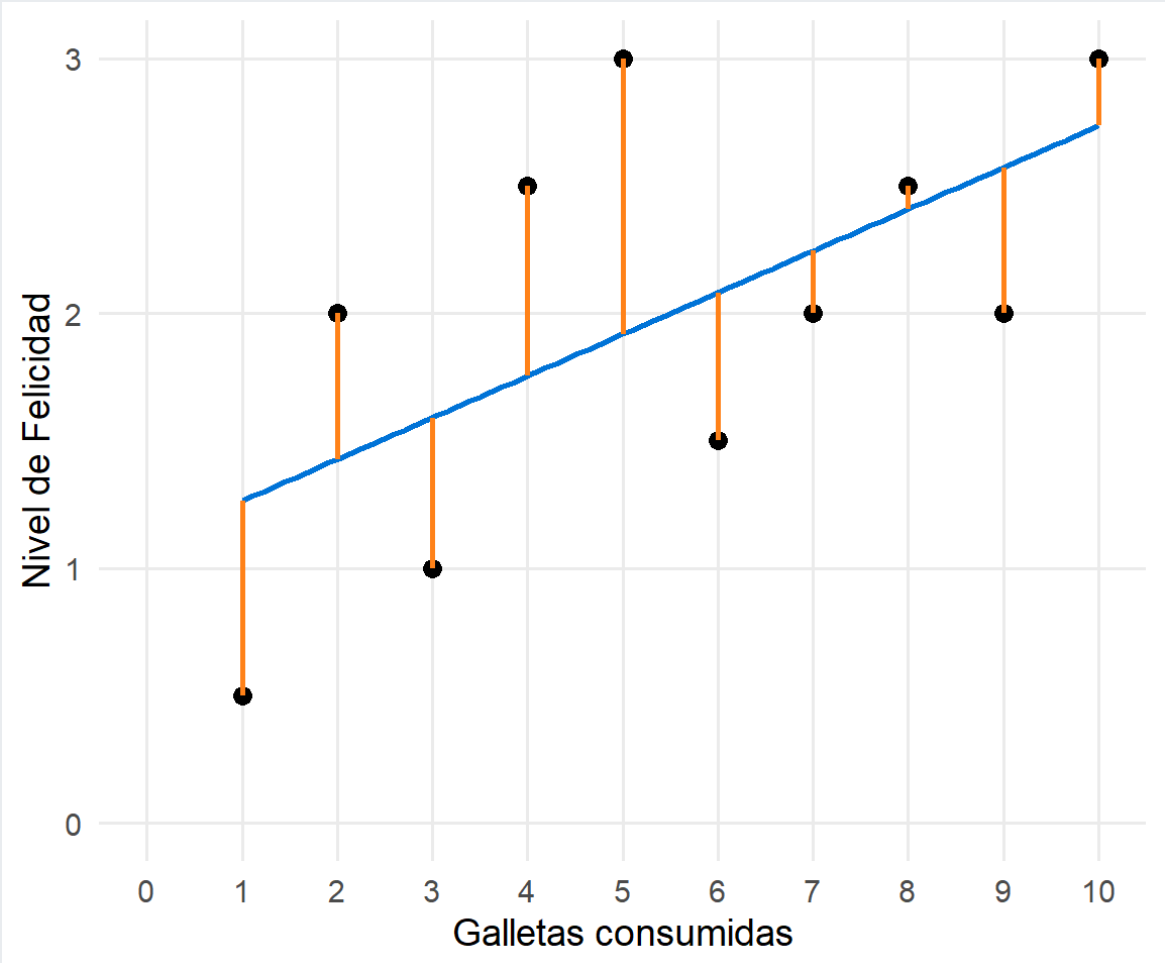
- El símbolo *gorro* ($\hat{\beta}_k$ o $\hat{y}$) significa que es una estimación y no un parámetro real

Modelo de Regresión

$$
\begin{aligned}
y &= \text{Modelo} + \text{Error} \\
&= f(x) + e
\end{aligned}
$$



Los errores



$$\text{Error} = \text{Observado} - \text{Predicho} = y - \hat{y}$$

La línea de los mínimos cuadrados ordinarios (MCO)

- El error para la observación i es:

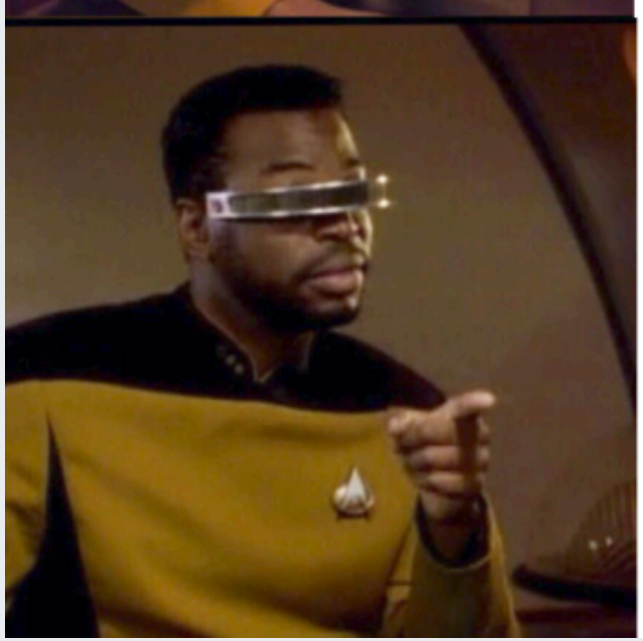
$$e_i = \text{Observado} - \text{Predicho} = y_i - \hat{y}_i$$

- La **suma de los residuos al cuadrado** es:

$$e_1^2 + e_2^2 + e_3^2 + .. + e_n^2$$

- La **línea de los mínimos cuadrados ordinarios** es la que minimiza la suma de los residuos al cuadrado:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum e_i^2$$



Estimación [editar]

Supongamos que b es un valor de "candidato" para el parámetro β. La cantidad $y_i - x_i'b$ se denomina residual para la i-ésima observación, mide la distancia vertical entre el punto de datos (x_i, y_i) y el hiperplano $y = x'b$, y por lo tanto se determina el grado de ajuste entre los datos reales y el modelo. La suma de cuadrados de los residuos (SSR) (también llamada la suma de cuadrados del error (ESS) o suma residual de cuadrados (RSS))[3] es una medida del ajuste del modelo general:

$$S(b) = \sum_{i=1}^{n} (y_i - x_i'b)^2 = (y - Xb)^T (y - Xb),$$

donde T denota la matriz de transposición . El valor de b que minimiza esta suma se llama el estimador MCO de β. La función S (b) es cuadrática en b con definida positiva de Hesse , y por lo tanto esta función posee un mínimo global único en $b = \hat{\beta}$, Que puede ser dada por la fórmula explícita:[4]

$$\hat{\beta} = \arg\min_{b \in \mathbb{R}^p} S(b) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

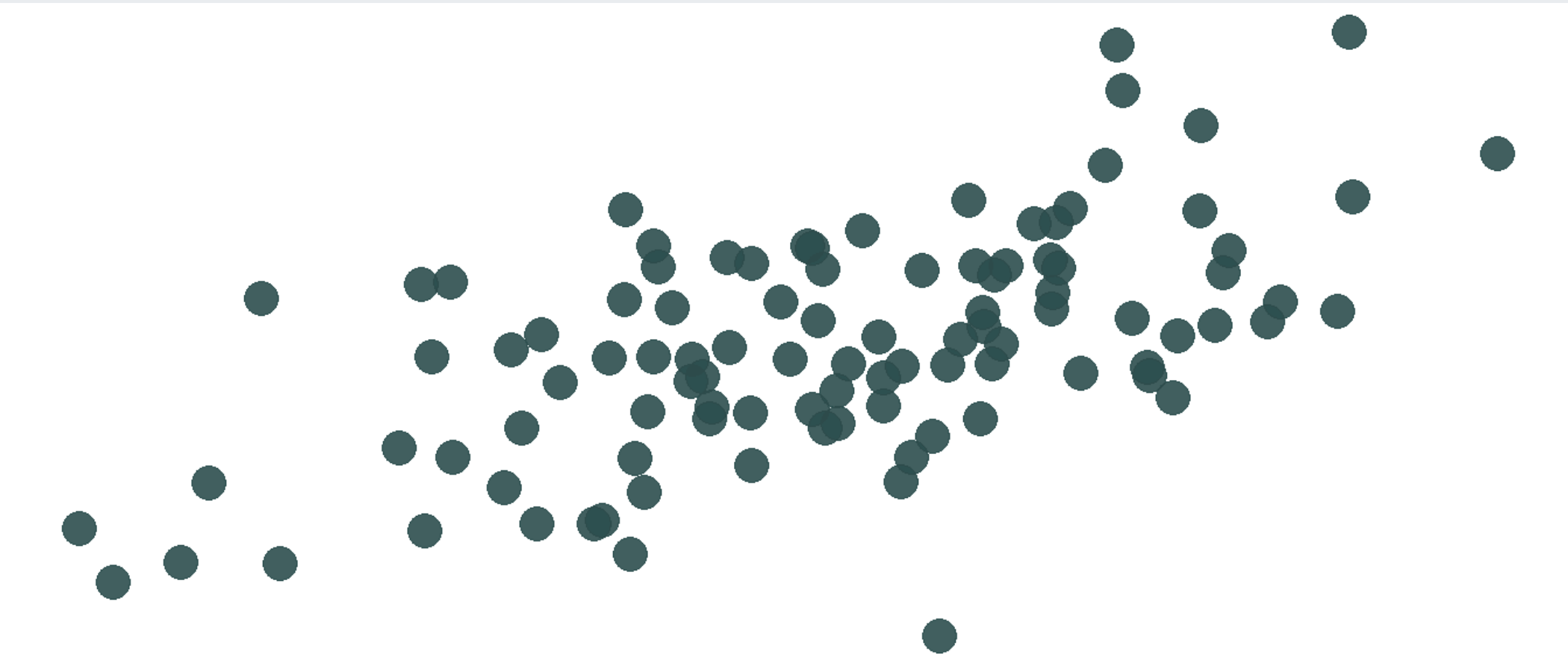
o de manera equivalente en forma de matriz,

$$\hat{\beta} = (X^T X)^{-1} X^T y \,.$$

```
modelo_felicidad <-
  lm(felicidad ~ galletas,
     data = galletas_datos)
```

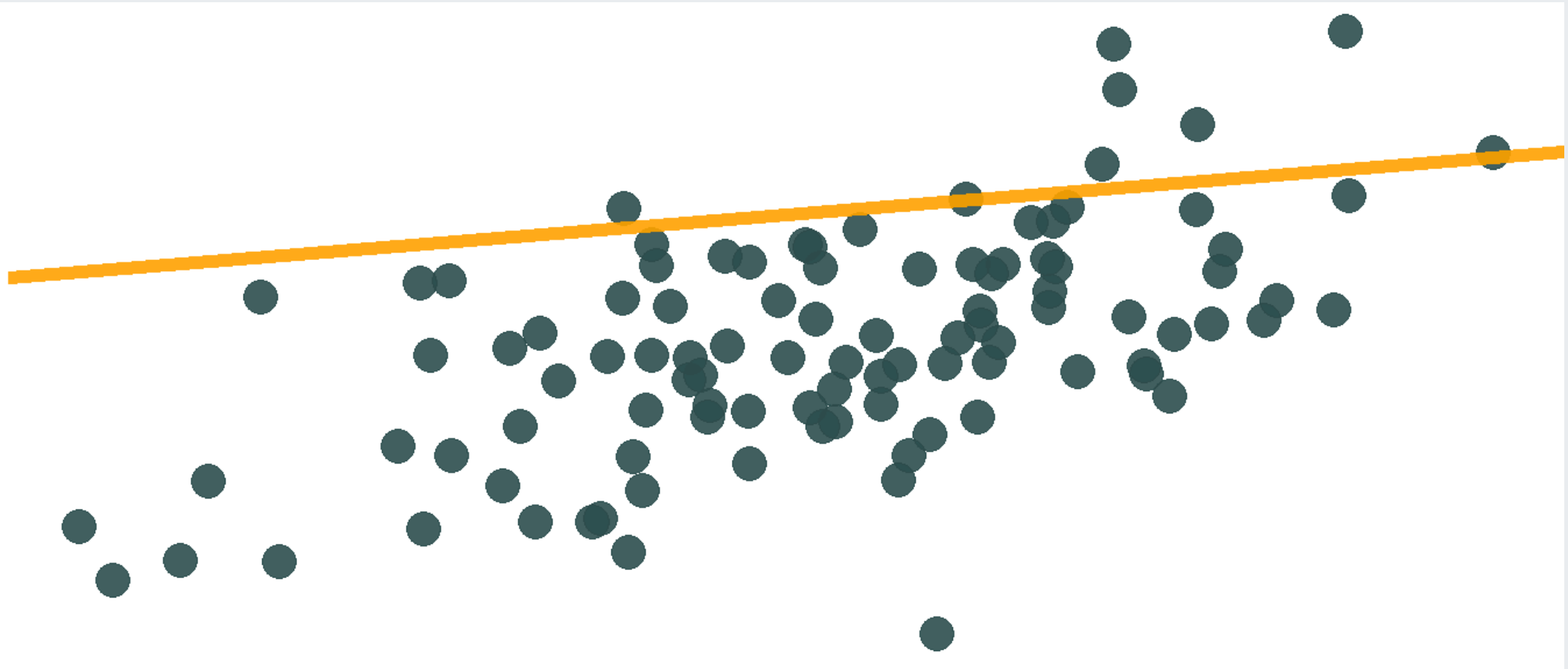
MCO vs Otras líneas

Volvamos a la relación felicidad y galletas…



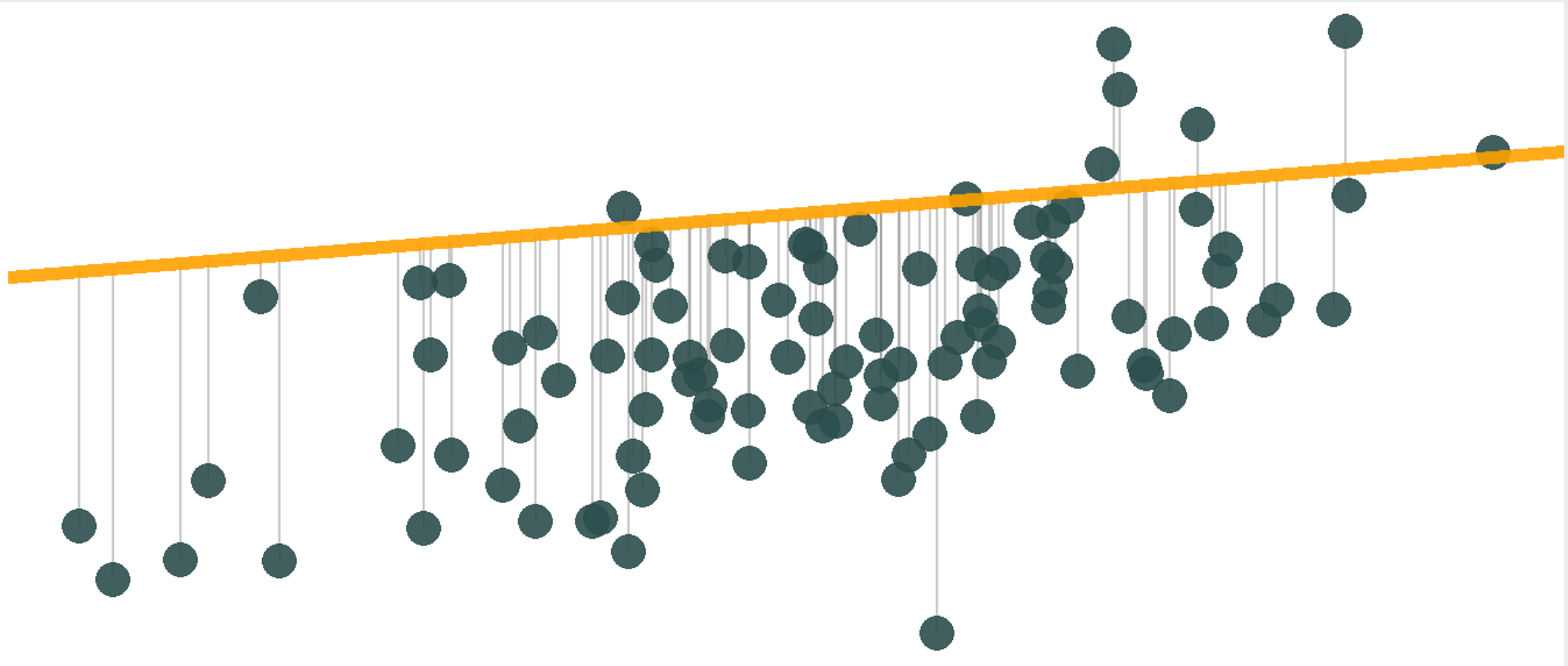
MCO vs Otras líneas

Para cualquier línea $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



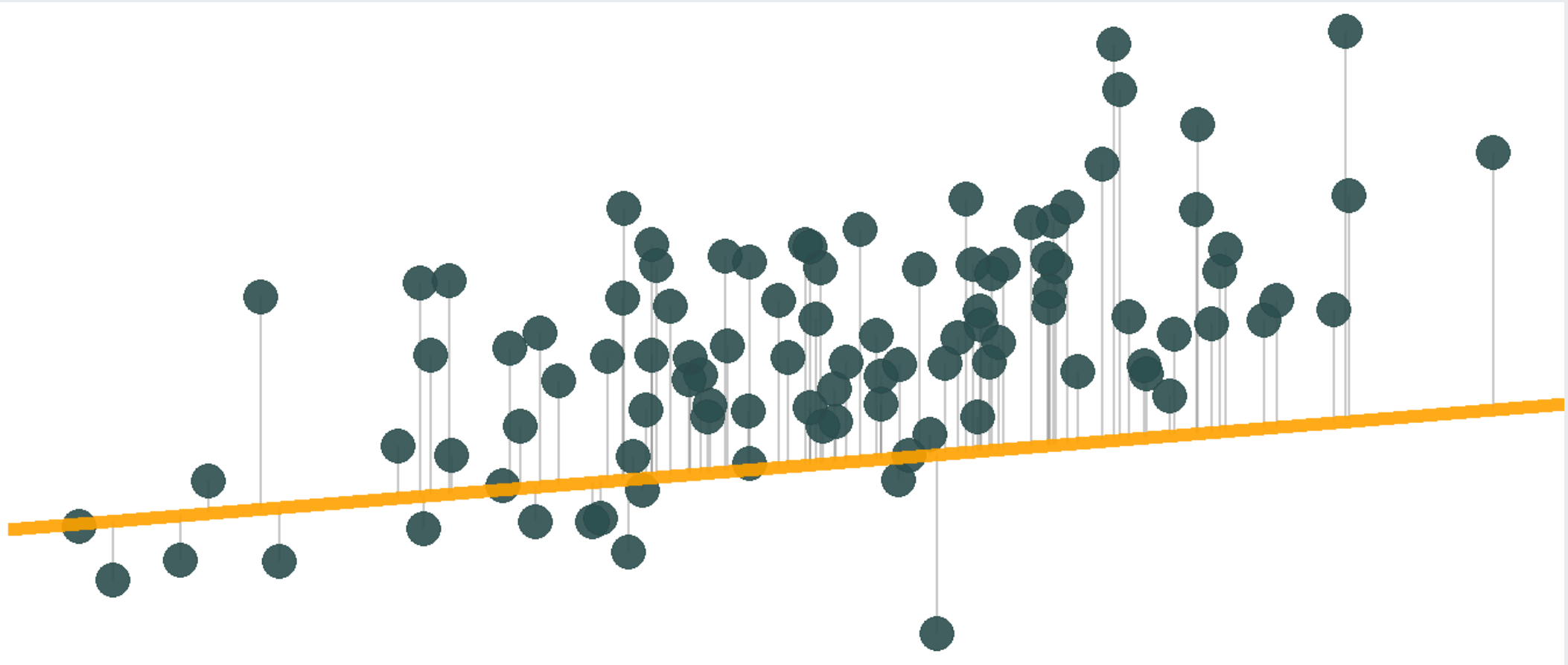
MCO vs Otras líneas

Para cualquier línea $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, podemos calcular $e_i = y_i - \hat{y}_i$



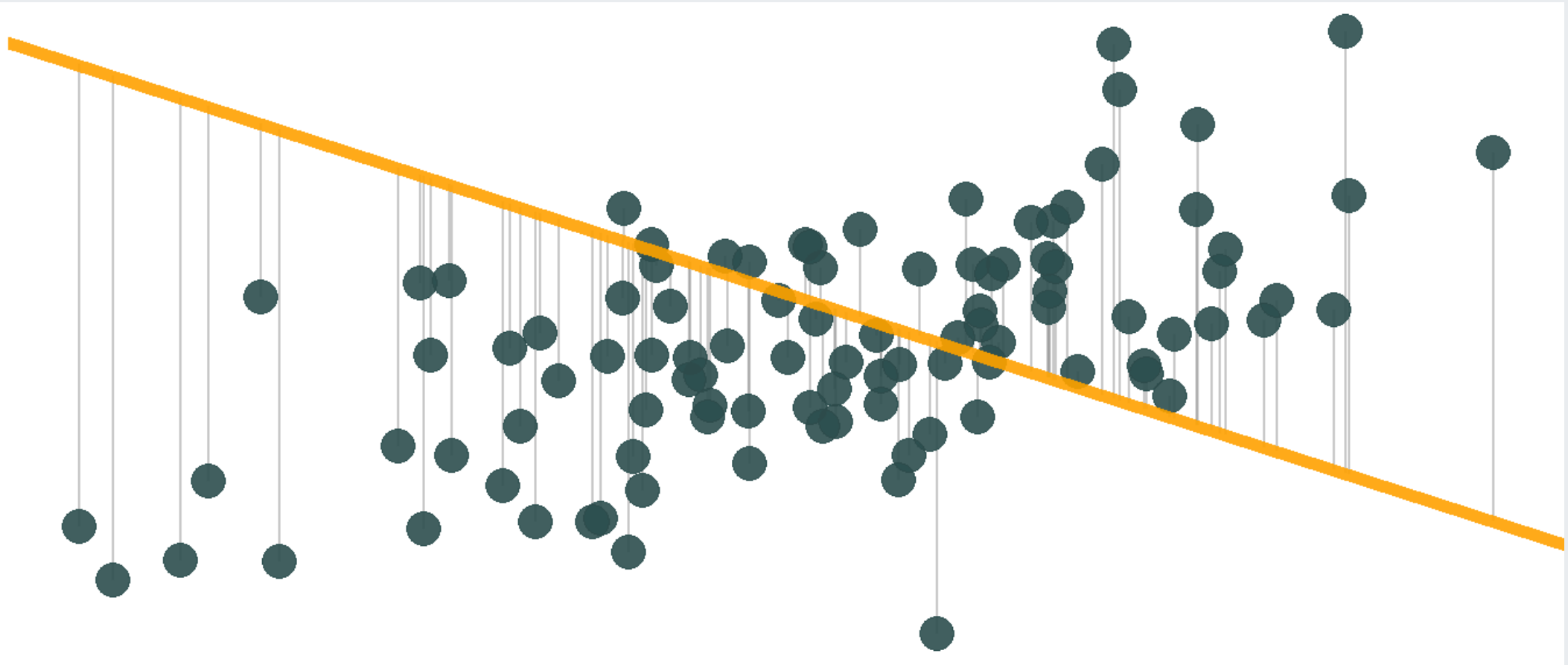
MCO vs Otras líneas

Para cualquier línea $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, podemos calcular $e_i = y_i - \hat{y}_i$



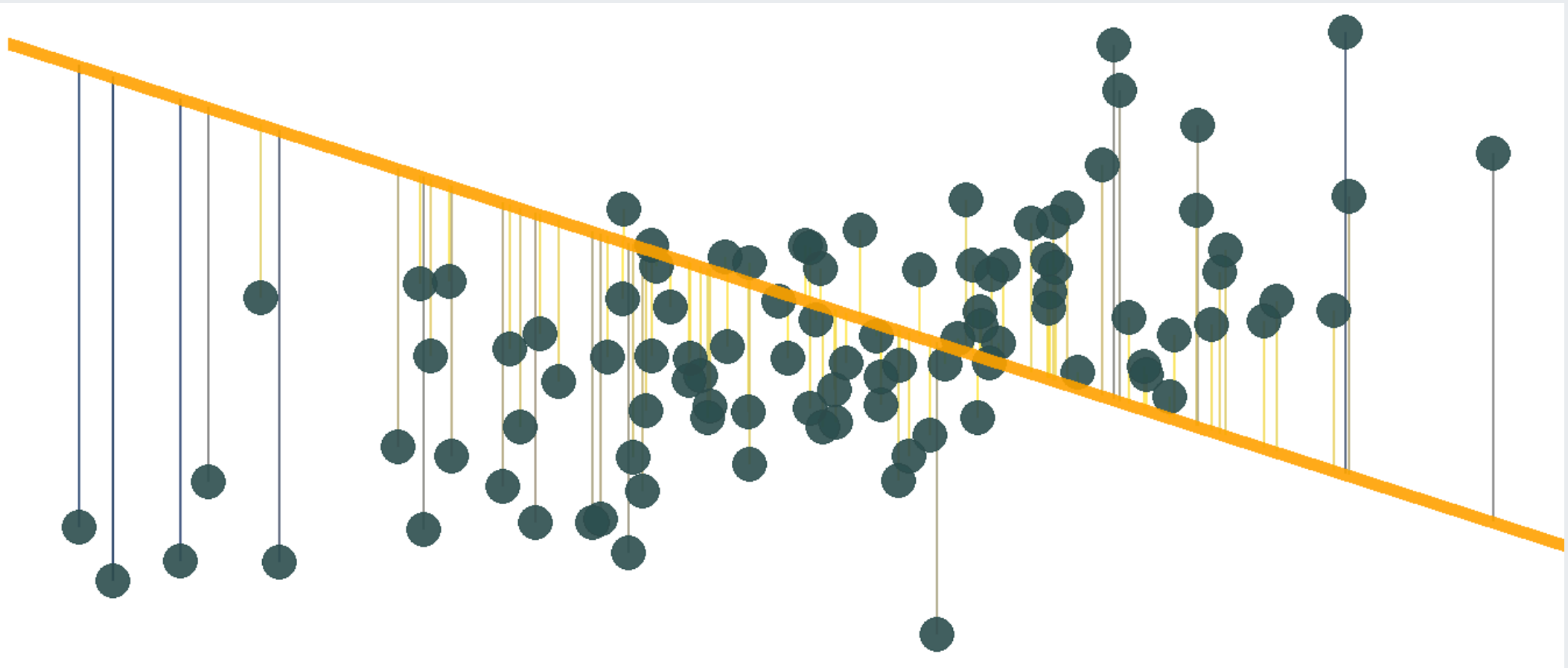
MCO vs Otras líneas

Para cualquier línea $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, podemos calcular $e_i = y_i - \hat{y}_i$



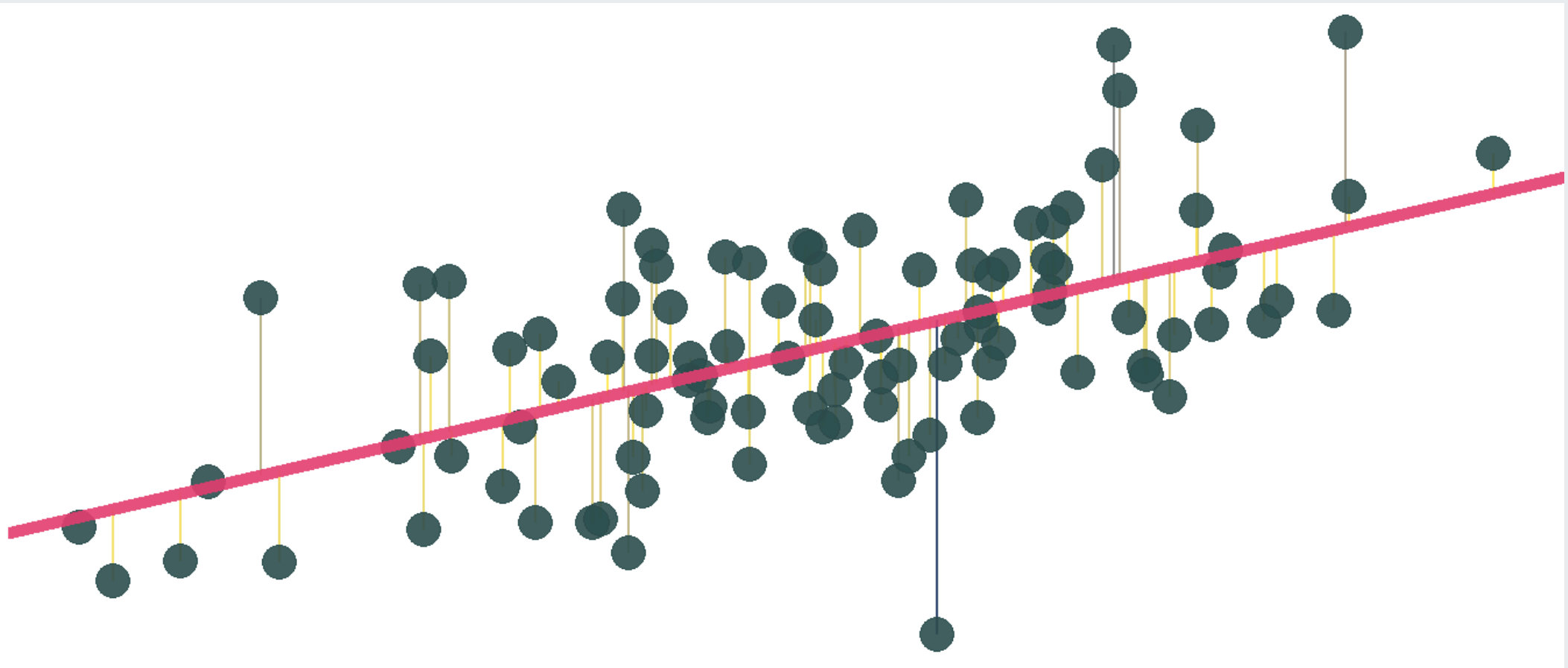
MCO vs Otras líneas

$\sum e_i^2$ calcula el cuadrado de los errores: errores más grandes reciben una penalización mayor



MCO vs Otras líneas

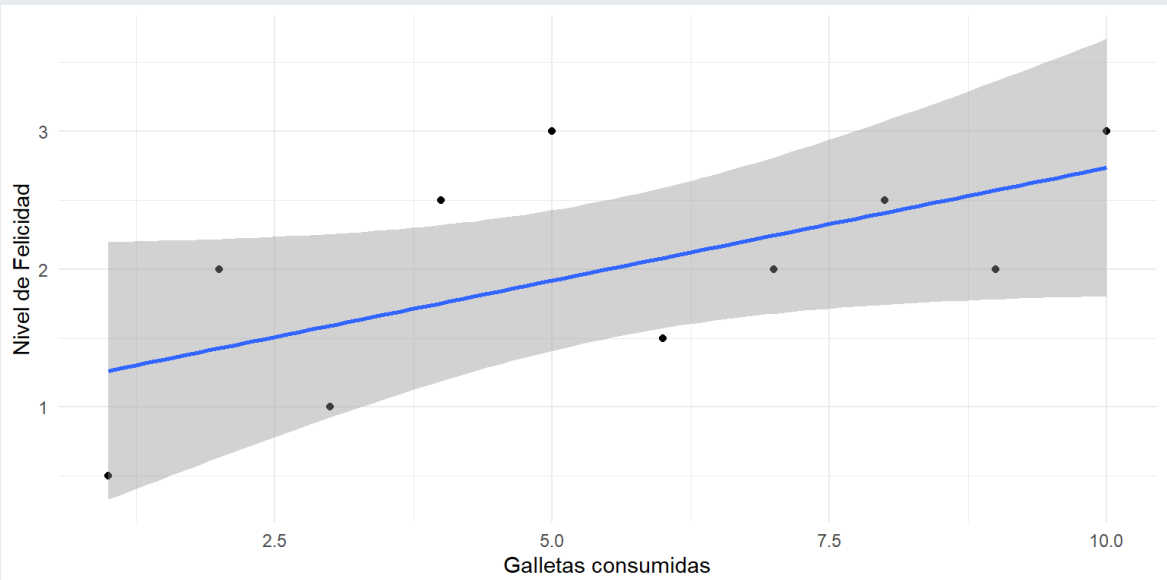
La línea de MCO es la combinación de $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan $\sum e_i^2$



¿Cómo gráficar la línea de regresión?

`geom_smooth(method="lm")` es la función dentro de ggplot para gráficar la línea de regresión y su respectivo intervalo de confianza.

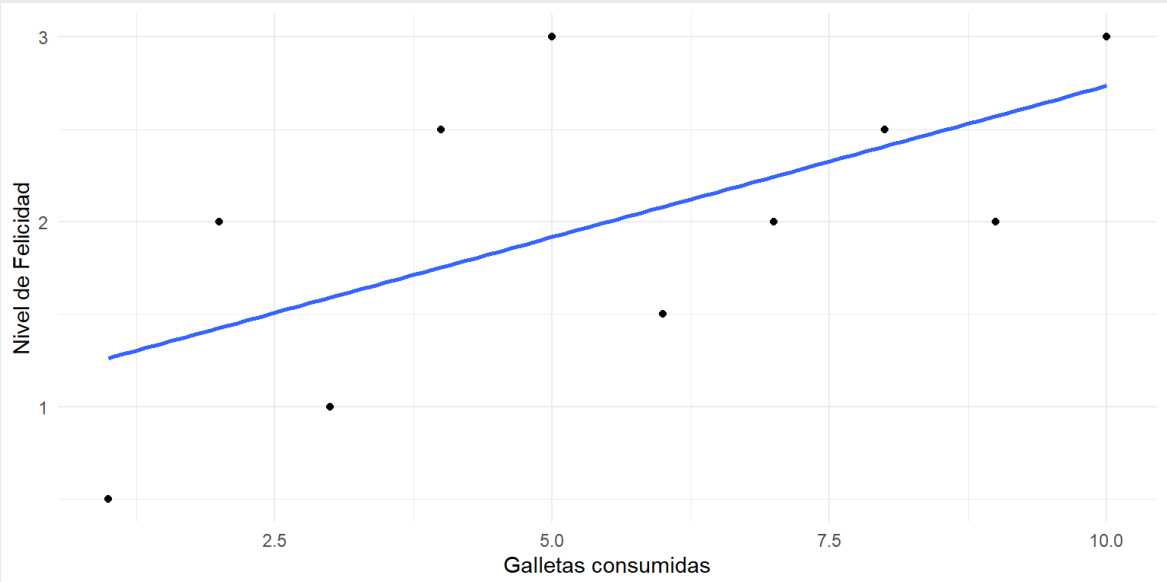
```
ggplot(galletas_datos,
       aes(x = galletas,
           y = felicidad))
  geom_point() +
  geom_smooth(method="lm")
  labs(x = "Galletas consu
       y = "Nivel de Felic
  theme_minimal()
```



¿Cómo gráficar la línea de regresión?

La opción `se = FALSE` elimina los intervalos de confianza.

```
ggplot(galletas_datos,
       aes(x = galletas,
           y = felicidad))
  geom_point() +
  geom_smooth(method="lm",
  labs(x = "Galletas consu
       y = "Nivel de Felic
  theme_minimal()
```



Estimar una regresión en R

- La función `lm()` estima una regresión lineal y la función `summary()` reporta los resultados:

```r
name_of_model <- lm(y ~ x, data = DATA)

summary(name_of_model)  # Para ver los detalles del modelo
```

- La función `tidy` del paquete `broom` reporta los resultados del modelo como un data frame para graficar:

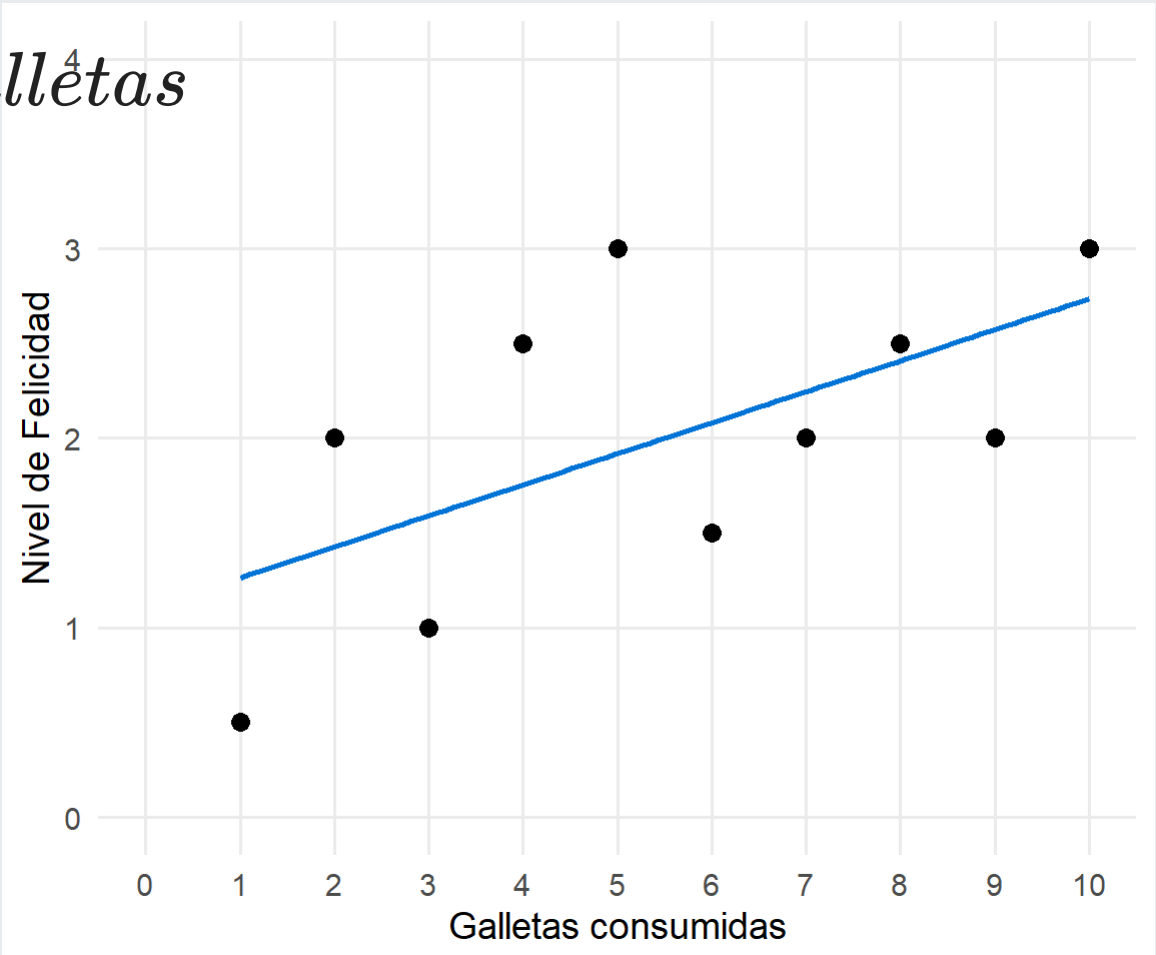
```r
library(broom)

tidy(name_of_model)
```

Modelando Galletas y Felicidad

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Gallétas$$

```r
modelo_felicidad <-
  lm(felicidad ~ galletas,
     data = galletas_datos
```



Modelando Galletas y Felicidad

La función `tidy` nos permite ver los:

- coeficientes
- errores estándar
- el estadístico t
- el p-value
- los IC

Para nuestro `modelo_felicidad`:

```r
tidy(modelo_felicidad, conf.int = TRUE)
```

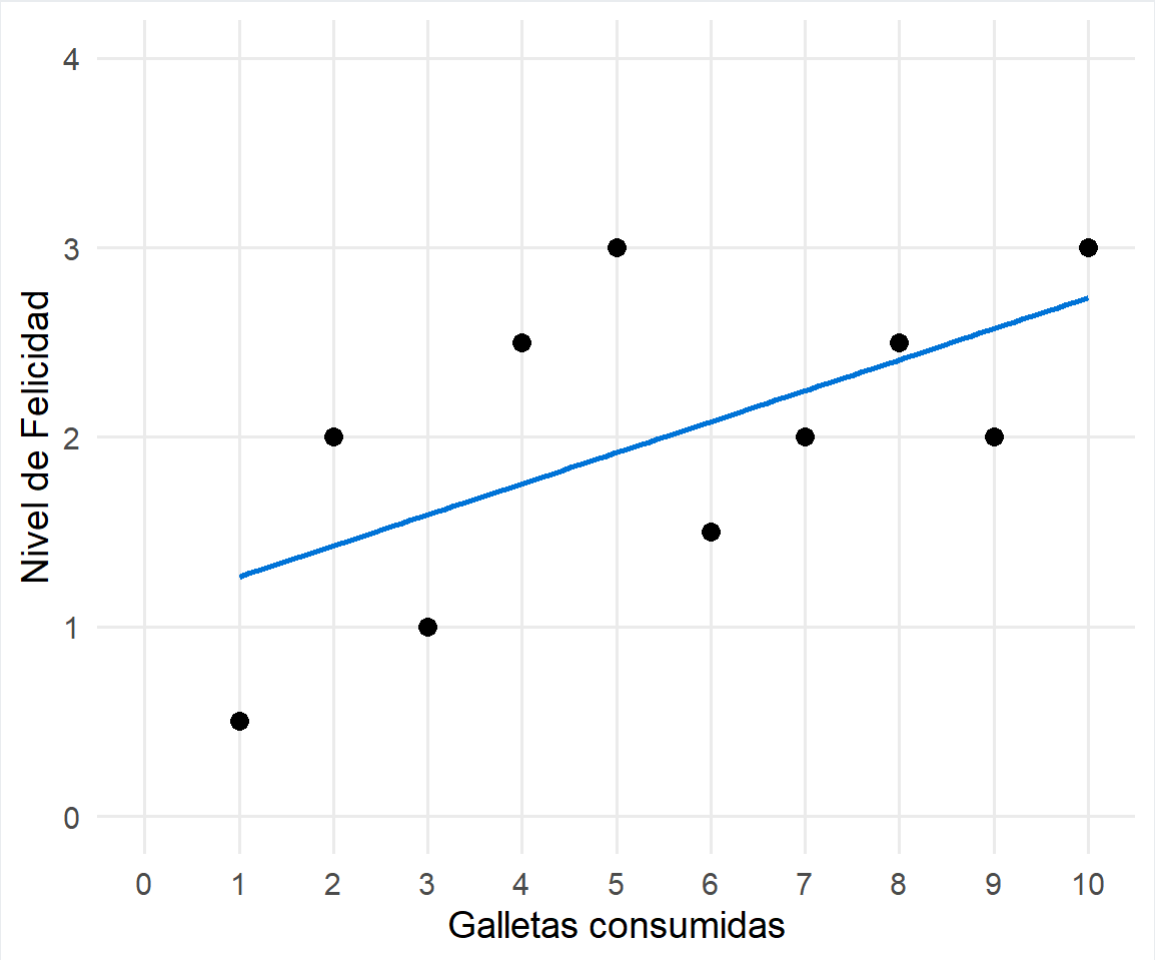
```
# A tibble: 2 × 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)    1.1       0.470      2.34  0.0475   0.0156     2.18
2 galletas       0.164     0.0758     2.16  0.0629  -0.0111     0.338
```

Traduciendo los resultados a matemáticas

```
# A tibble: 2 × 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)    1.1
2 galletas       0.164
```

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Galletas$$

$$\widehat{Felicidad} = 1.1 + 0.16 \times Galletas$$



Interpretación de los coeficientes

Un incremento en una unidad de x está *asociado* con un incremento (o reducción) promedio de β_1 unidades en y

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Galletas$$

$$\widehat{Felicidad} = 1.1 + 0.16 \times Galletas$$

- En *promedio*, una galleta adicional está asociado a un aumento en la felicidad de 0.16 unidades
- Si no hay consumo de galletas, esperamos que el puntaje de felicidad sea 1.1 unidades

¿Es el intercepto importante?

- La interpretación del intercepto es importante si en el contexto de los datos:
  1. La variable independiente puede tomar valores iguales o cercanos a cero
  2. La variable independiente tiene valores cercanos a cero en los datos observados
- En caso contrario, el intercepto no tiene ninguna interpretación práctica
- Veremos más ejemplos sobre esto más adelante…

Ejercicio 2

Según la sabiduría popular en Hollywood, el recaudo durante el primer fin de semana es un fuerte predictor del éxito comercial de una película.

1. Grafiquen la línea de regresión entre el recaudo total en Estados Unidos y el recaudo en el primer fin de semana para evaluar esta creencia.
2. Estimen la siguiente regresión y llámenla `hollywood_model`:

$$\widehat{\text{US Total Gross}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Opening Gross}$$

3. Usando la función `tidy` reporten los valores de los coeficientes.
4. ¿Cuál es la interpretación de $\hat{\beta}_1$ en este caso?¿Y de $\hat{\beta}_0$?

Predicción

$$\widehat{\text{US Total Gross}} = 5,108,220 + 3.12 \times \text{Opening Gross}$$

Según nuestro modelo, ¿cuál sería el recaudo en US de una película cuyo recaudo en el primer fin de semana fue de $50,000,000?

$$\begin{aligned}\widehat{\text{US Gross}} &= 5,108,220 + 3.12 \times \text{Opening Gross} \\ &= 5,108,220 + 3.12 \times 50,000,000 \\ &= 161,108,220\end{aligned}$$

Predicción con R

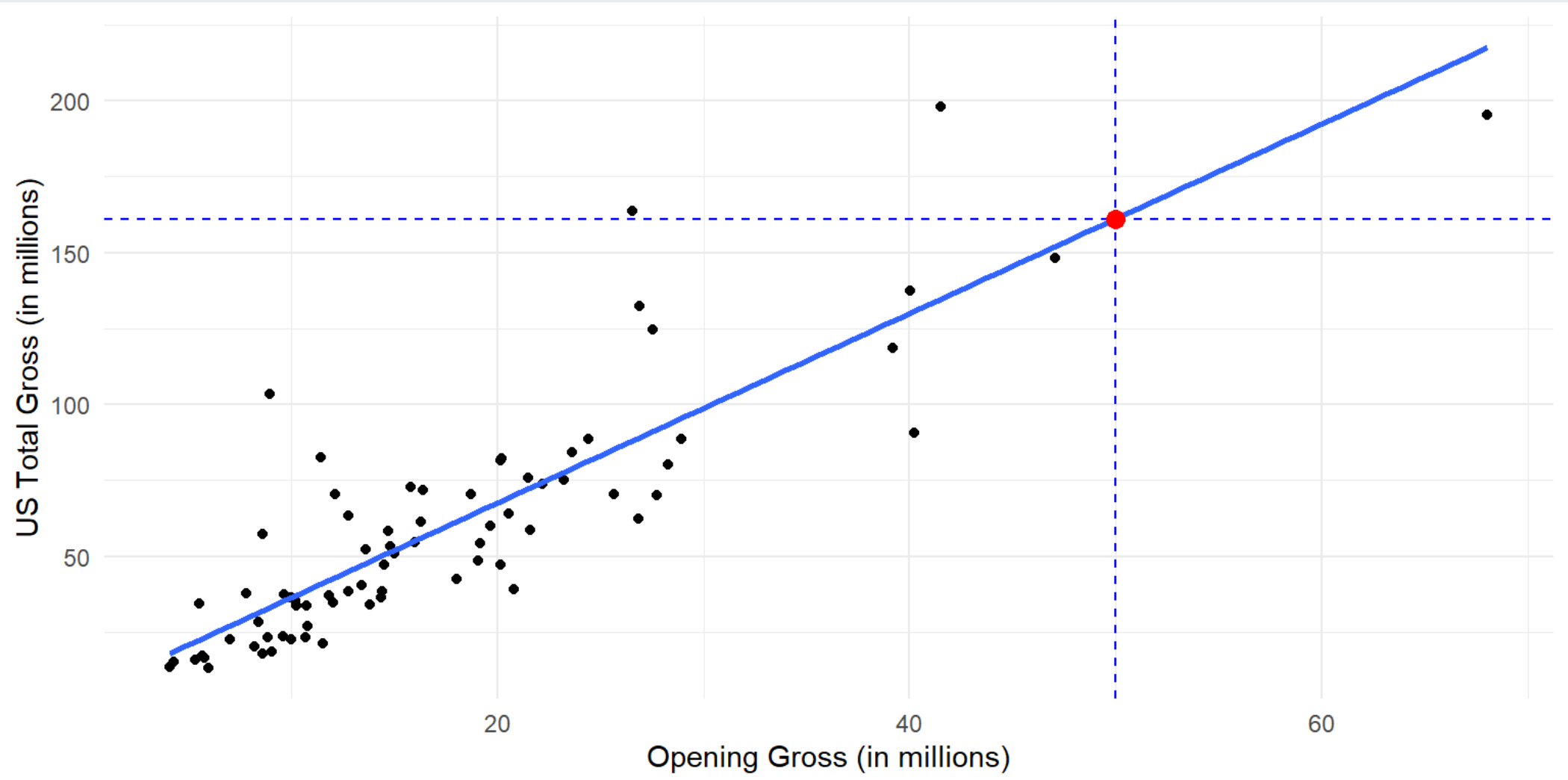
El comando `predict()` nos permite predecir $\widehat{\text{US Gross}}$ para uno o varios valores:

```r
# Creamos los valores para los cuales queremos predecir
valores_opening <- data.frame(opening_gross = c(20000000,40000000,50000000)


# Predice los valores con los coeficientes estimados
# en hollywood_model
predict(hollywood_model, newdata = valores_opening)
```

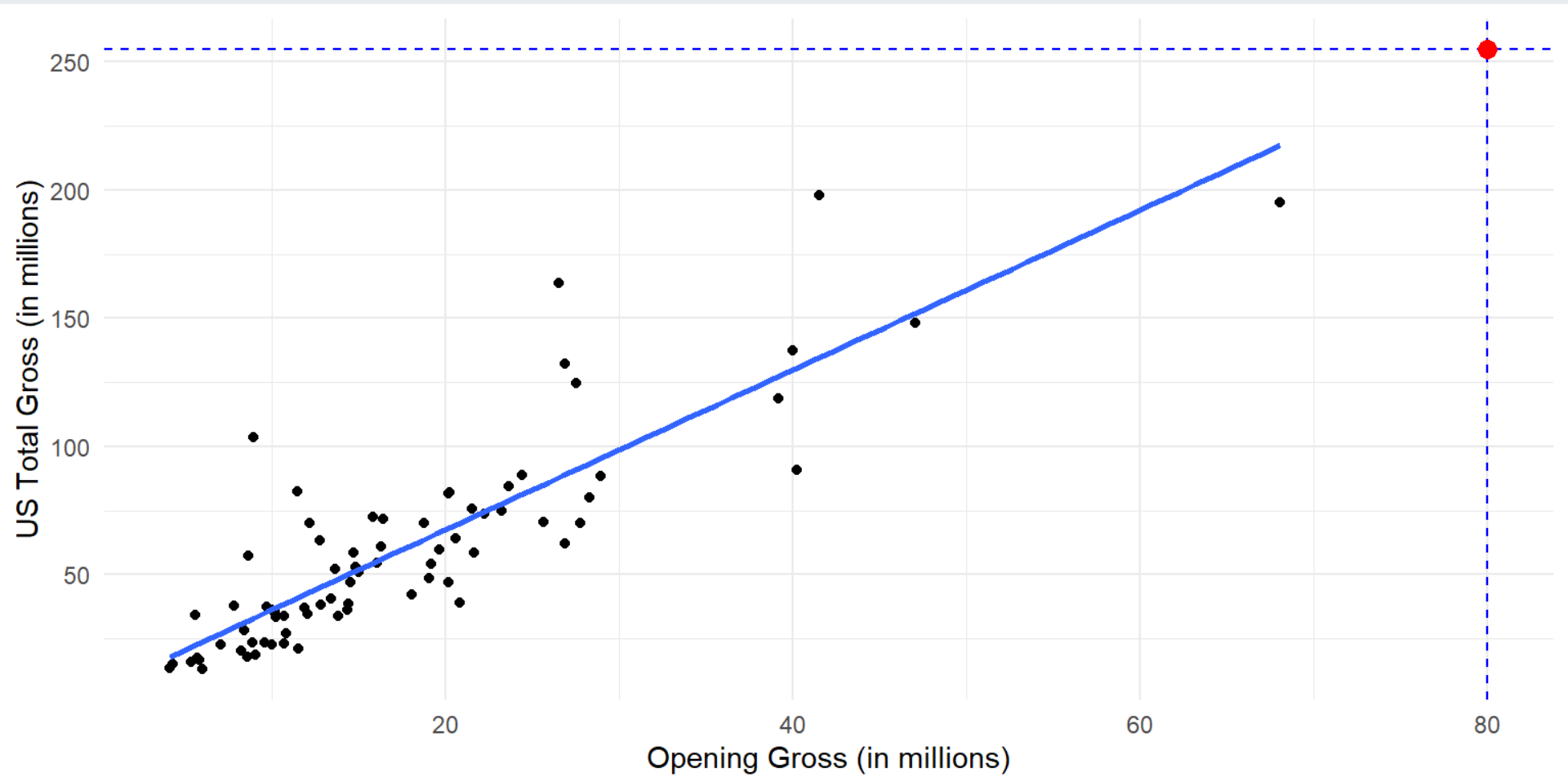
```
       1          2          3 
67520606 129932991 161139184 
```

Predicción


200
150
100
50
US Total Gross (in millions)
20
40
60
Opening Gross (in millions)

¿Es posible la extrapolación?

Extrapolar es tratar de predecir y fuera del rango de valores de x. Es posible pero no aconsejable.



Inferencia de los coeficientes

Cuando trabajamos con distribuciones muestrales, la idea era que:

$$\bar{x} \xrightarrow{\text{🤞 ojalá 🤞}} \mu$$

De igual manera, en el modelo de regresión queremos:

$$\hat{\beta} \xrightarrow{\text{🤞 ojalá 🤞}} \beta$$

Inferencia de los coeficientes

$$\widehat{\text{US Total Gross}} = 5,108,220 + 3.12 \times \text{Opening Gross}$$

- Es β_1 diferente de cero?

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

```
# A tibble: 2 × 7
  term            estimate   std.error statistic  p.value    conf.low
conf.high
  <chr>              <dbl>       <dbl>     <dbl>    <dbl>       <dbl>
<dbl>
1 (Intercept)   5108220.   4502660.         1.13 2.60e- 1 -3865567.
1.41e7
2 opening_gross       3.12       0.218      14.3  7.07e-23        2.69
3.56e0
```

Más pruebas de hipótesis

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t = \frac{3.12 - 0}{0.218} = 14.3 > t_{\alpha=0.05} = 1.96$$

- Rechazamos la H_0 a un nivel de significancia del 5%!
- El p-value es menor a 0.05:

```r
tidy(hollywood_model, conf.int = TRUE)[2,5]<0.05
```

```
     p.value
[1,]    TRUE
```

- El intervalo de confianza al 95% no incluye el cero

Regresión Lineal Múltiple

Regresión Múltiple

No estamos limitados a una sola variable explicativa!

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n$$

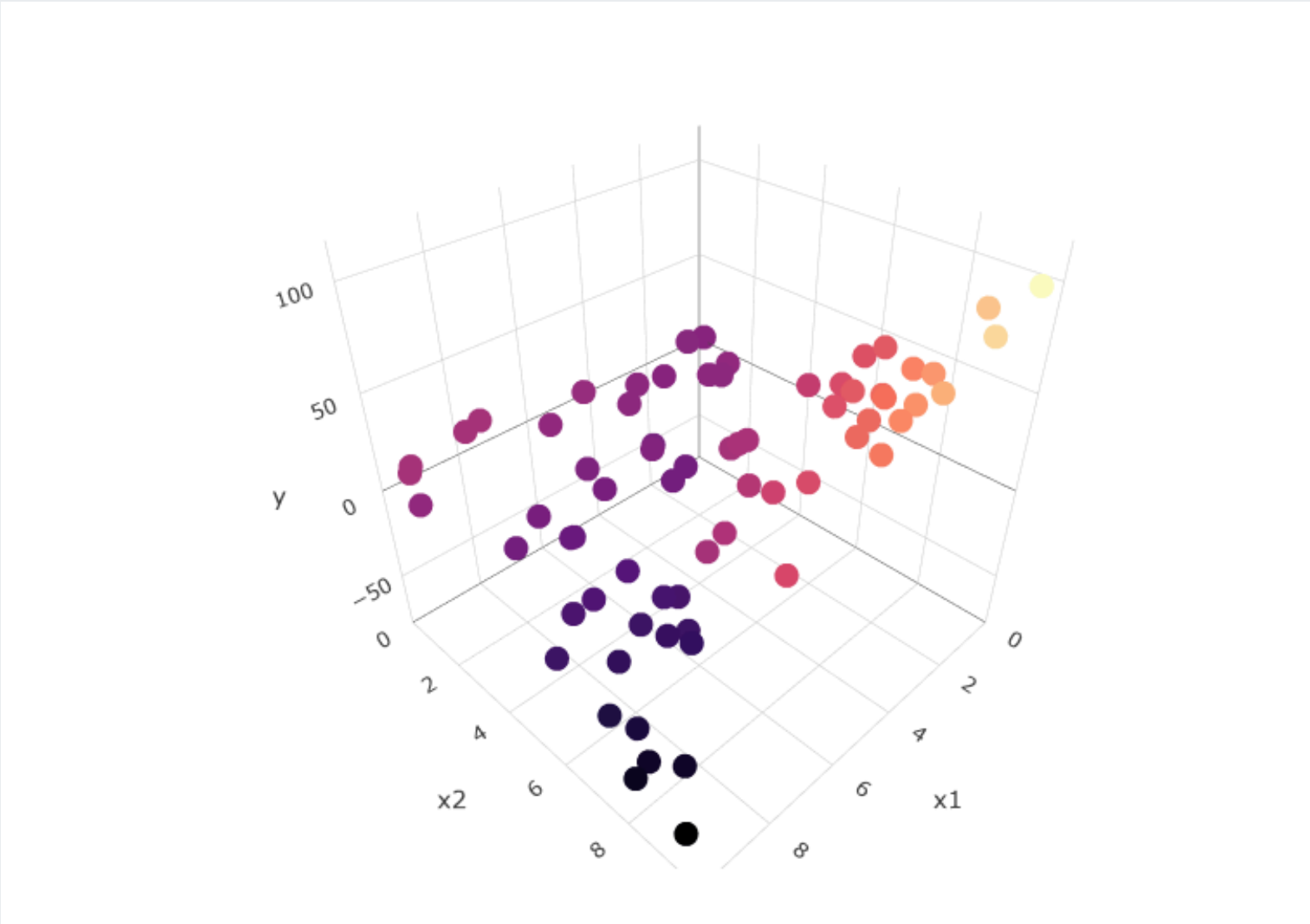
Podríamos pensar que el verdadero modelo de felicidad es:

$$\widehat{\text{Felicidad}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Galletas} + \hat{\beta}_2 \times \text{Estudiante}$$

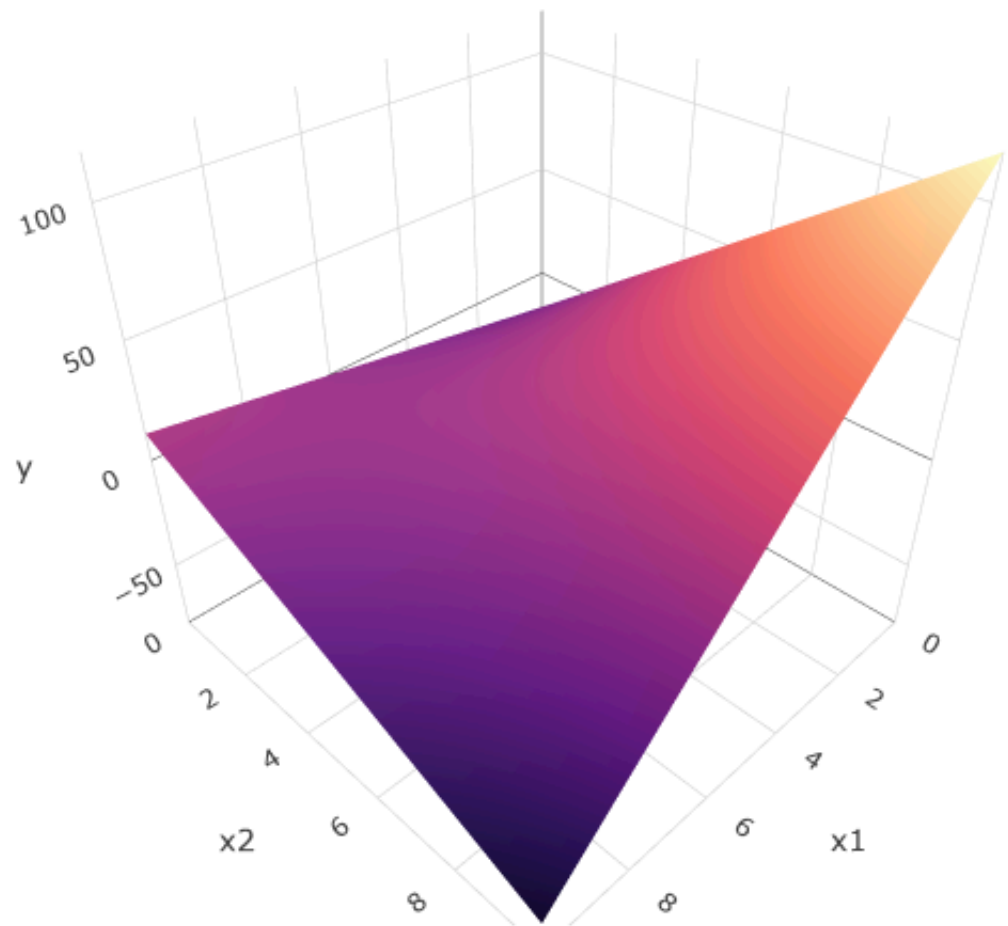
O que para predecir el recaudo de una película se necesita el siguiente modelo:

$$\widehat{\text{US Gross}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Opening Gross} + \hat{\beta}_2 \text{Budget} + \hat{\beta}_3 \text{Sequel}$$

¿Cómo visualizar la relación $y = f(x_1, x_2) + \varepsilon$?



¿Cómo visualizar la “recta” en una regresión múltiple?



Variables Categóricas vs Variables Continuas

En los dos modelos antes explicados hay dos tipos de variables:

Variables Categóricas

- Estudiante
- Sequel

Variables Continuas

- Galletas
- Opening Gross

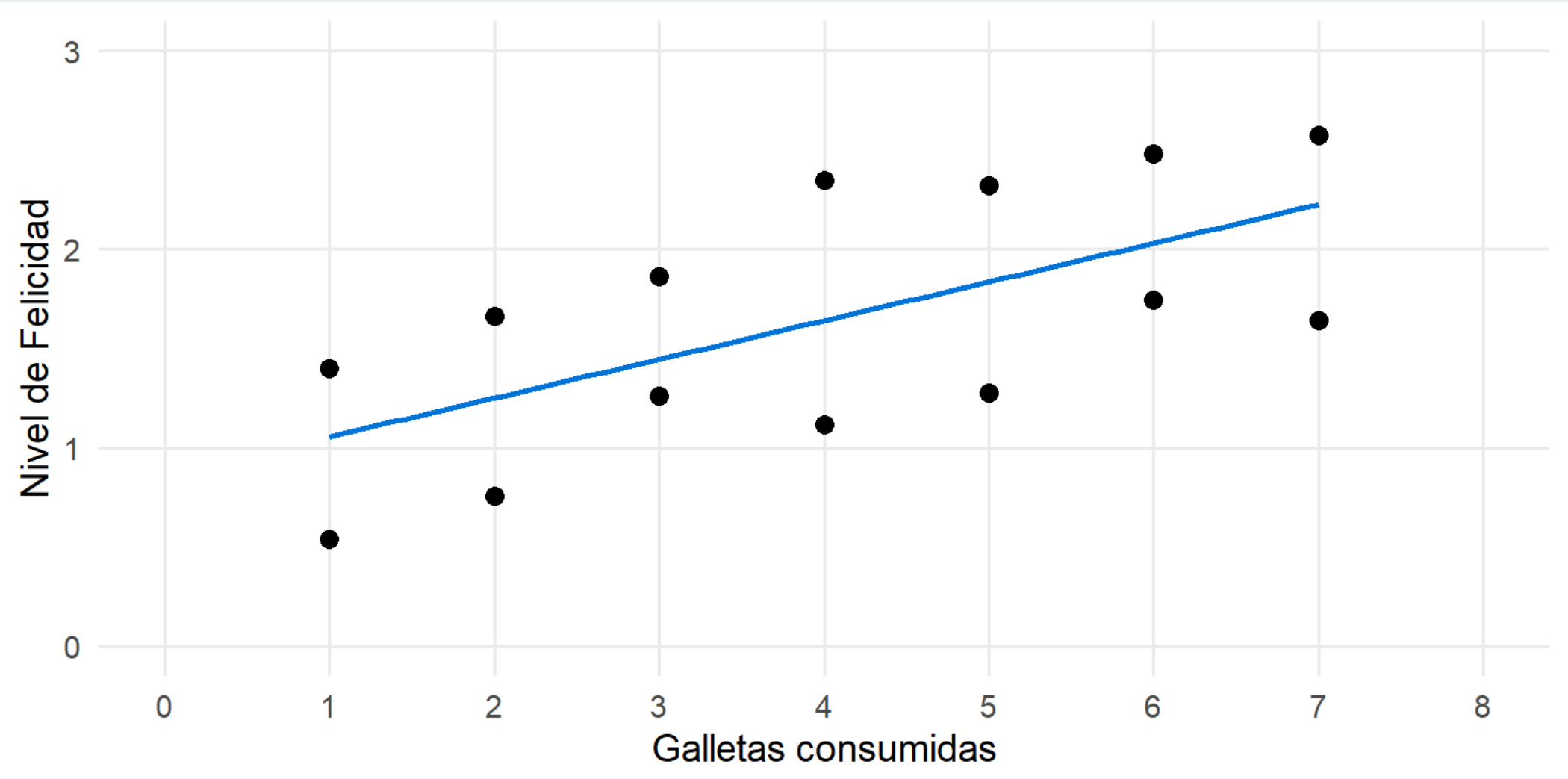
Variables Categóricas vs Variables Continuas

Variables Categóricas vs Variables Continuas

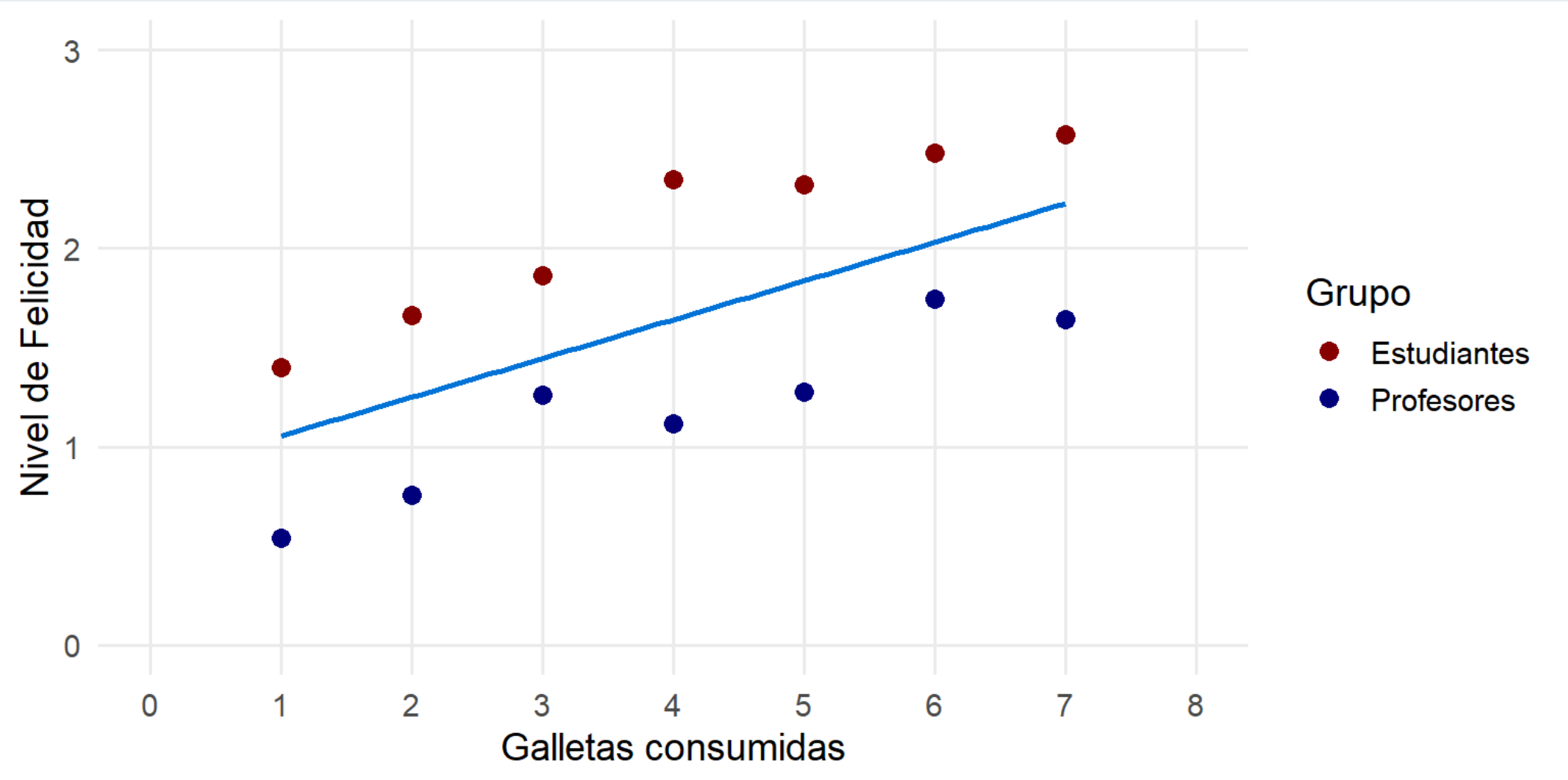
Galletas, Felicidad, Estudiantes y Profesores

galletas	felicidad	estudiante
1	0.5379049	0
2	0.7539645	0
3	1.2617417	0
4	1.1141017	0
5	1.2758575	0
6	1.7430130	0
7	1.6421832	0
1	1.3969878	1
2	1.6626294	1
3	1.8608676	1
4	2.3448164	1
5	2.3219628	1
6	2.4801543	1
7	2.5721365	1

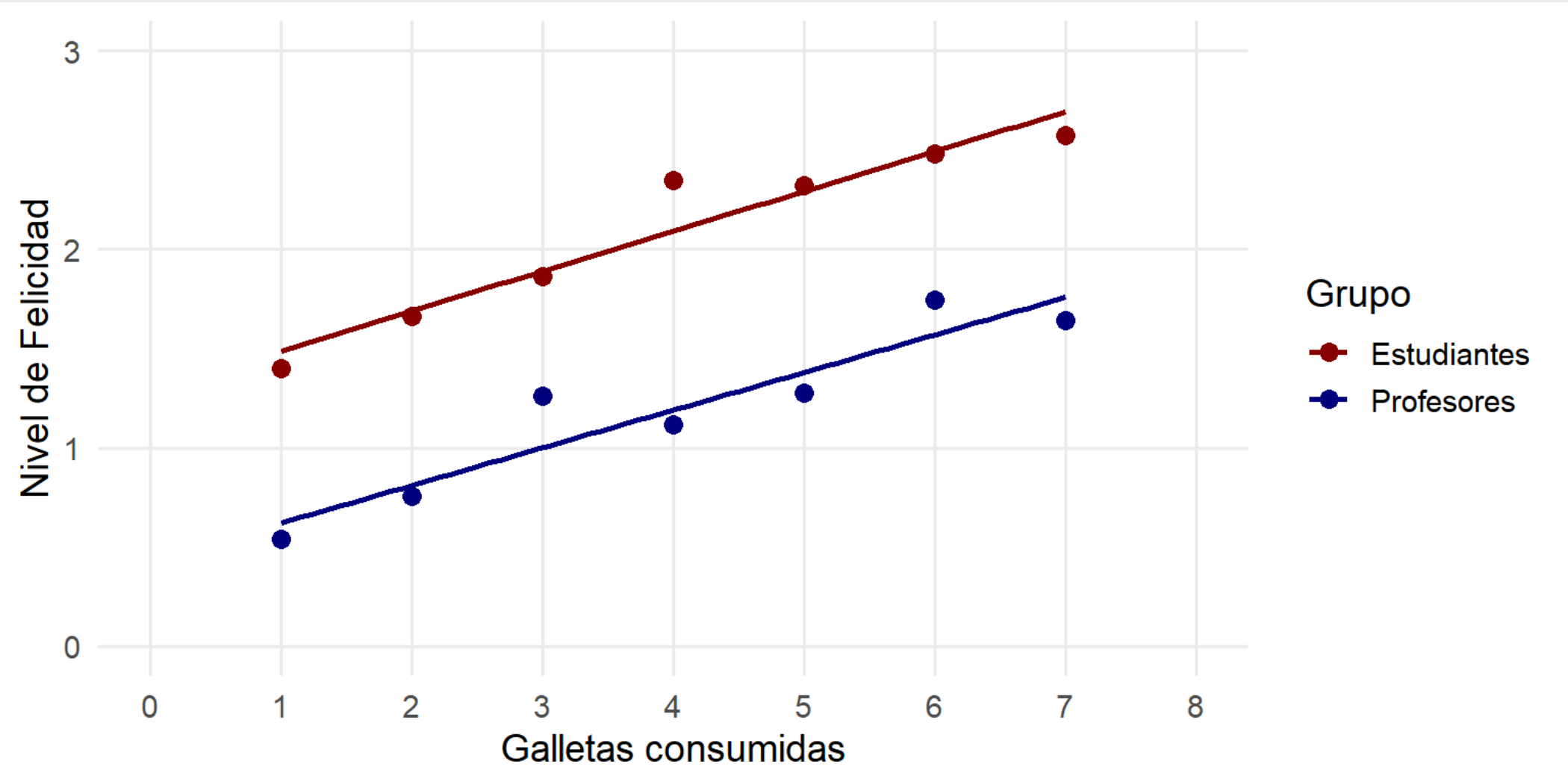
Galletas y Felicidad



¿Qué pasa si diferenciamos entre profesores y estudiantes?



Ambos grupos parecen tener pendientes diferentes...



Variables Categóricas

¿Cómo interpretar el modelo si agregamos una variable categórica $Estudiante$ que sea igual a 1 si la observación es de un estudiante?

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Galletas + \hat{\beta}_2 \times Estudiante$$

El intercepto para las observaciones de los profesores será $\hat{\beta}_0$ porque $Estudiante = 0$

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Galletas$$

Variables Categóricas

¿Cómo interpretar el modelo si agregamos una variable categórica $Estudiante$ que sea igual a 1 si la observación es de un estudiante?

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Galletas + \hat{\beta}_2 \times Estudiante$$

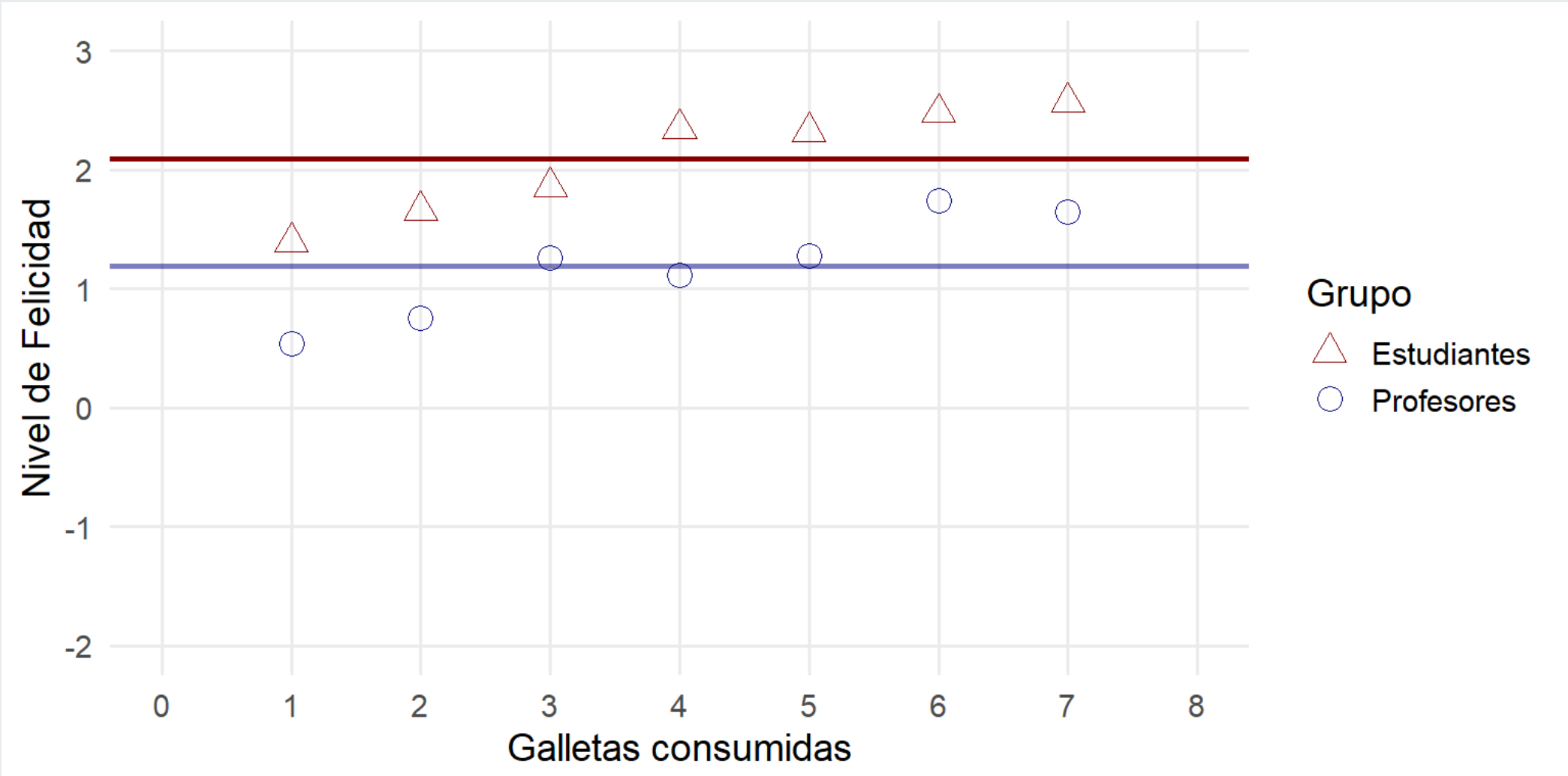
El intercepto para las observaciones de los estudiantes será $\hat{\beta}_0 + \hat{\beta}_2$ porque $Estudiante = 1$

$$\widehat{Felicidad} = \hat{\beta}_0 + \hat{\beta}_1 \times Galletas + \hat{\beta}_2$$

En promedio, los estudiantes tienen un nivel de felicidad $\hat{\beta}_2$ más alto que los profesores ($\hat{\beta}_2 > 0$ según la gráfica)

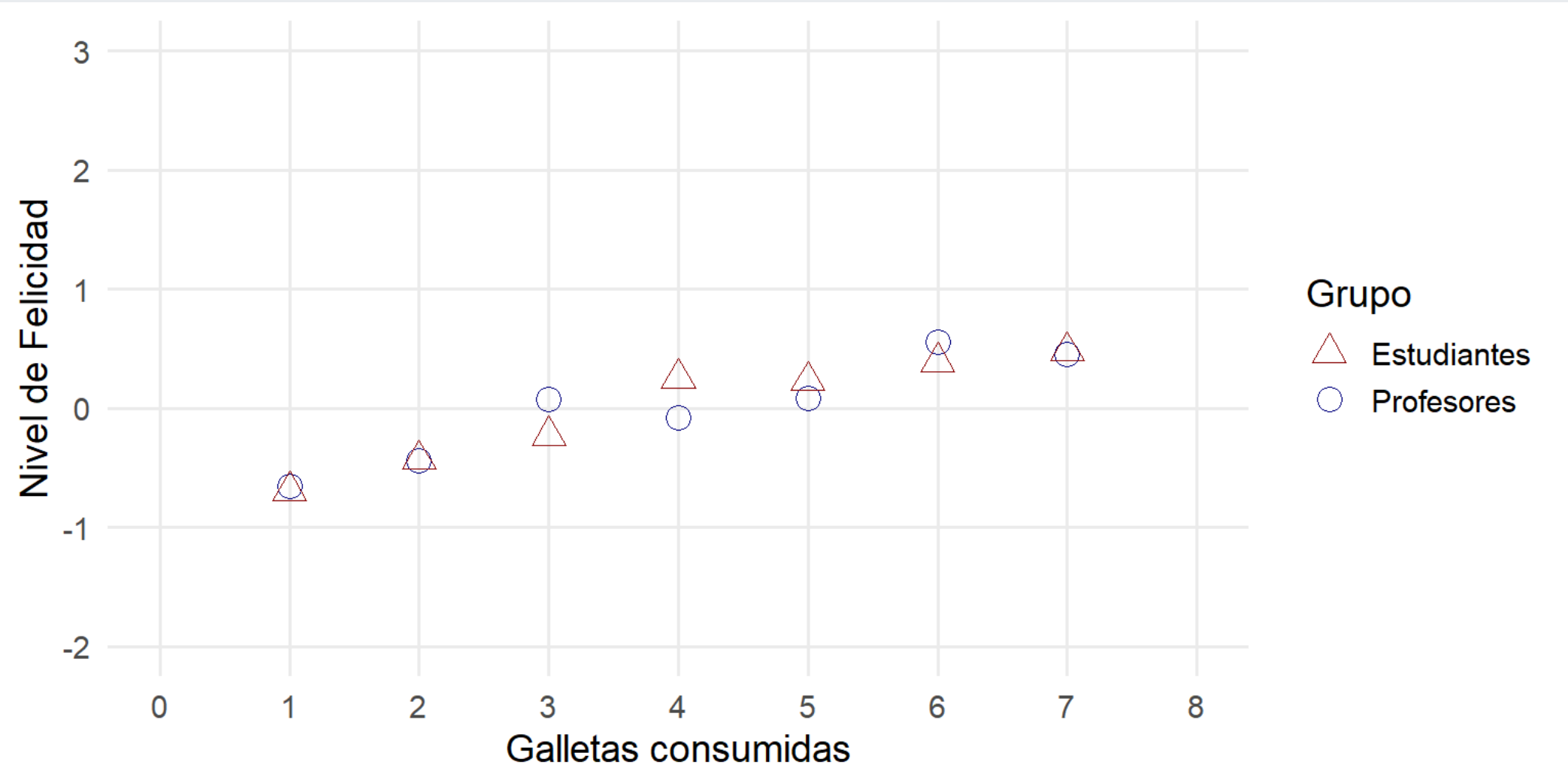
Otra manera de visualizar la variable categórica:

El intercepto y $Estudiante$ controlan por las medias de ambos grupos:



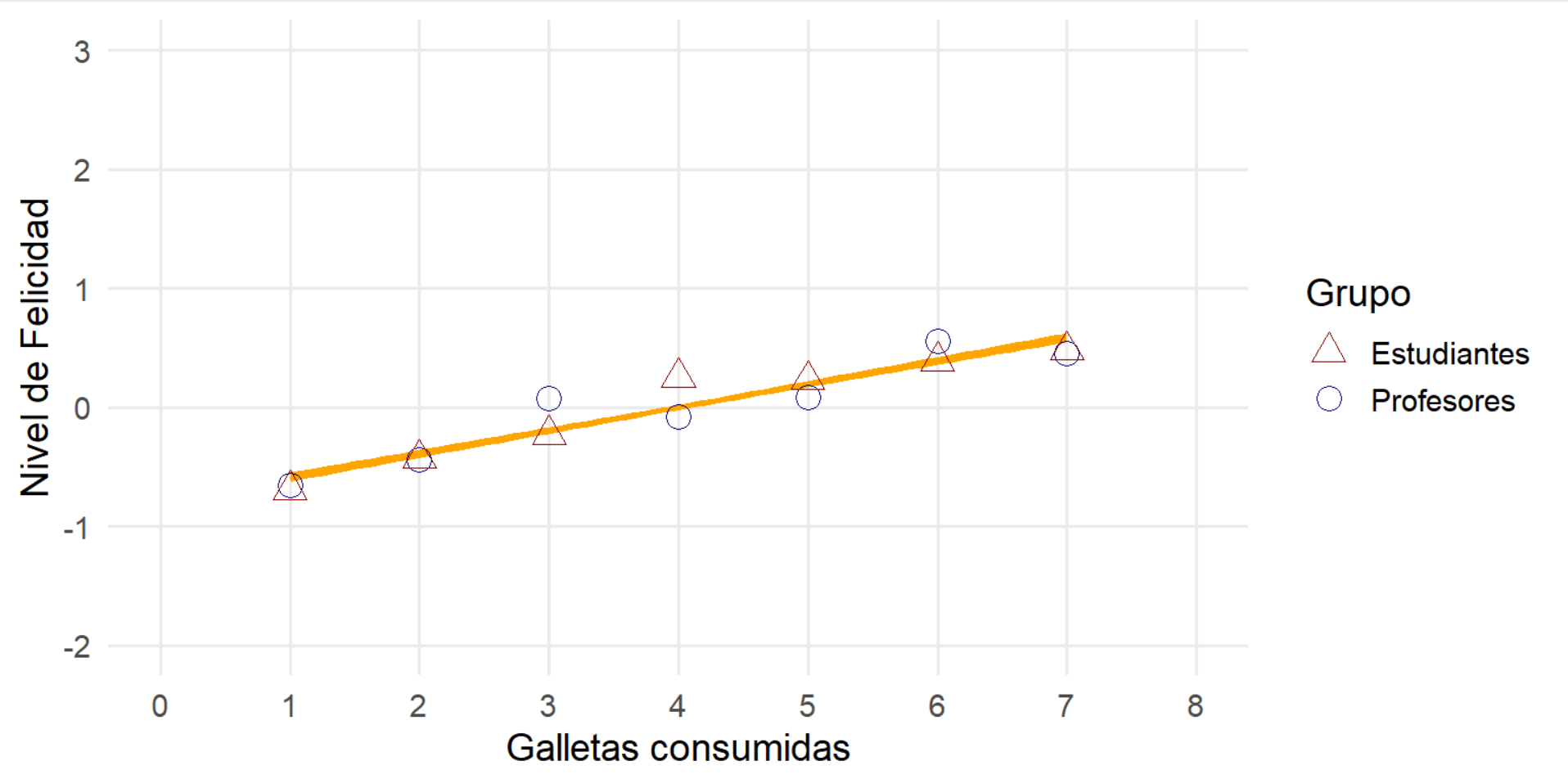
Otra manera de visualizar la variable categórica:

Luego de remover las medias de ambos grupos:



Otra manera de visualizar la variable categórica:

$\hat{\beta}_1$ estima la relación de interés luego de controlar por $Estudiante$:



Estimar una regresión múltiple en R

- Nuevamente usamos la función `lm()`, simplemente separamos las variables independientes con un `+`:

```
name_of_model <- lm(y ~ x1 + x2 + x3, data = DATA)

tidy(name_of_model)
```

Ejercicio 3

1. Estimen la siguiente regresión con el nombre `hollywood_model_2`:

$$\widehat{\text{US Gross}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Opening Gross} + \hat{\beta}_2 \text{Budget} + \hat{\beta}_3 \text{Sequel}$$

2. Usando la función `tidy` respondan: ¿cuáles coeficientes son estadísticamente diferentes de cero?
3. Completen el siguiente código para predecir el recaudo total para una secuela con recaudo del fin de semana promedio y presupuesto promedio:

```
multiples_valores <- data.frame(opening_gross=_________,
                                budget=__________,
                                sequel=__)
predict(hollywood_model_2, newdata = multiples_valores)
```

Predicción en Regresión Múltiple

Así como en el caso con una variable, usamos el comando `predict()` para predecir $\widehat{\text{US Gross}}$. En este caso, necesitamos al menos un valor para cada variable que está en la regresión.

Hagamos una predicción para una de las observaciones en nuestros datos. En este caso, para la película "The Holiday":

```r
# Seleccionemos las 3 variables dependientes para The Holiday
the_holiday <- hollywood |>
  filter(movie=="The Holiday") |>
  select(opening_gross, budget, sequel)

# Usamos el comando predict nuevamente
y_gorro <- predict(hollywood_model, newdata = the_holiday)
y_gorro
```

```
       1
59596928
```

Predicción en Regresión Múltiple

- ¿Es precisa nuestra estimación? Comparemos nuestro estimado con el recaudo real de la película:

```r
# El valor observado
y <- hollywood |>
  filter(movie == "The Holiday") |>
  pull(us_gross)
y
```

```
[1] 63224849
```

- El error para "The Holiday" será:

$$e = y - \hat{y} = \text{US Gross} - \widehat{\text{US Gross}}$$

```r
y - y_gorro
```

```
      1
3627921
```

Filtrar la variación

- Cada x en el modelo explica una porción de la variación en y
- La interpretación acá es más complicada que en el modelo de regresión simple porque sólo se puede mover una variable a la vez

Interpretación para variables continuas

Manteniendo todo lo demás constante, un incremento de una unidad en x está asociado con un incremento/reducción promedio de β_k en y

$$\widehat{\text{US Gross}} = -8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budg}$$
$$- 11,929,834 \times \text{Sequel}$$

Manteniendo todo lo demás constante, un incremento de un dólar en el recaudo del primer fin de semana está asociado con un incremento promedio de 2.99 dólares en el recaudo total en US

Interpretación para variables categóricas

Manteniendo todo lo demás constante, y es, en promedio, β_k unidades mayor/menor para x_{k} comparado con x_{omitida}

$$\widehat{\text{US Gross}} = -8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budg}$$
$$- 11,929,834 \times \text{Sequel}$$

Manteniendo todo lo demás constante, las sequelas están asociadas a un recaudo promedio menor, en aproximadamente $11.9 millones, comparadas con las películas que no son secuelas

Variable categóricas con más de 2 niveles

$$\widehat{\text{US Gross}} = -8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budge}$$
$$- 11,929,834 \times \text{Sequel} - 15,000,000 \times \text{Trilogy}$$

Si es la primera película $Sequel = Trilogy = 0$, el modelo es:

$$\widehat{\text{US Gross}} = -8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budge}$$

Variable categóricas con más de 2 niveles

$$\widehat{\text{US Gross}} = -8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budg}$$
$$- 11,929,834 \times \text{Sequel} - 15,000,000 \times \text{Trilogy}$$

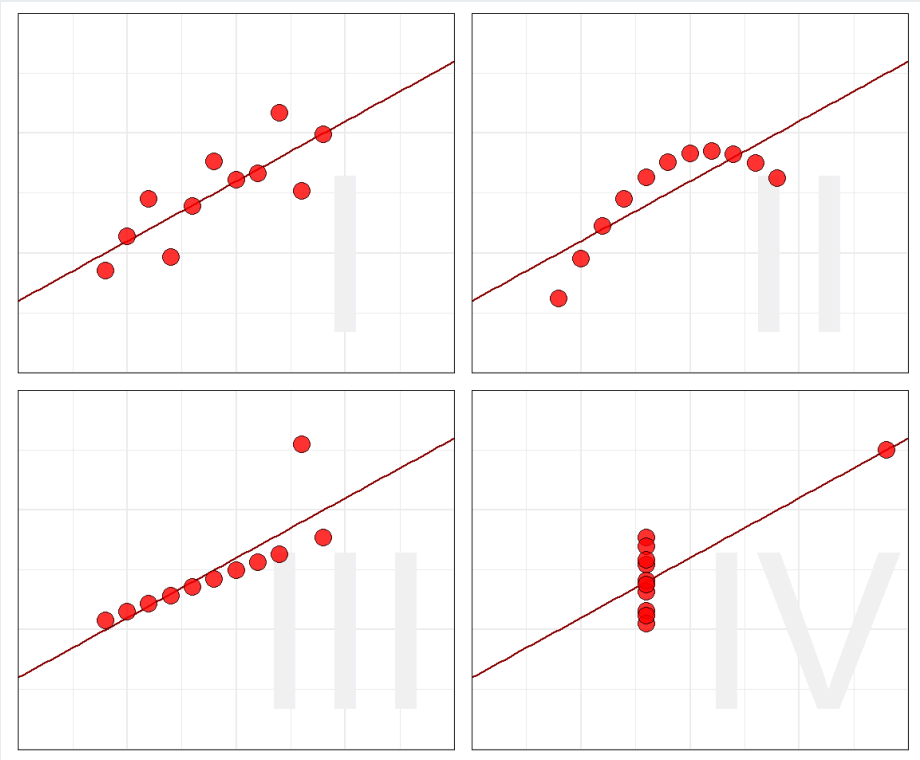
Si es trilogía, entonces $Sequel = 0$ y $Trilogy = 1$. En este caso, el modelo es:

$$\widehat{\text{US Gross}} = -8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budg}$$
$$- 15,000,000$$

Manteniendo lo demás constante, estimamos que una trilogía tendrá, en promedio, un recaudo 15 millones de dólares menor que una primera entrega

Volvamos a Anscombe...

Si estimamos $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times x$



En cada regresión, $\hat{\beta}_1 = 0.5$. Graficar los datos es crucial!