

Analítica de Datos

Regresión Lineal

Carlos Cardona Andrade

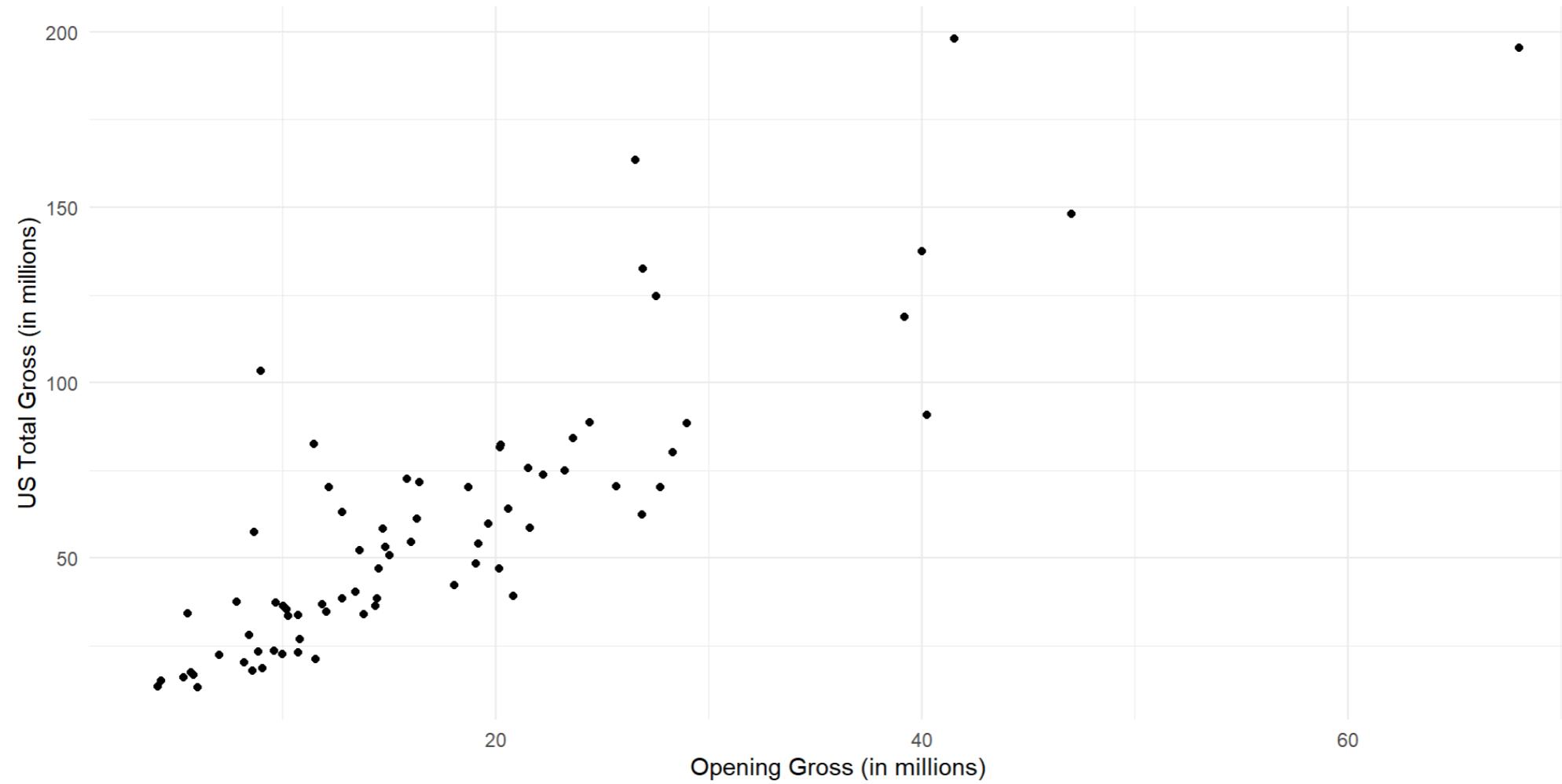
Tabla de contenido

1. Correlación
2. Regresión Lineal Simple
3. Regresión Lineal Múltiple

Correlación

US Total Gross vs Opening Gross

► Code



La correlación en R

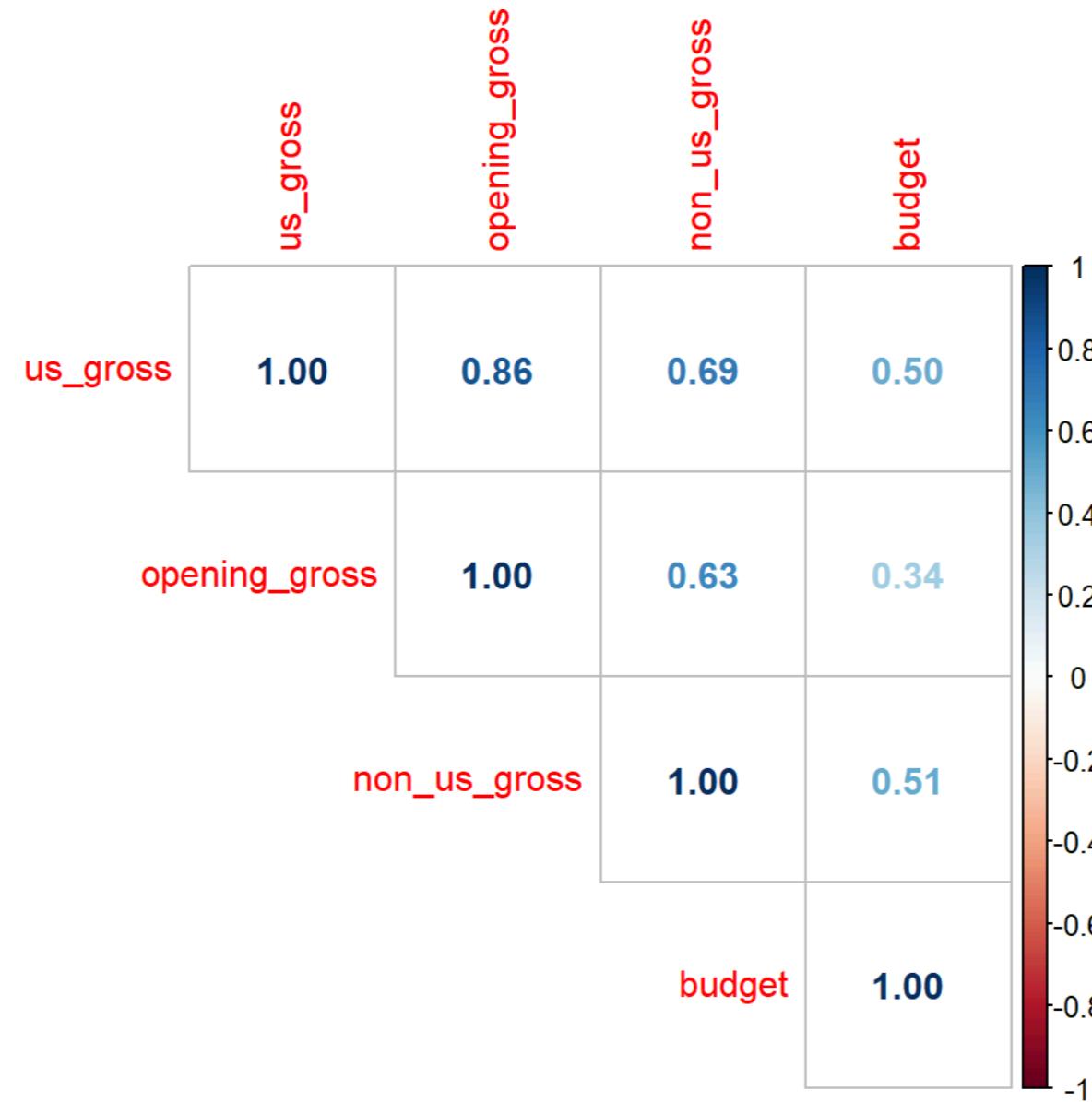
- Podemos utilizar un diagrama de dispersión para realizar un primer análisis de la relación entre dos variables
- El coeficiente de correlación (lineal) es utilizado para medir la fuerza de la asociación (lineal) entre dos variables
- La correlación entre el recaudo en US y el recaudo el primer fin de semana:

```
1 cor(hollywood$us_gross, hollywood$opening_gross)  
[1] 0.8586015
```

Función corrplot

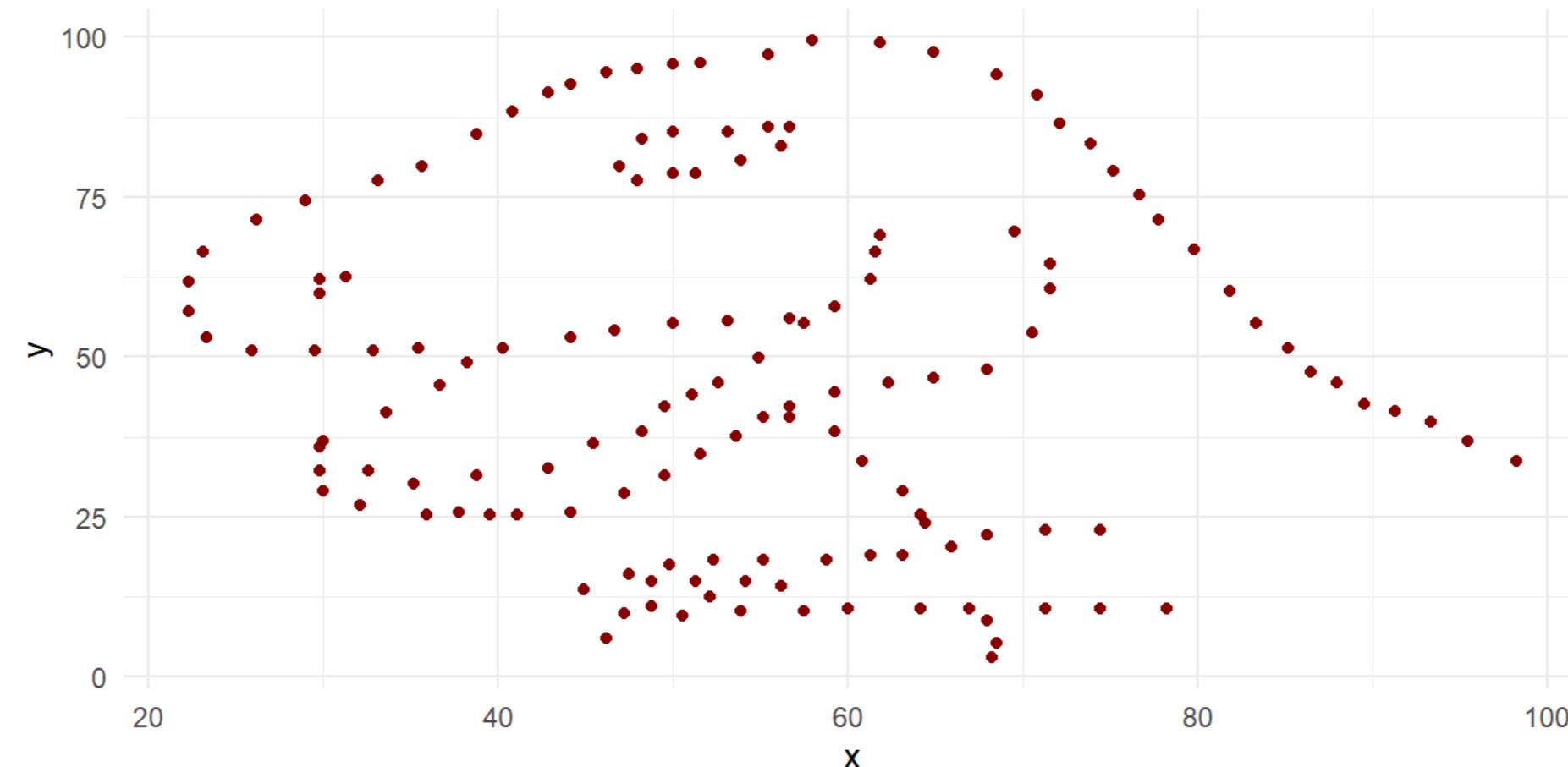
```
1 # RECUERDEN INSTALAR EL PAQUETE PRIMERO!!!
2 # install.packages("corrplot")
3
4 # Cargamos el paquete
5 library(corrplot)
6
7 # Creemos una base de datos temporal sólo con estas 4 variables
8 hollywood_sub <- hollywood %>% select( us_gross, opening_gross, non_us_gross, budget)
9
10 # Creemos la matriz de correlaciones
11 corrplot(cor(hollywood_sub ),
12           method = "number",
13           type = "upper")
```

Función corrplot



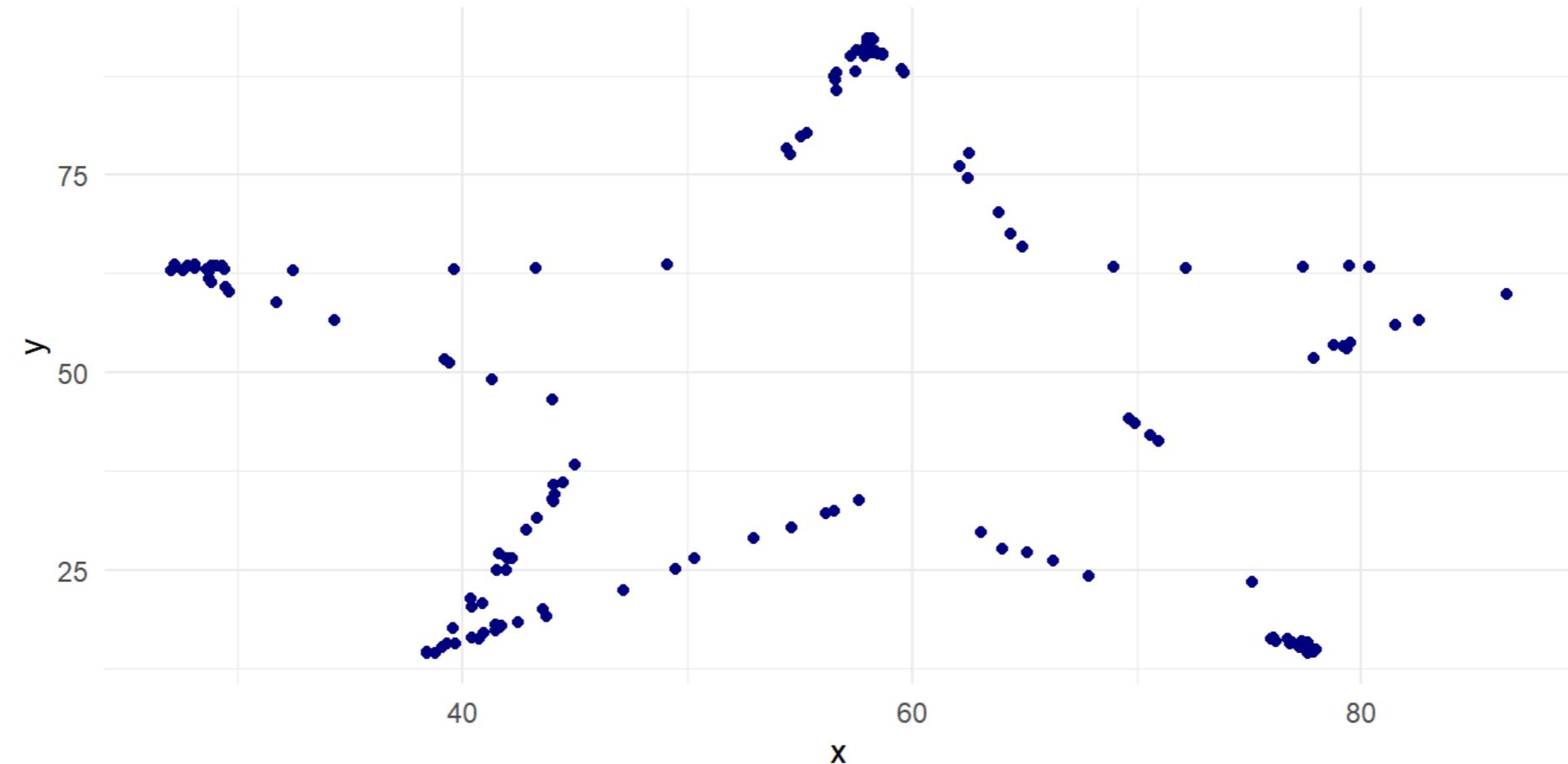
La correlación no es suficiente

```
# A tibble: 1 x 5
  mean_x  mean_y std_dev_x std_dev_y corr_x_y
  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1   54.3     47.8     16.8     26.9   -0.0645
```



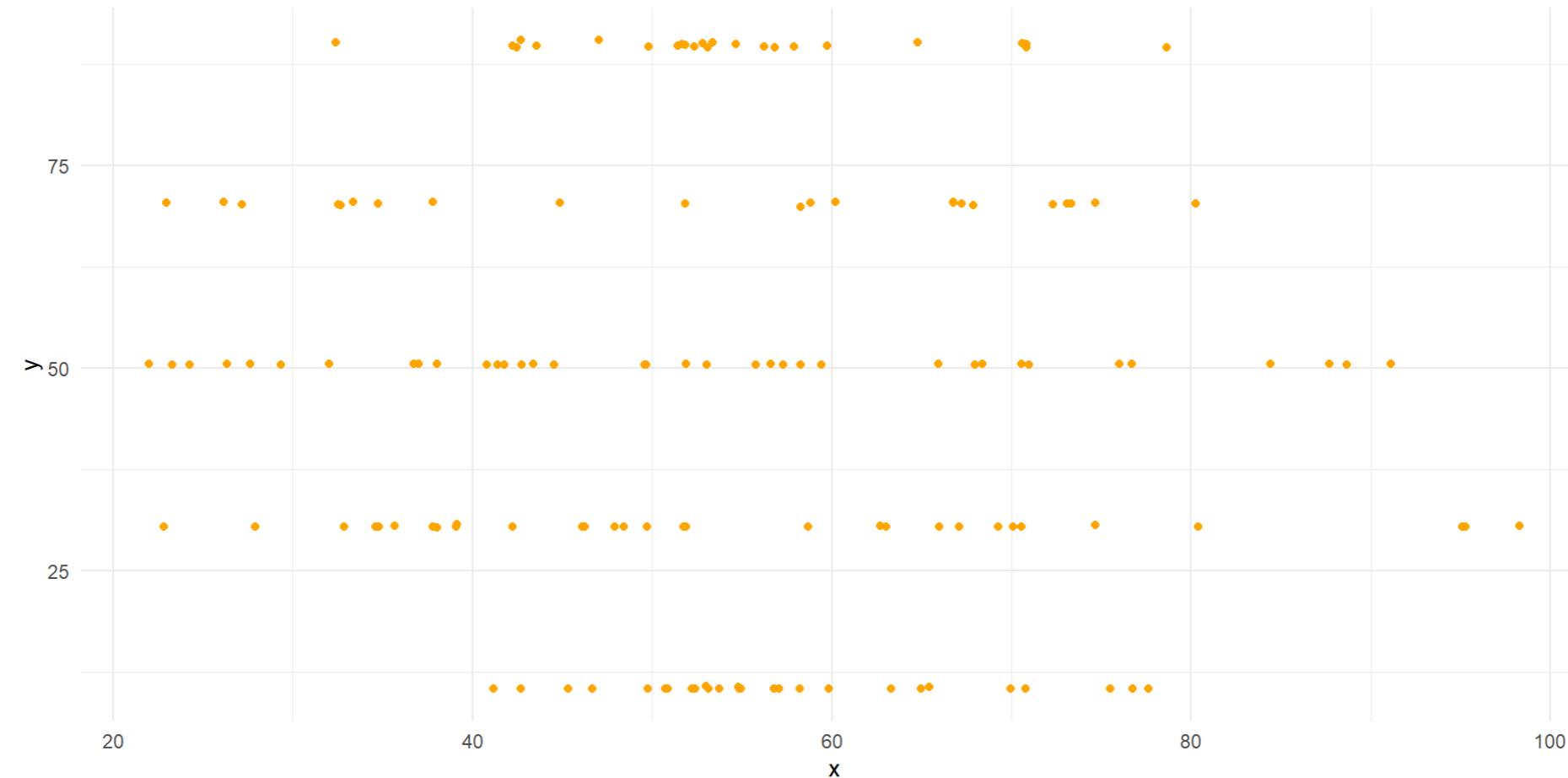
La correlación no es suficiente

```
# A tibble: 1 x 5
  mean_x  mean_y std_dev_x std_dev_y corr_x_y
  <dbl>    <dbl>     <dbl>     <dbl>      <dbl>
1   54.3     47.8     16.8     26.9    -0.0630
```



La correlación no es suficiente

```
# A tibble: 1 x 5
  mean_x  mean_y std_dev_x std_dev_y corr_x_y
  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 54.3     47.8     16.8     26.9   -0.0617
```



Regresión Lineal

Simple

Partes esenciales de una regresión

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

¿Por qué 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

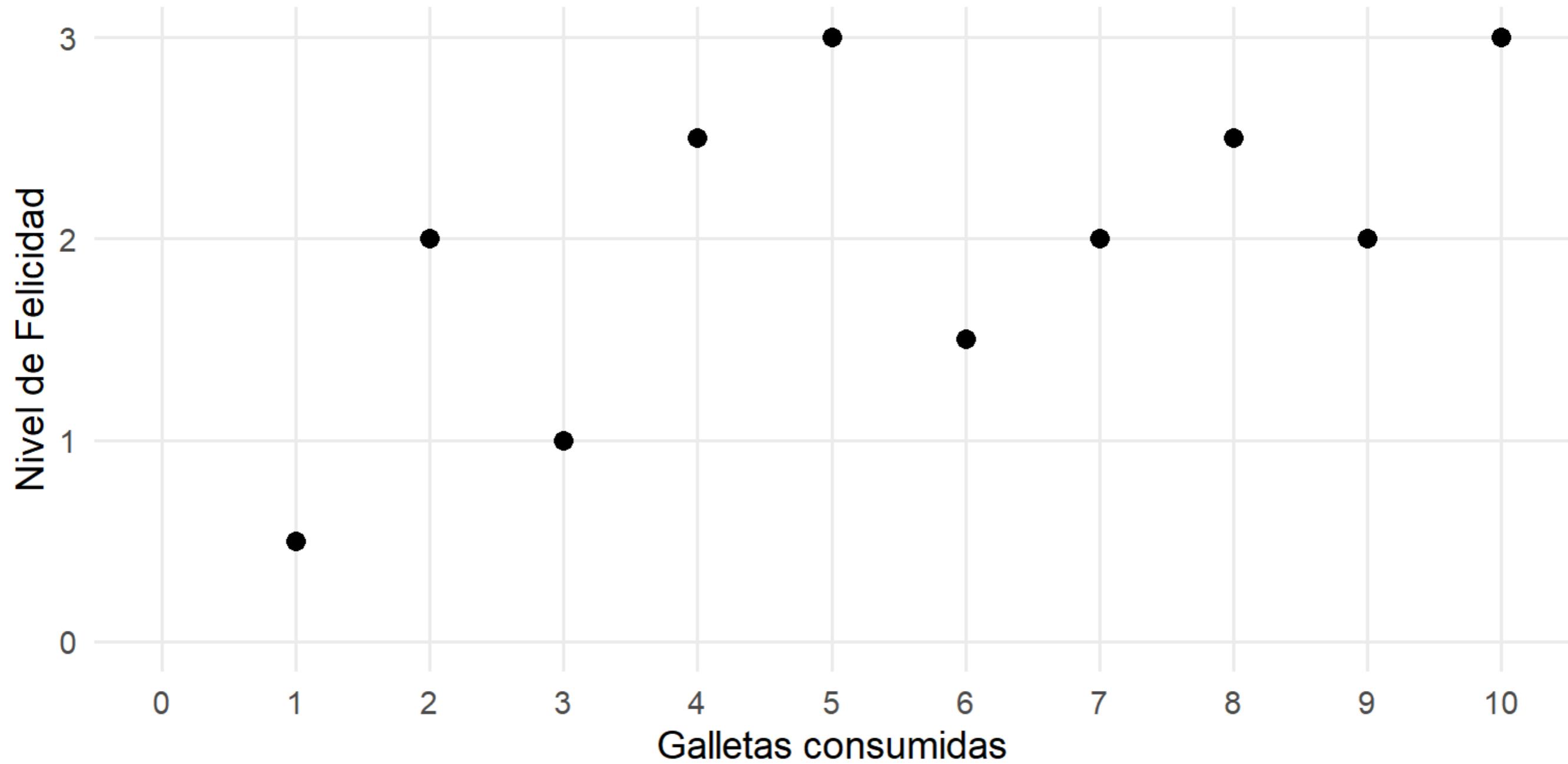
- Explicar el efecto de X en Y
- Nos enfocamos en X

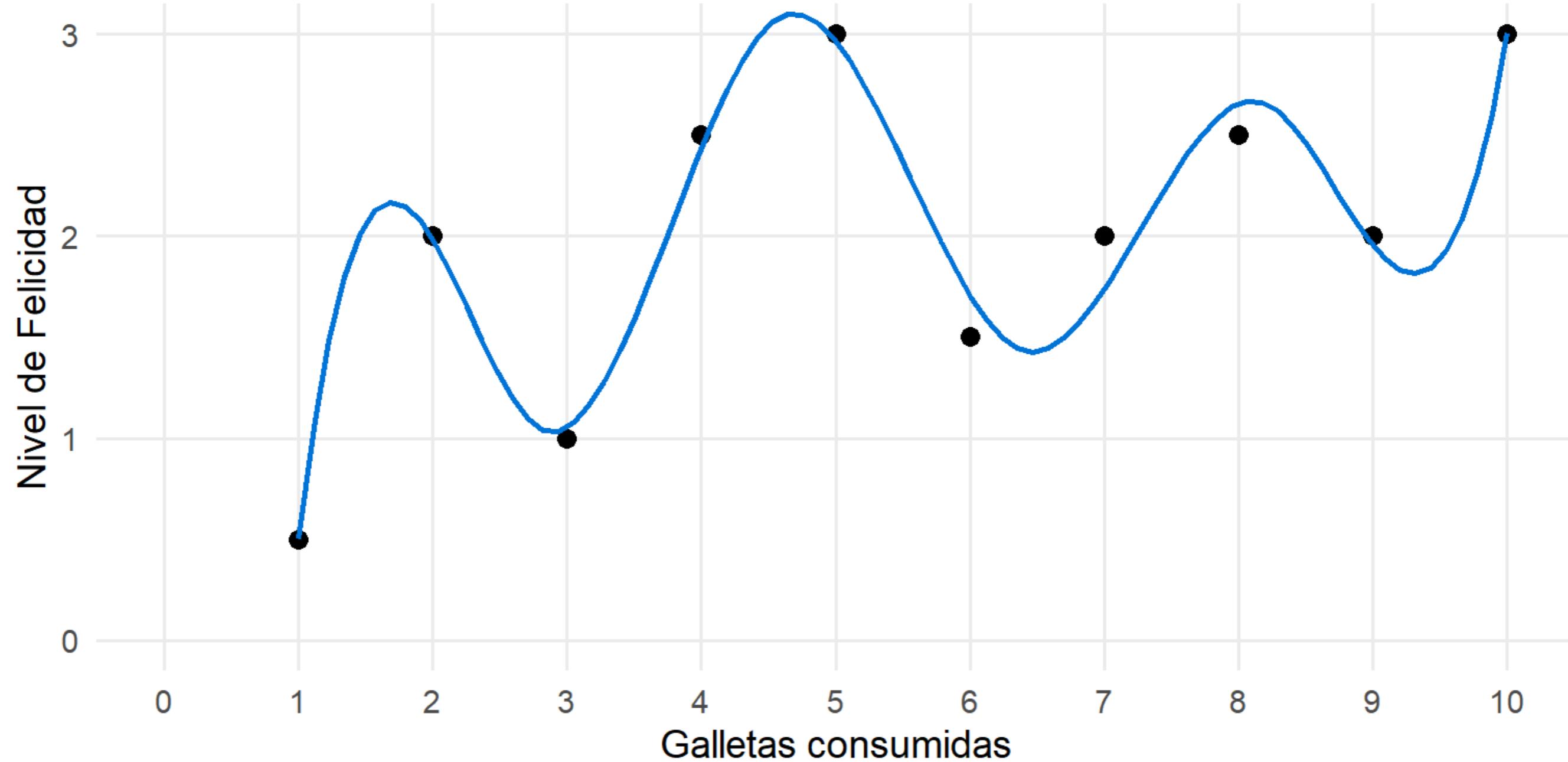
¿Cómo?

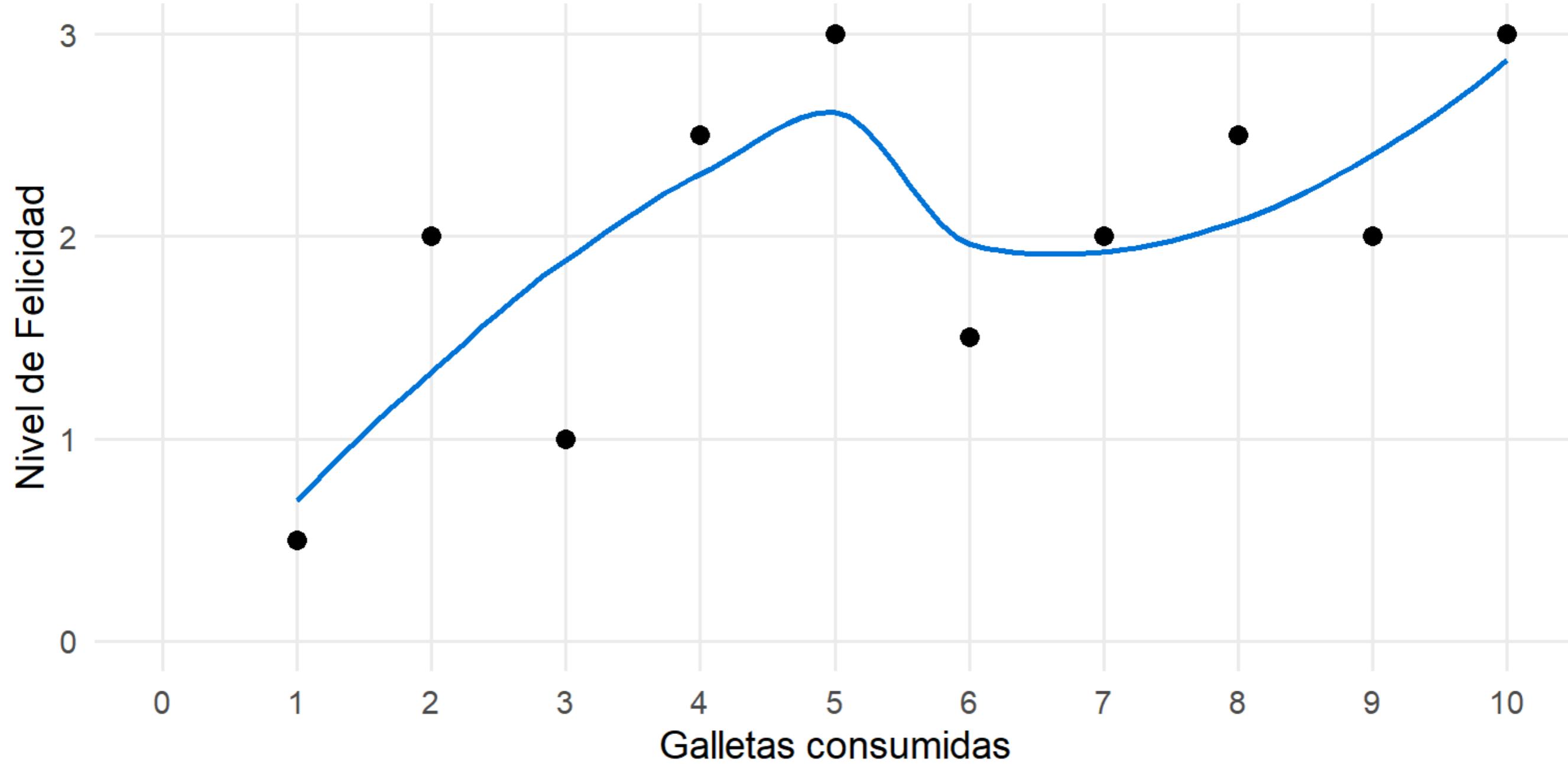
1. Graficar X y Y
2. Dibujar una recta que se aproxime a la relación observada (ojalá funcione para datos que no están en la muestra)
3. Estimar los números que componen esa recta
4. Interpretar esos números

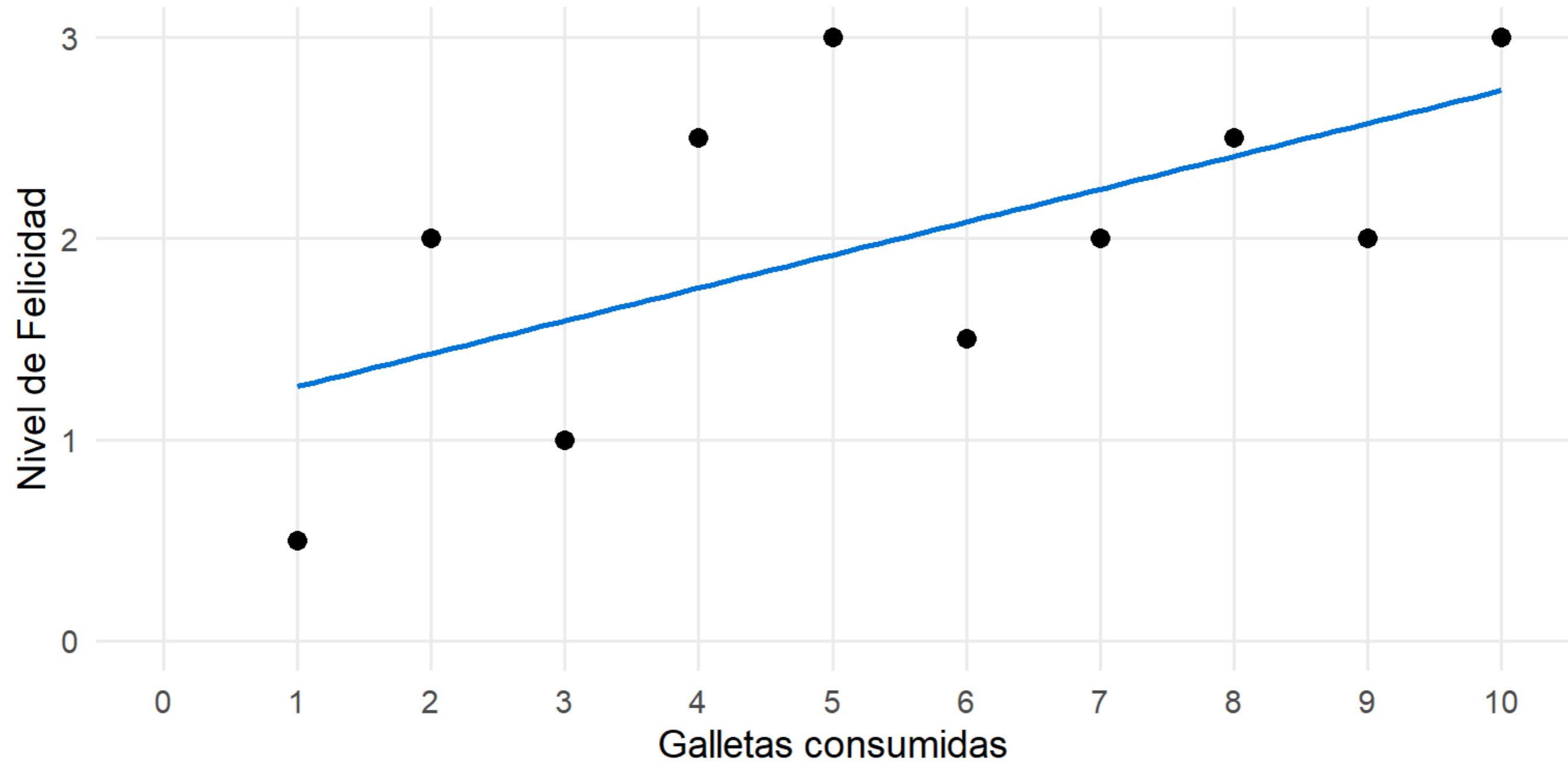
Galletas y Felicidad

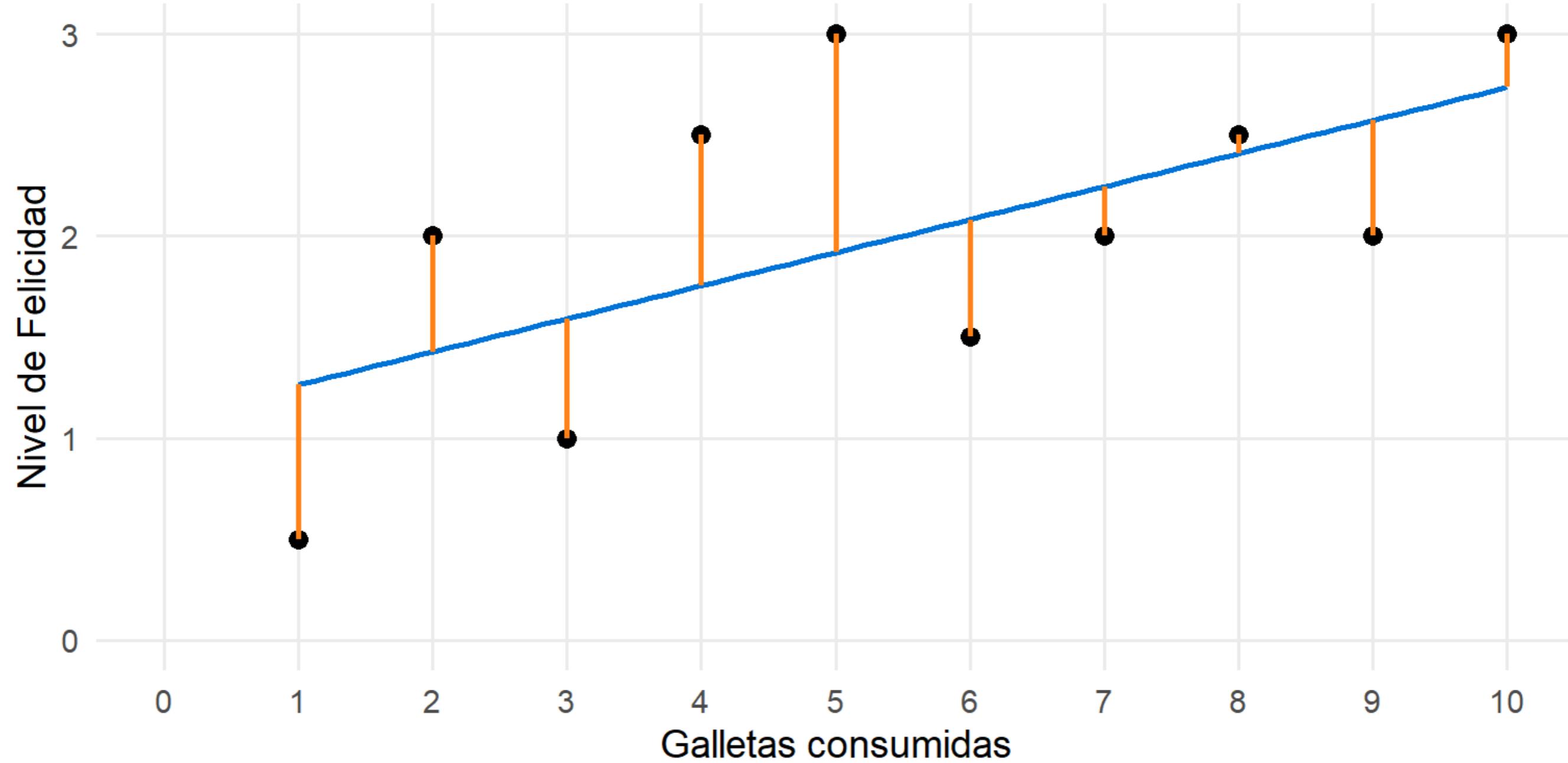
felicidad	galletas
0.5	1
2.0	2
1.0	3
2.5	4
3.0	5
1.5	6
2.0	7
2.5	8
2.0	9
3.0	10

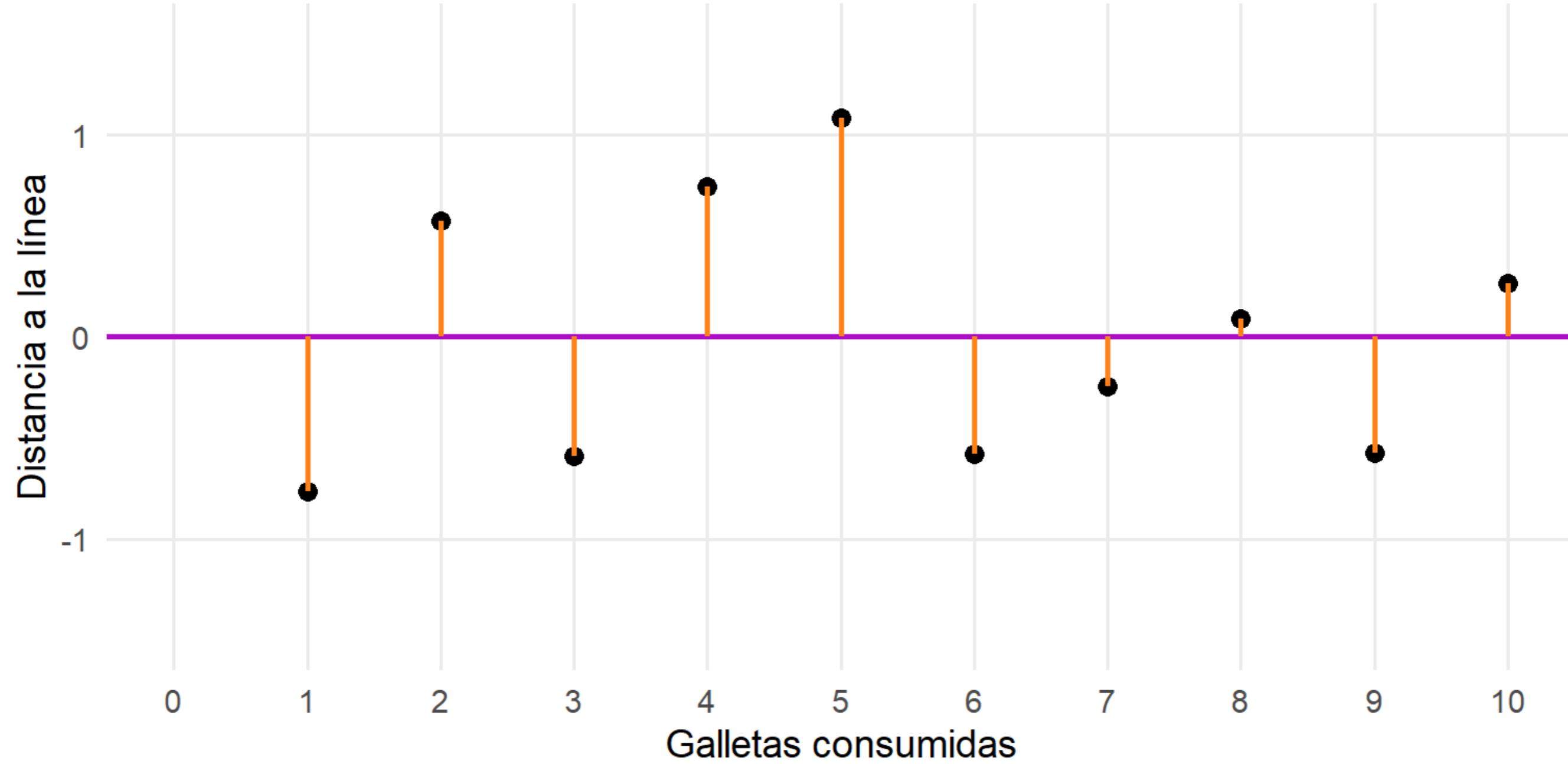




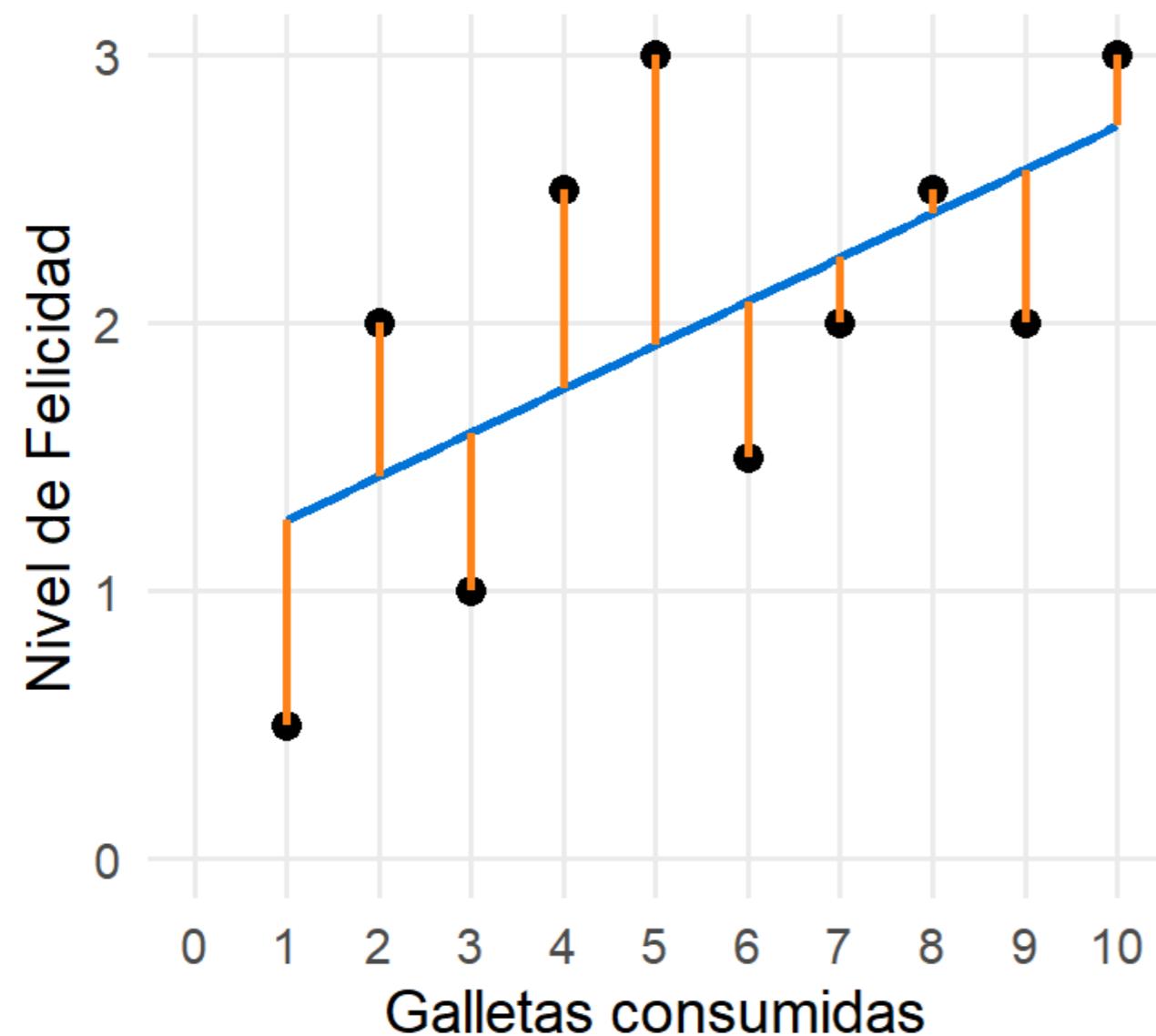




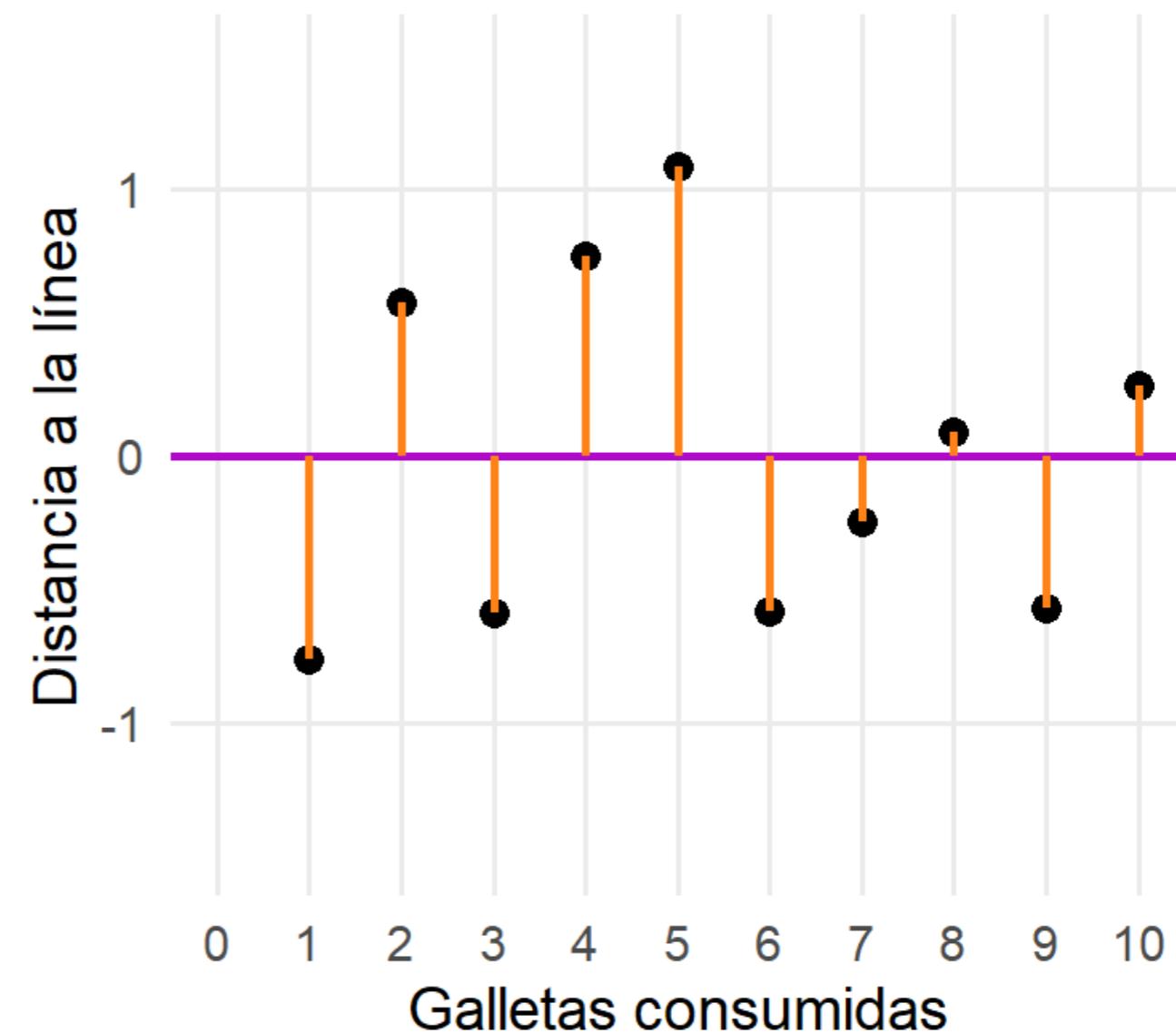




Galletas y Felicidad



Residuos



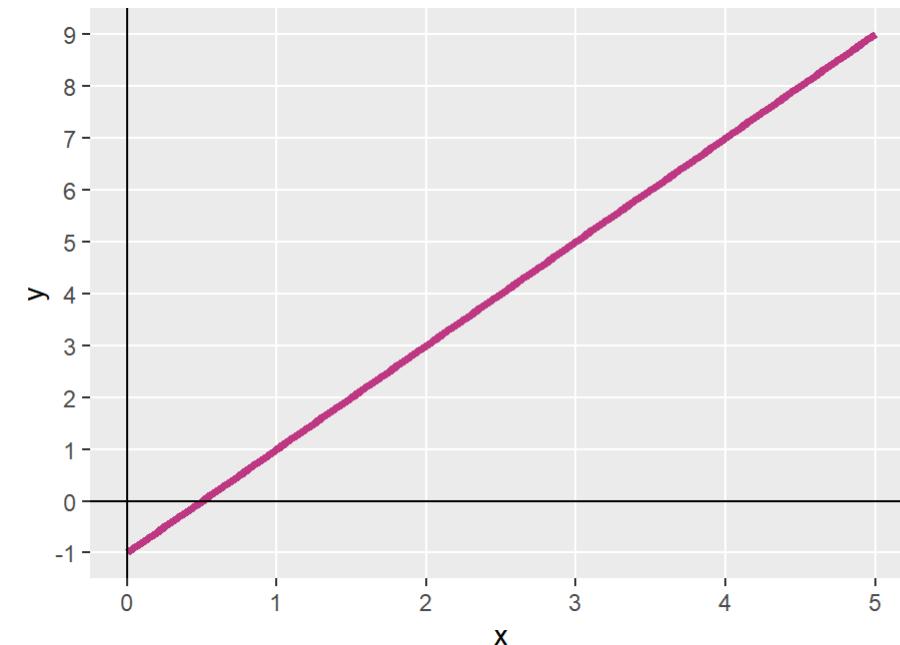
Pendiente de una recta

$$y = mx + b$$

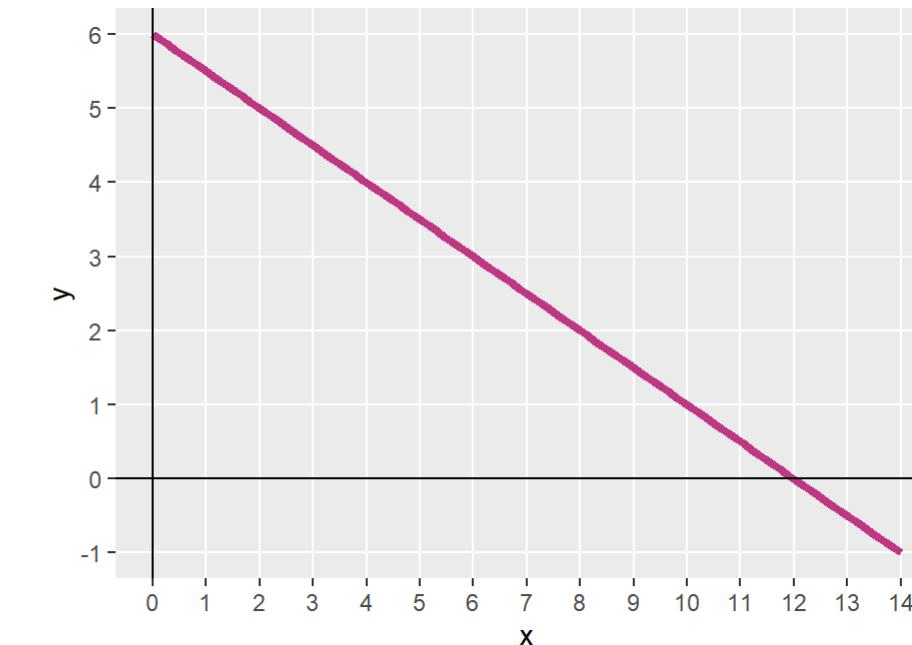
y	Un número
x	Un número
m	La pendiente $\frac{\Delta y}{\Delta x}$
b	El intercepto con y

Pendiente de una recta

$$y = 2x - 1$$



$$y = -0.5x + 6$$



Regresión lineal simple

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

- β_1 : la pendiente verdadera de la relación entre X y Y
- β_0 : el intercepto verdadero de la relación entre X y Y
- ε : el error

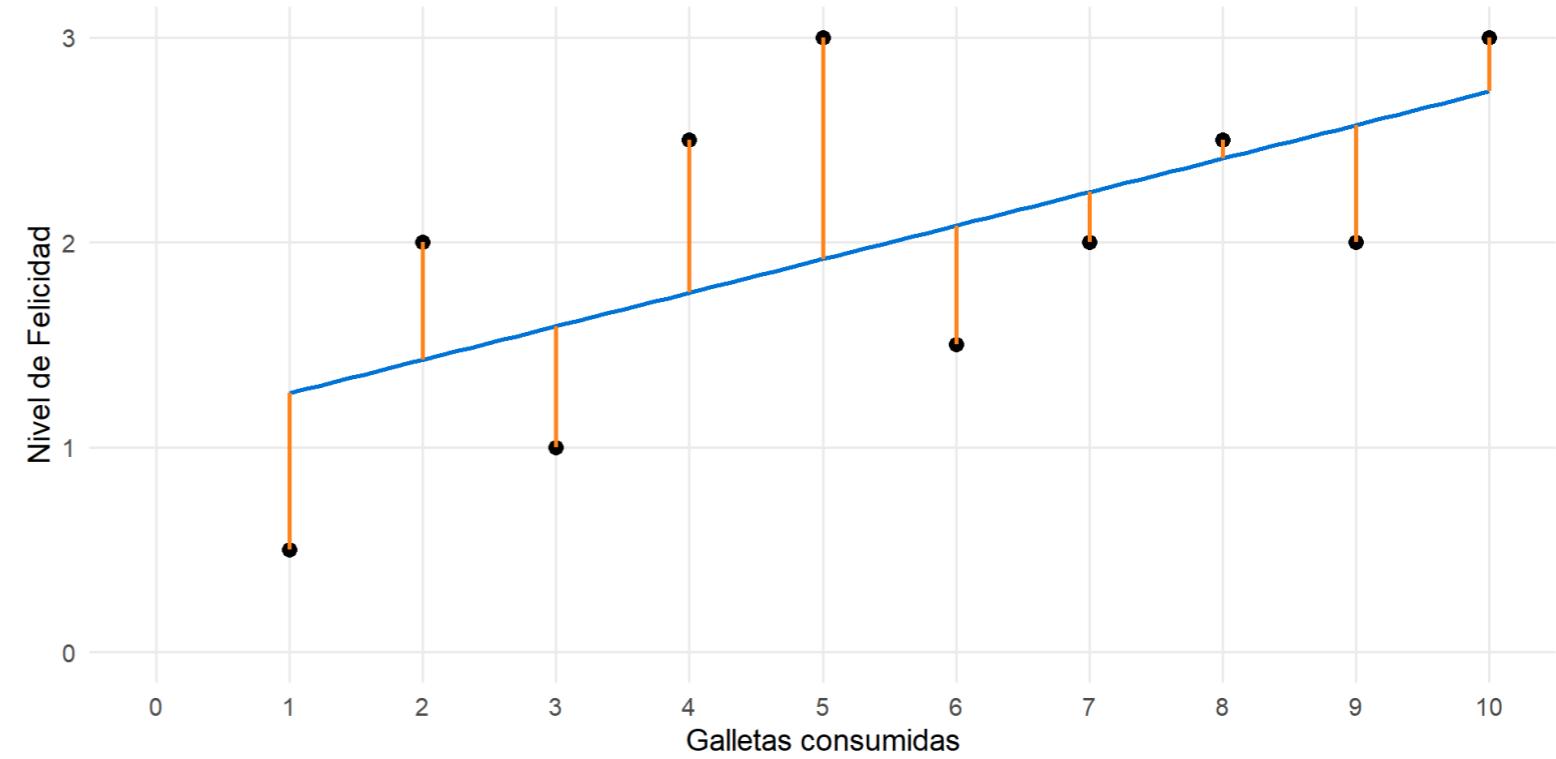
Regresión lineal simple

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

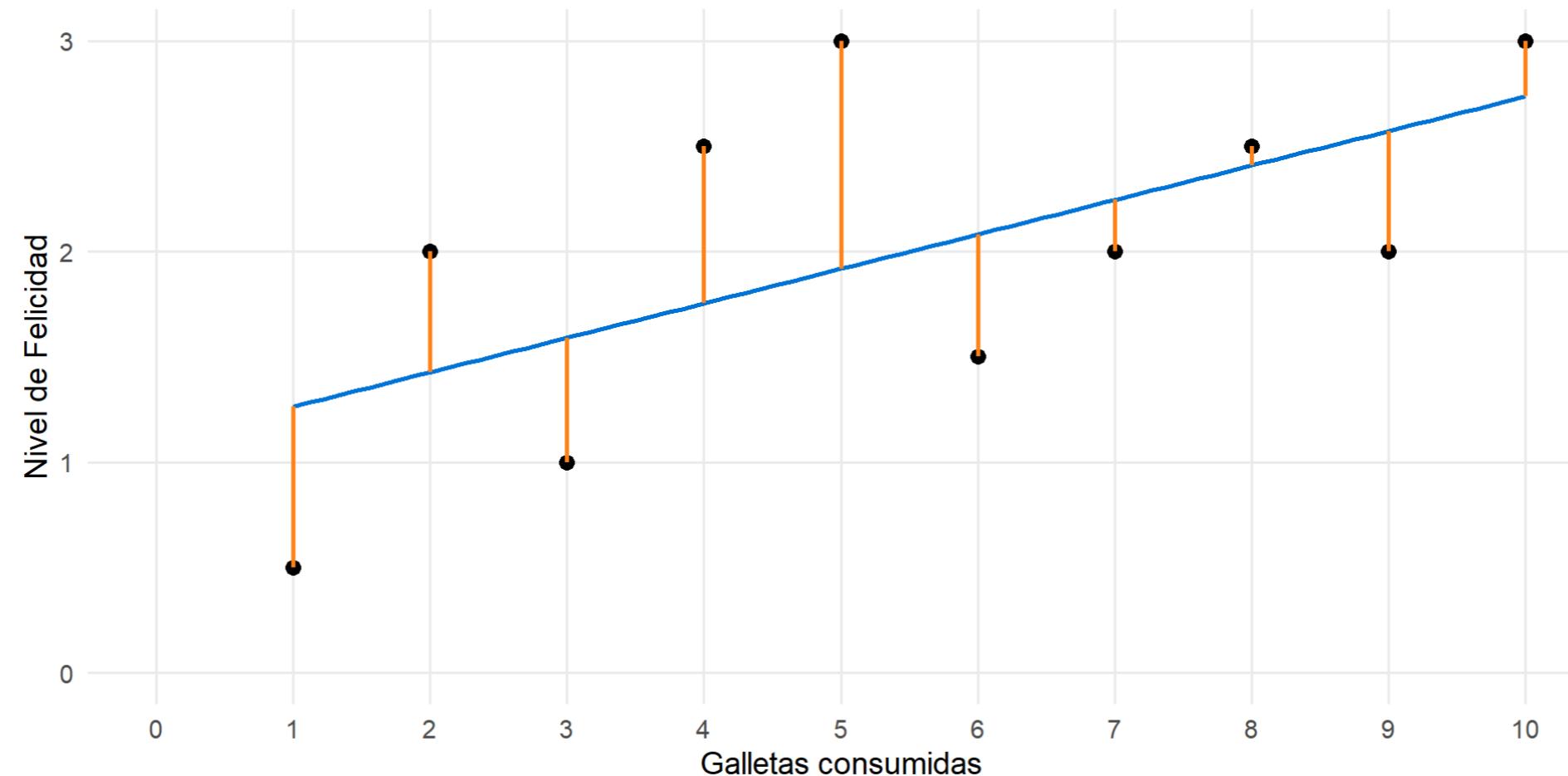
- $\hat{\beta}_1$: la pendiente estimado de la relación entre X y Y
- $\hat{\beta}_0$: el intercepto estimado de la relación entre X y Y
- No hay error!!!

Modelo de Regresión

$$\begin{aligned}Y &= \text{Modelo} + \text{Error} \\&= f(X) + \varepsilon\end{aligned}$$



Residuos



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

La línea de los mínimos cuadrados

- El residuo para la observación i^{th} es:

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

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

$$e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2$$

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



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)
```

Construyendo modelos en R

- La sintaxis para los resultados del modelo es:

```
1 name_of_model <- lm(Y ~ X, data = DATA)
2
3 summary(name_of_model) # Para ver los detalles del modelo
```

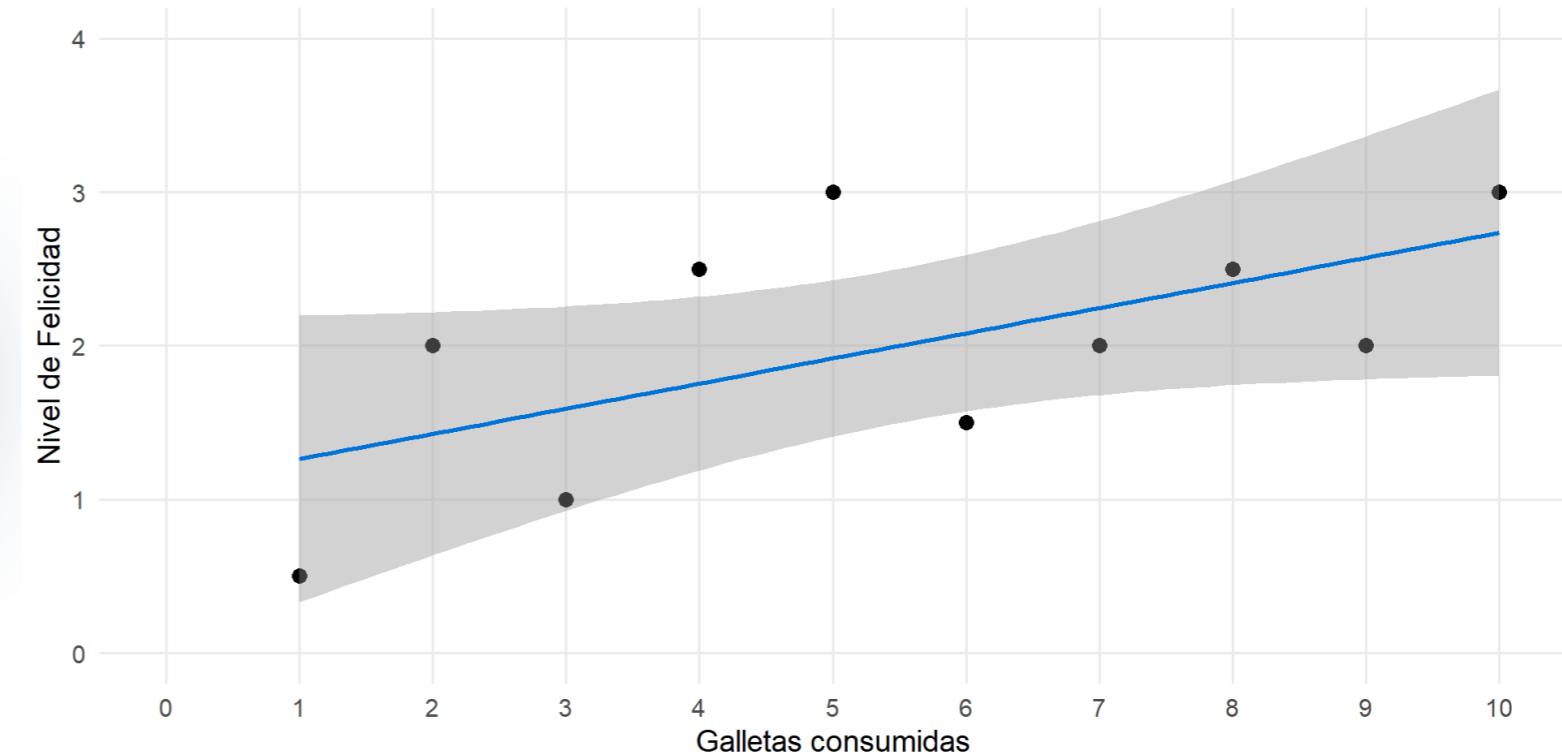
- Otras opciones para evaluar el modelo son:

```
1 library(broom)
2
3 # Convierte los resultados del modelo a un data frame para graficar
4 tidy(name_of_model)
5
6 # Convierte los parámetros que evalúan el modelo a un data frame
7 glance(name_of_model)
```

Modelando Galletas y Felicidad

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

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



Modelando Galletas y Felicidad

Podemos ver los coeficientes, error estándar, p-value e IC:

```
1 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
```

Para ver aspectos evaluando el ajuste del modelo:

```
1 glance(modelo_felicidad)

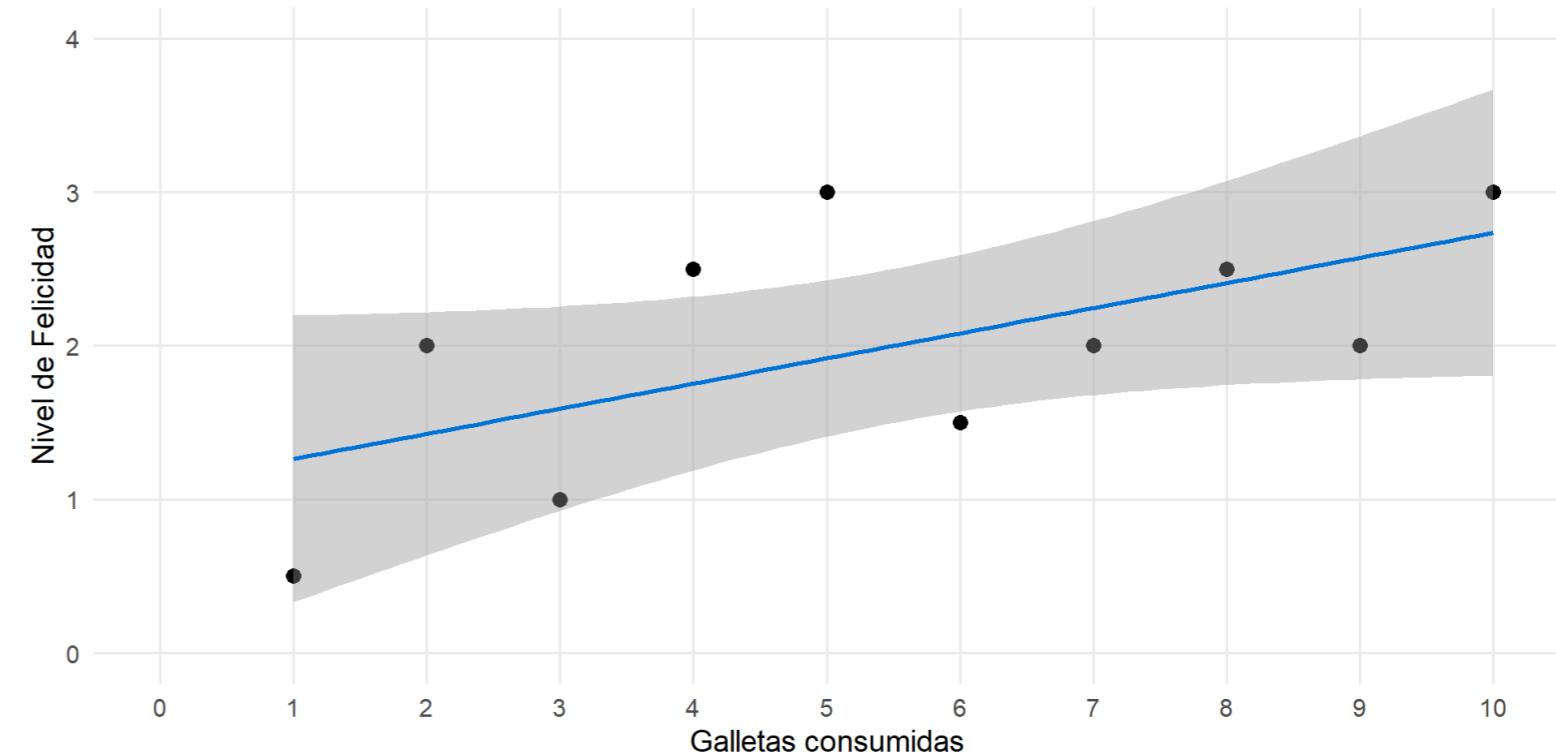
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic p.value      df logLik     AIC     BIC
  <dbl>          <dbl> <dbl>     <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl>
1 0.368         0.289 0.688     4.66   0.0629     1  -9.34  24.7  25.6
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Traduciendo los resultados a matemáticas

```
# A tibble: 2 x 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 aumento en la felicidad de 0.16 unidades
- Si no hay consumo de galletas, esperamos que el puntaje de felicidad sea 1.1 unidades

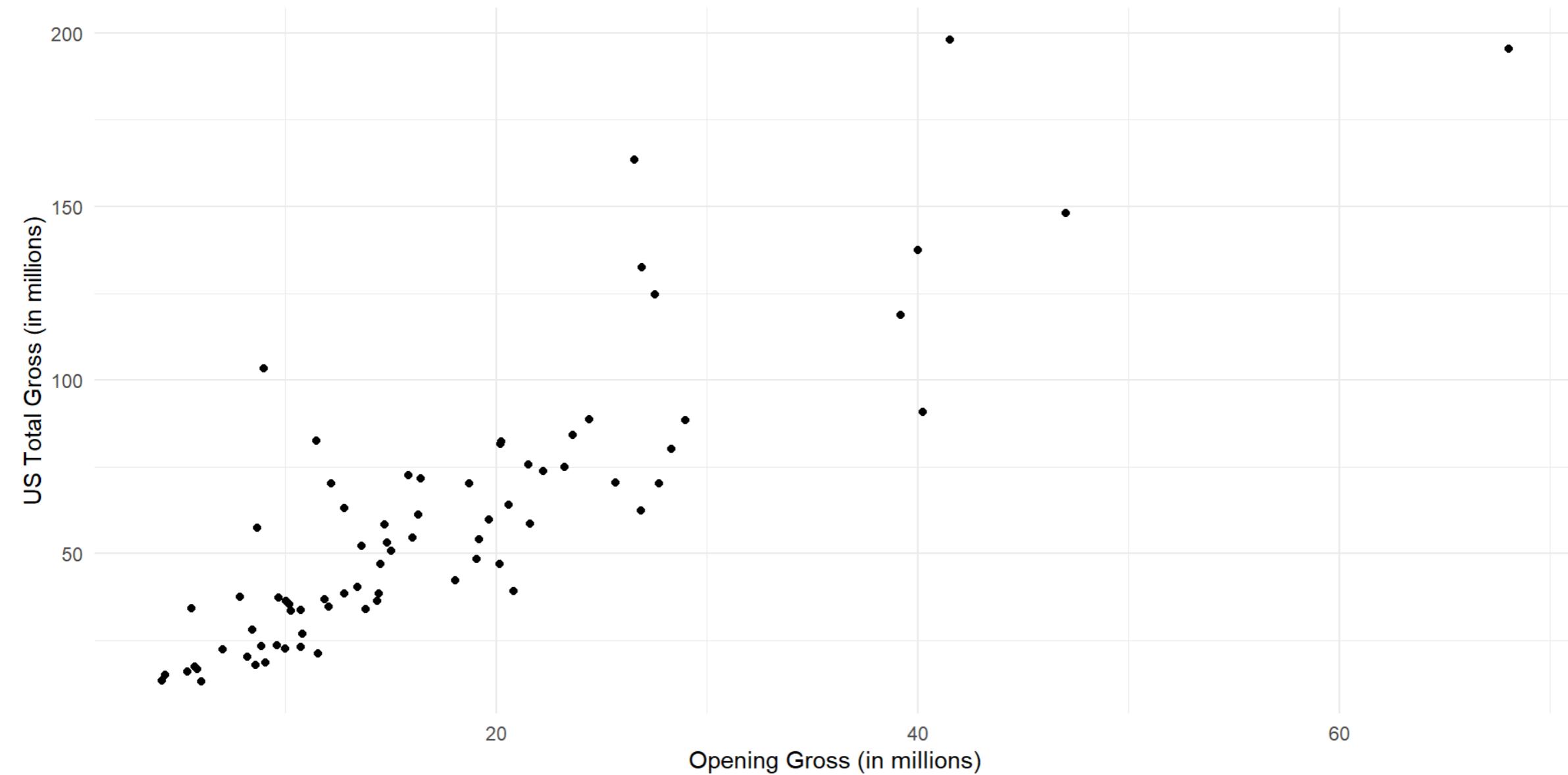
¿Es el intercepto importante?

- 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...

Volvamos a Hollywood Rules

- Según la sabiduría popular en Hollywood, el recaudo durante el primer fin de semana es un fuerte predictor del éxito comercial de una película
- Grafiquemos la relación entre el recaudo en Estados Unidos y el recaudo en el primer fin de semana para evaluar esta creencia:

Volvamos a Hollywood Rules



Volvamos a Hollywood Rules

```
1 hollywood_model <- lm(us_gross ~ opening_gross, data=hollywood)
2 tidy(hollywood_model, conf.int = TRUE)

# A tibble: 2 x 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
```

Entonces nuestro modelo lineal es:

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

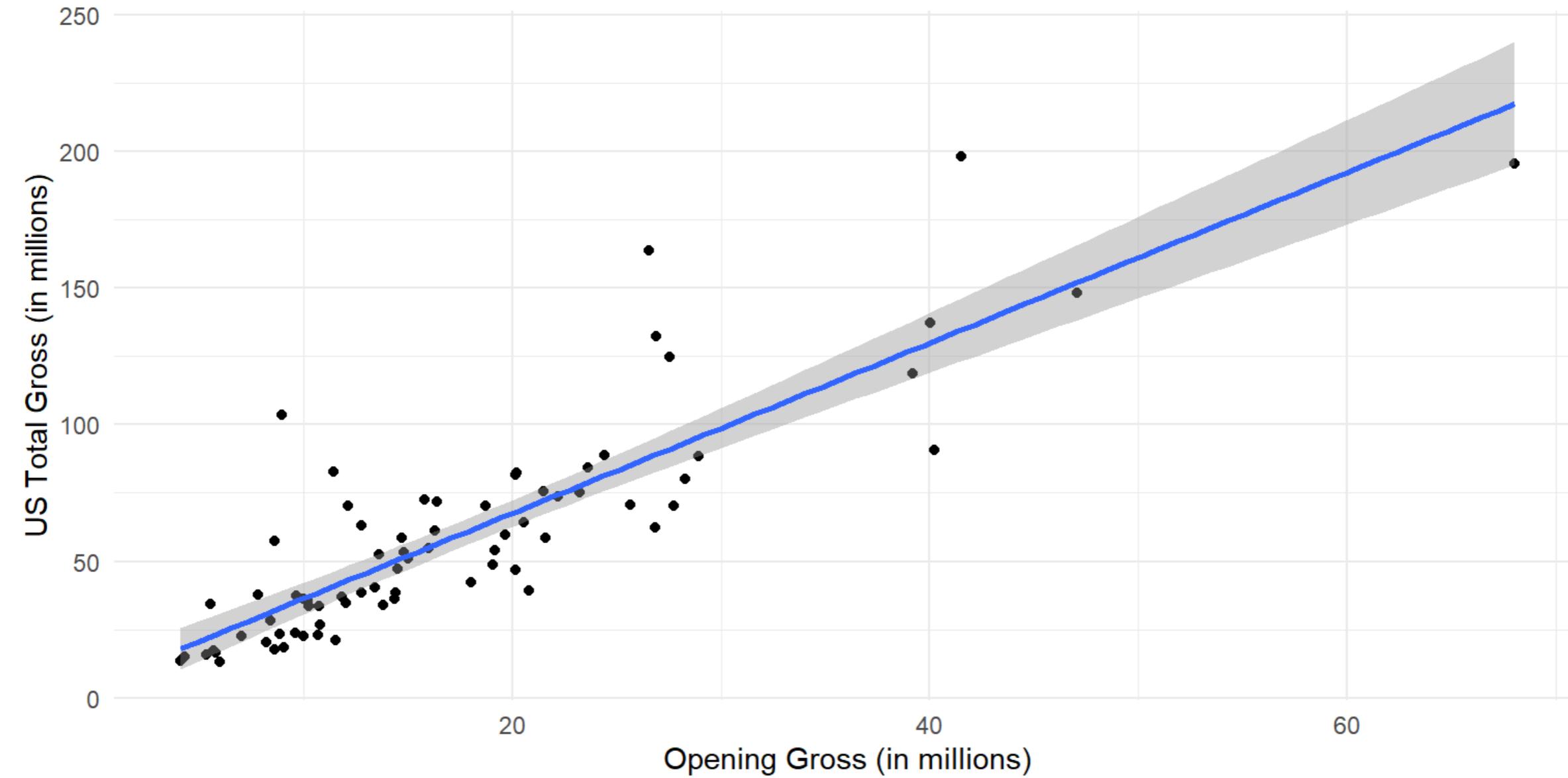
¿Cuál es la interpretación de $\hat{\beta}_1$ en este caso? ¿Y de $\hat{\beta}_0$?

¿Cómo graficar la línea de regresión?

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

```
1 ggplot(data = hollywood, aes(x = (opening_gross / 1000000), y = (us_gross / 1000000))) +  
2   geom_point() +  
3   geom_smooth(method="lm") +  
4   labs(  
5     x = "Opening Gross (in millions)",  
6     y = "US Total Gross (in millions)"  
7   ) +  
8   theme_minimal()
```

¿Cómo graficar la línea de regresión?



Predicción

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?

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

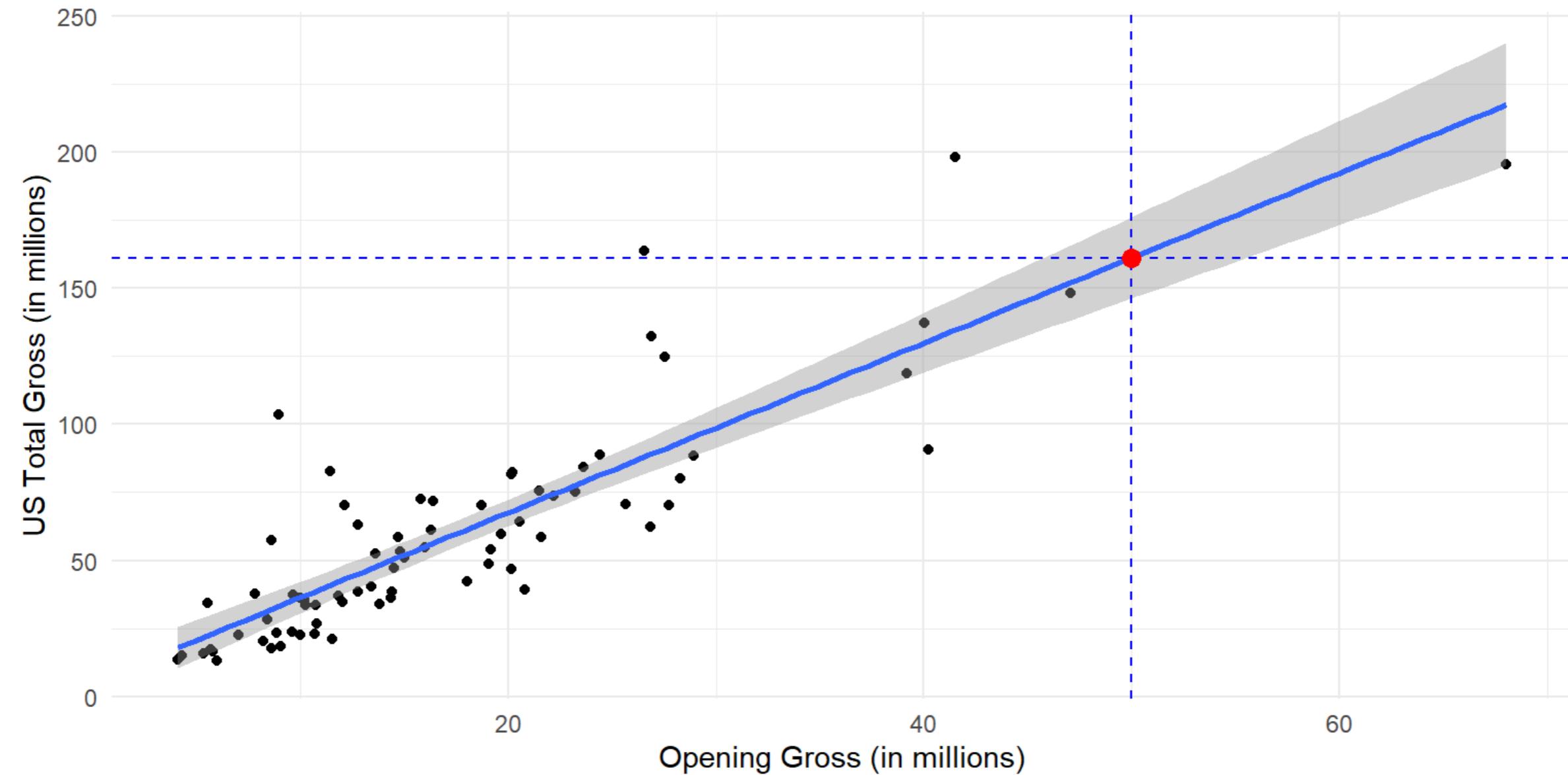
Predicción con R

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

```
1 # Creamos los valores para los cuales queremos predecir
2 valores_opening <- data.frame(opening_gross = c(20000000,40000000,50000000))
3
4
5 # Predice los valores con los coeficientes estimados
6 # en hollywood_model
7 predict(hollywood_model, newdata = valores_opening)
```

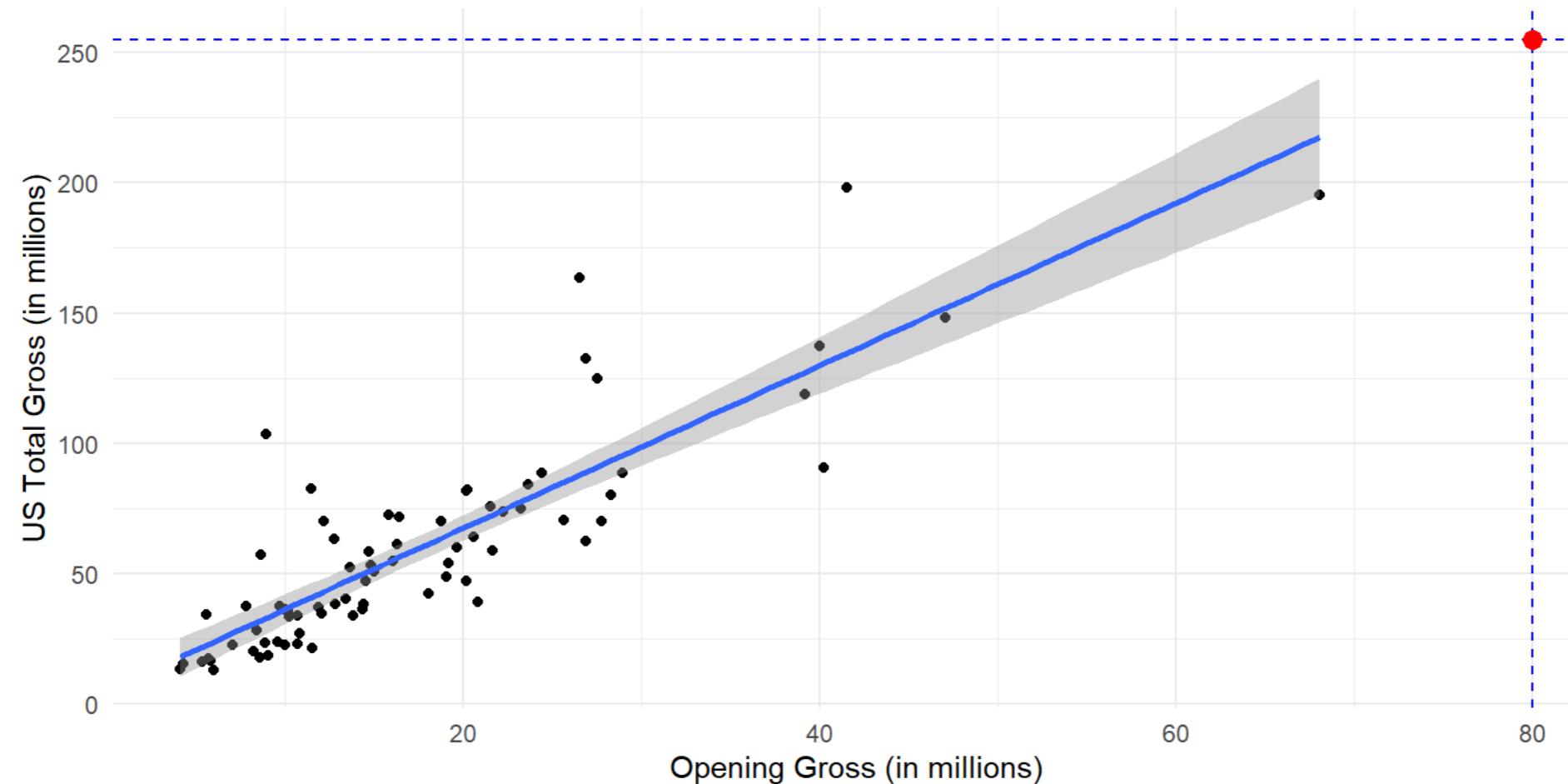
1	2	3
67520606	129932991	161139184

Predicción



¿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?

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

# A tibble: 2 x 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

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$Z = \frac{3.12 - 0}{0.218} = 14.3 > Z_{\frac{\alpha}{2}} = 1.96$$

- Rechazamos la H_0 a un nivel de significancia del 5%!
- El p-value es 7.07e-23 (en notación científica), el cual es mucho menor a 0.05.

Regresión Lineal

Múltiple

Regresión Múltiple

No estamos limitados a una sola variable explicativa!

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

```
1 hollywood_model <- lm(us_gross ~ opening_gross + budget + sequel, data=hollywood)
```

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

Regresión Múltiple

```
1 hollywood_model <- lm(us_gross ~ opening_gross + budget + sequel, data=hollywood)
2 tidy(hollywood_model, conf.int = TRUE)
```

```
# A tibble: 4 x 7
  term            estimate  std.error statistic  p.value  conf.low  conf.high
  <chr>          <dbl>    <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
1 (Intercept) -8785254.   5438753.   -1.62 1.11e- 1 -1.96e+7  2.06e+6
2 opening_gross      2.99     0.232     12.9 3.11e-20  2.52e+0  3.45e+0
3 budget           0.356     0.101     3.52 7.53e- 4  1.54e-1  5.57e-1
4 sequel          -11929834.  7788684.   -1.53 1.30e- 1 -2.75e+7  3.60e+6
```

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

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.

```
1 multiples_valores <- data.frame(opening_gross=50000000,  
2                                     budget=100000000,  
3                                     sequel=1)  
4 predict(hollywood_model, newdata = multiples_valores)
```

```
1  
164132342
```

Predicción en Regresión Múltiple

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

```
1 # Seleccionemos las 3 variables dependientes para The Holiday  
2 the_holiday <- hollywood %>%  
3   filter(movie=="The Holiday") %>%  
4   select(opening_gross, budget, sequel)  
5  
6 # Usamos el comando predict nuevamente  
7 predicho <- predict(hollywood_model, newdata = the_holiday)  
8 predicho
```

```
1  
59596928
```

Predicción en Regresión Múltiple

- ¿Es precisa nuestra estimación?

```
1 # El valor observado  
2 observado <- hollywood %>%  
3   filter(movie == "The Holiday") %>%  
4   pull(us_gross)  
5 observado
```

```
[1] 63224849
```

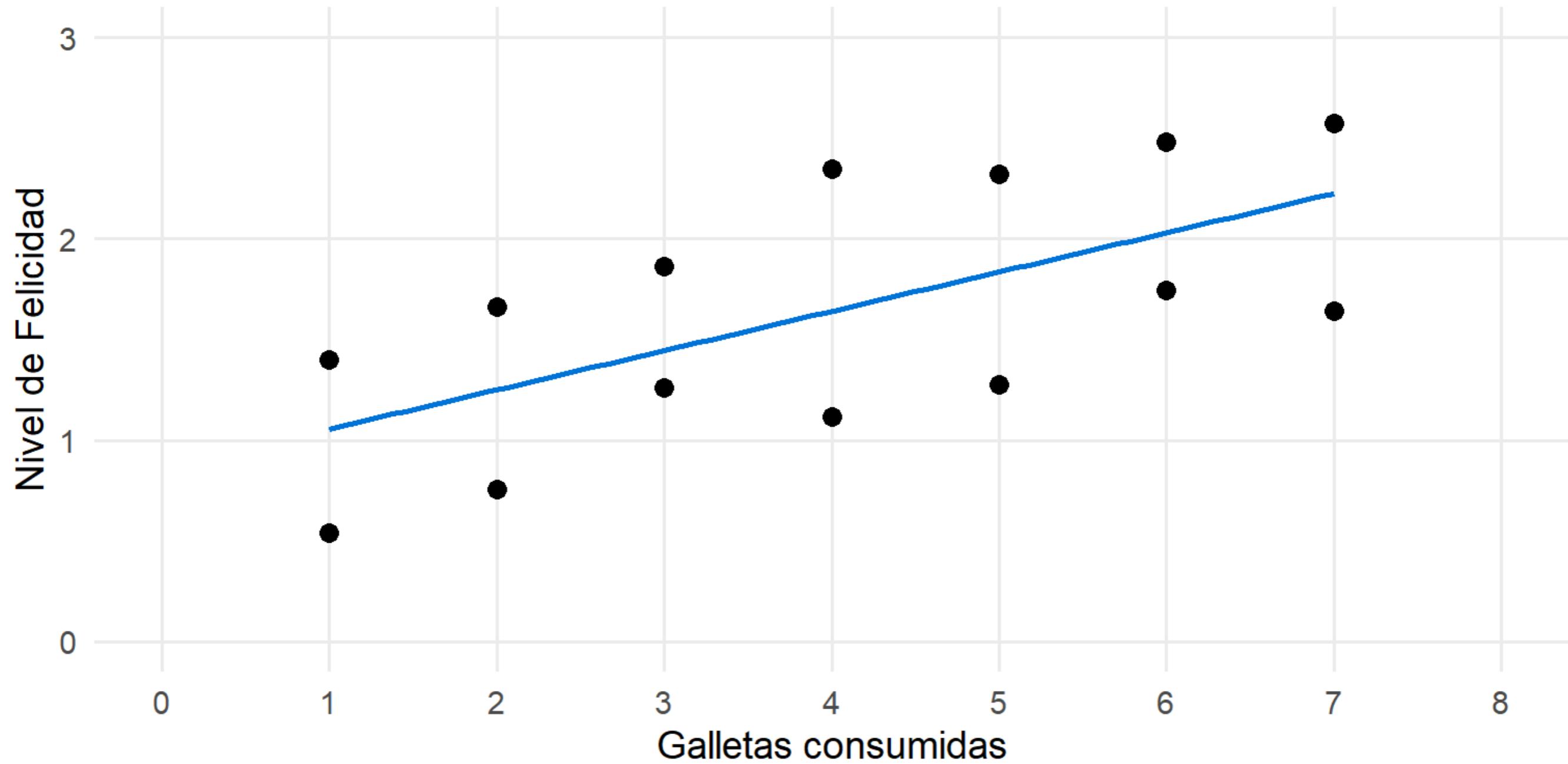
- El residuo para “The Holiday” será:

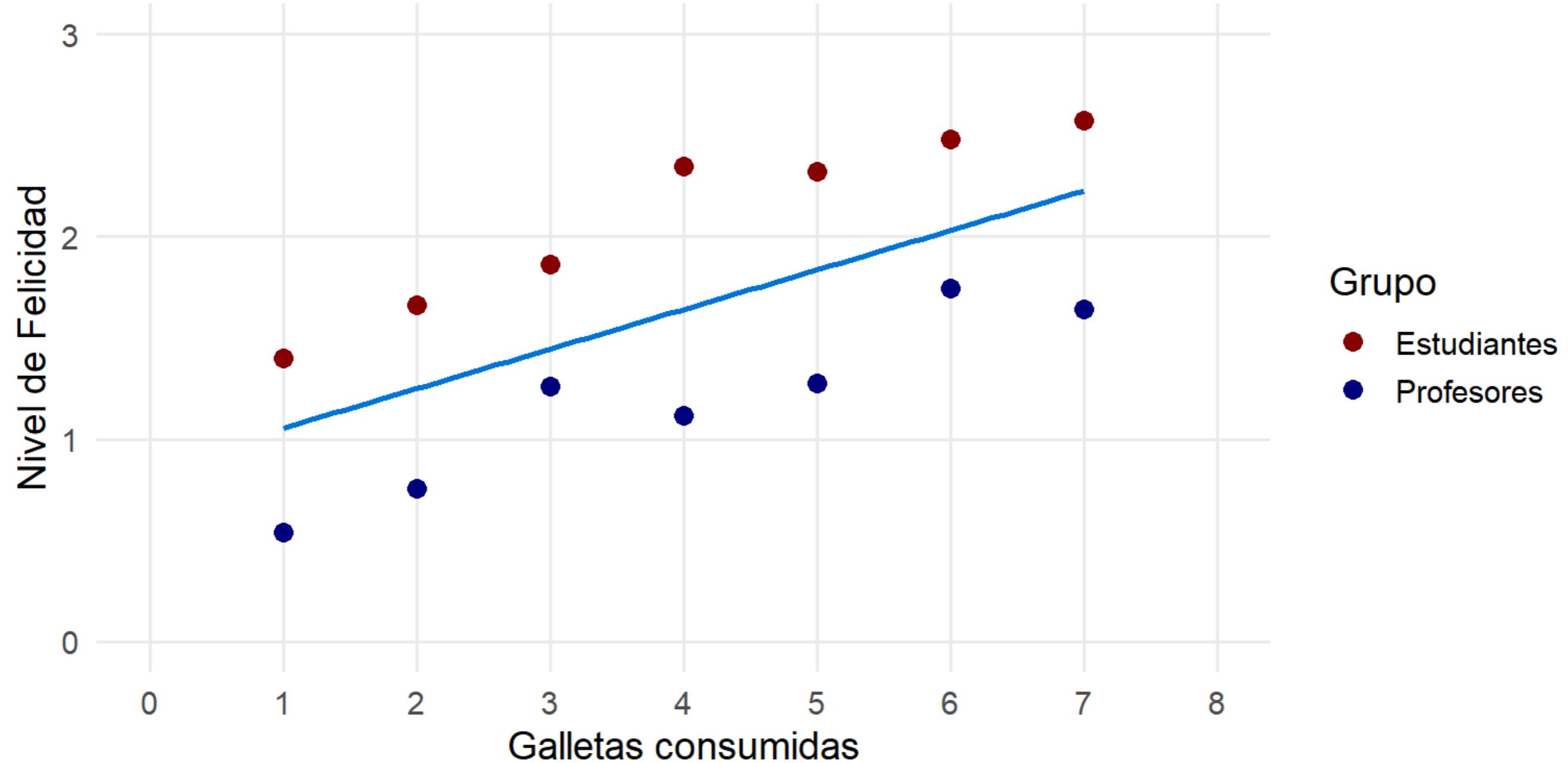
```
1 observado-predicho
```

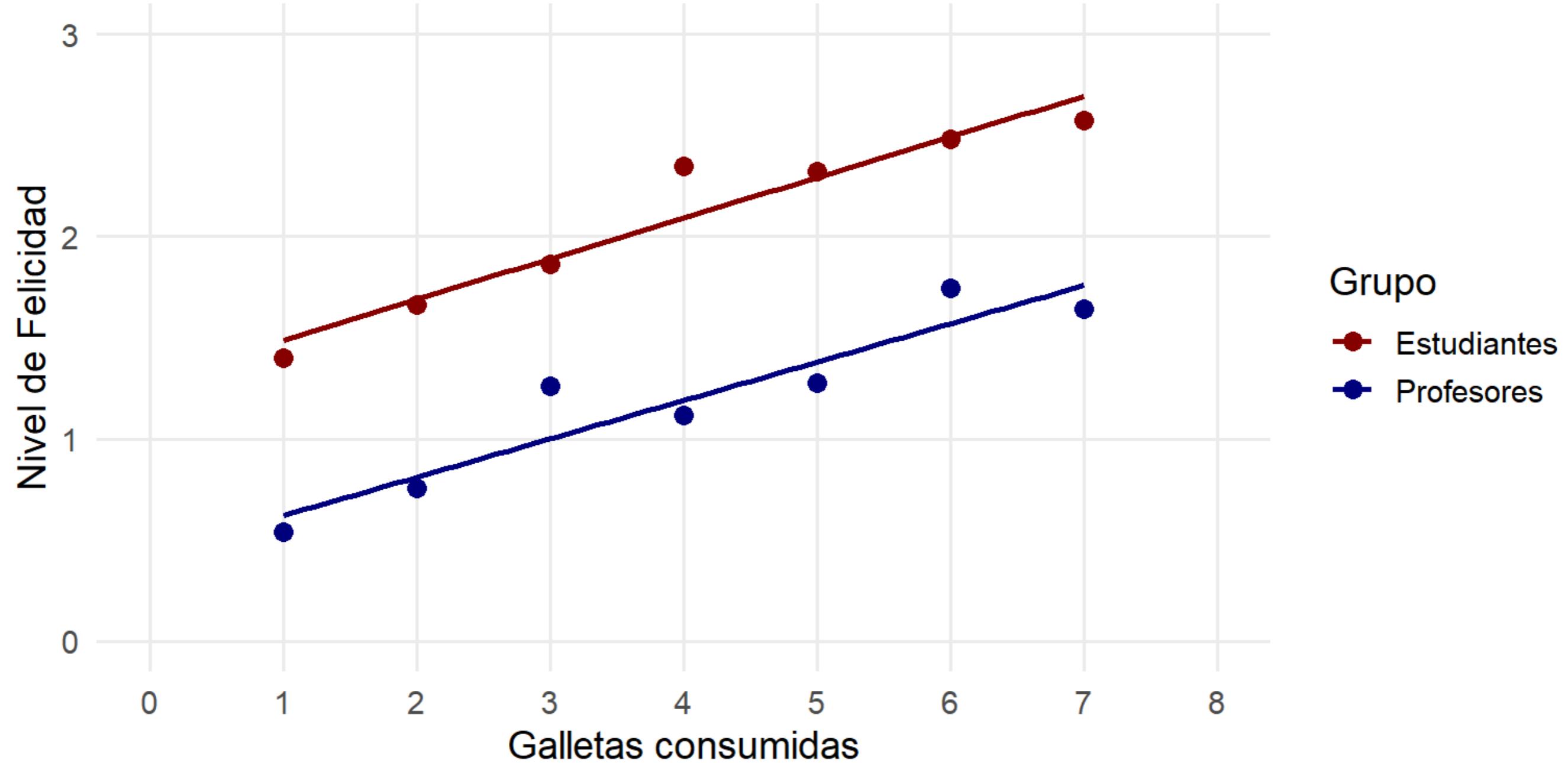
```
1  
3627921
```

Variables Categóricas vs Variables Continuas

Variables Categóricas vs Variables Continuas







Variables Categóricas

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

- El intercepto para las observaciones de los profesores será $\hat{\beta}_0$ porque $Estudiante = 0$
- El intercepto para las observaciones de los estudiantes será $\hat{\beta}_0 + \hat{\beta}_2$ porque $Estudiante = 1$

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_n en Y

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

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, \mathbf{Y} es, en promedio, β_n unidades mayor/menor para \mathbf{X}_n comparado con $\mathbf{X}_{\text{omitida}}$

$$\begin{aligned}\widehat{\text{US Gross}} = & - 8,785,254 + 2.99 \times \text{Opening Gross} + 0.356 \times \text{Budget} \\ & - 11,929,834 \times \text{Sequel}\end{aligned}$$

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

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

Si es la primera película $\text{Sequel} = \text{Trilogy} = 0$, el modelo es:

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

Variable categóricas con más de 2 niveles

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

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

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

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

¿Qué tan bueno es el modelo?

R-Squared

El R^2 es el porcentaje de la varianza de la variable dependiente explicada por el modelo de regresión

$$R^2 = \text{Corr}(x, y)^2 = \text{Corr}(y, \hat{y})$$

- Está entre 0 (nuestro modelo no predice nada) y 1 (predicción perfecta)
- No tiene unidad de medida

¿Qué tan bueno es el modelo?

Con la función `glance()` podemos ver diferentes aspectos que evalúan el modelo:

```
1 glance(hollywood_model)

# A tibble: 1 x 12
  r.squared adj.r.squared      sigma statistic  p.value    df logLik     AIC     BIC
  <dbl>        <dbl>      <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 0.790       0.782 18824686.     89.3 4.91e-24     3 -1361. 2731. 2743.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Este modelo de regresión explica el 79% de la varianza del recaudo total en US

```
1 hollywood_model <- lm(us_gross ~ opening_gross + budget + sequel, data=hollywood)
```

