Analítica de los Negocios

Más sobre Regresión Lineal

Carlos Cardona Andrade

Plan para hoy

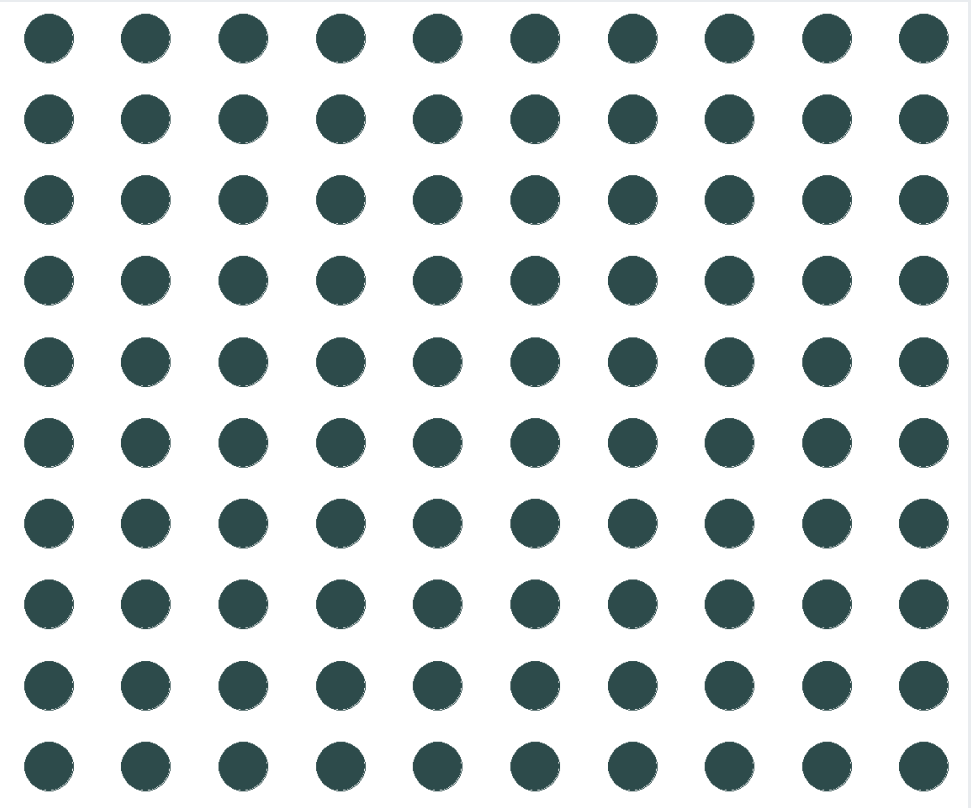
1. Población, Muestra y Regresión
2. Condiciones del modelo
3. Consideraciones para una Regresión
4. Transformación Logarítmica

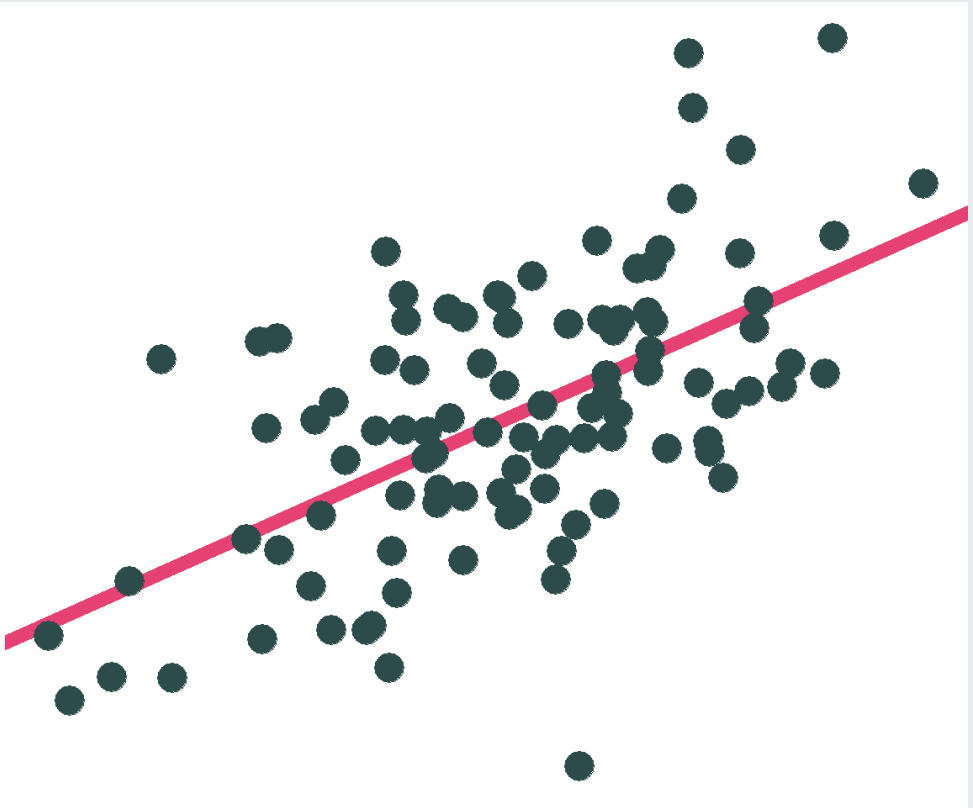
Población, Muestra y Regresión

Regresión Poblacional - Ejemplo

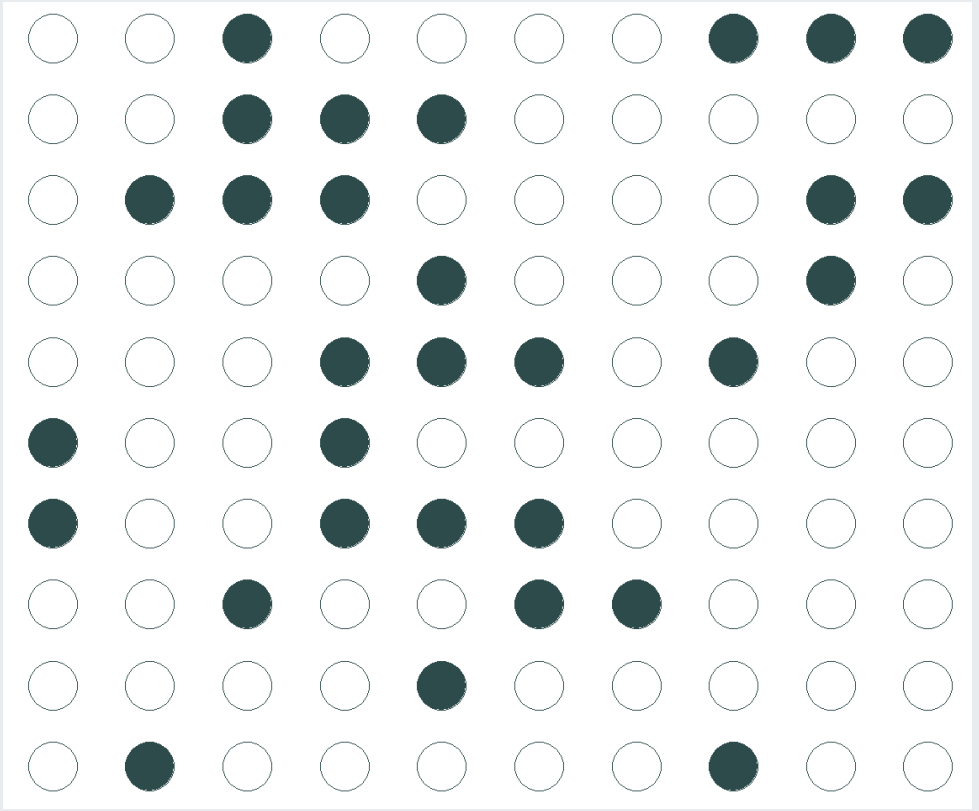
- 100 estudiantes están tomando Analítica de los Negocios y queremos saber el efecto de comer galletas en la felicidad para todos ellos. Asumamos que tenemos los datos para todos los estudiantes.
- Luego tomemos muestras de 30 estudiantes y miremos cómo los coeficientes de la siguiente regresión varían de acuerdo a la muestra:

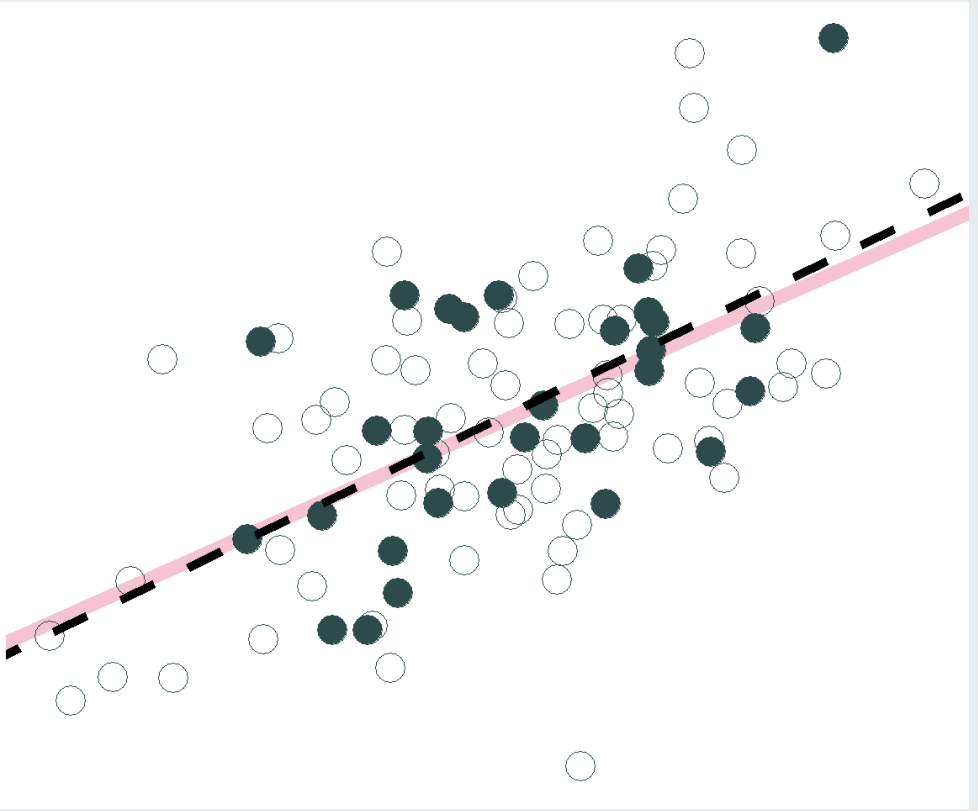
Regresión Poblacional



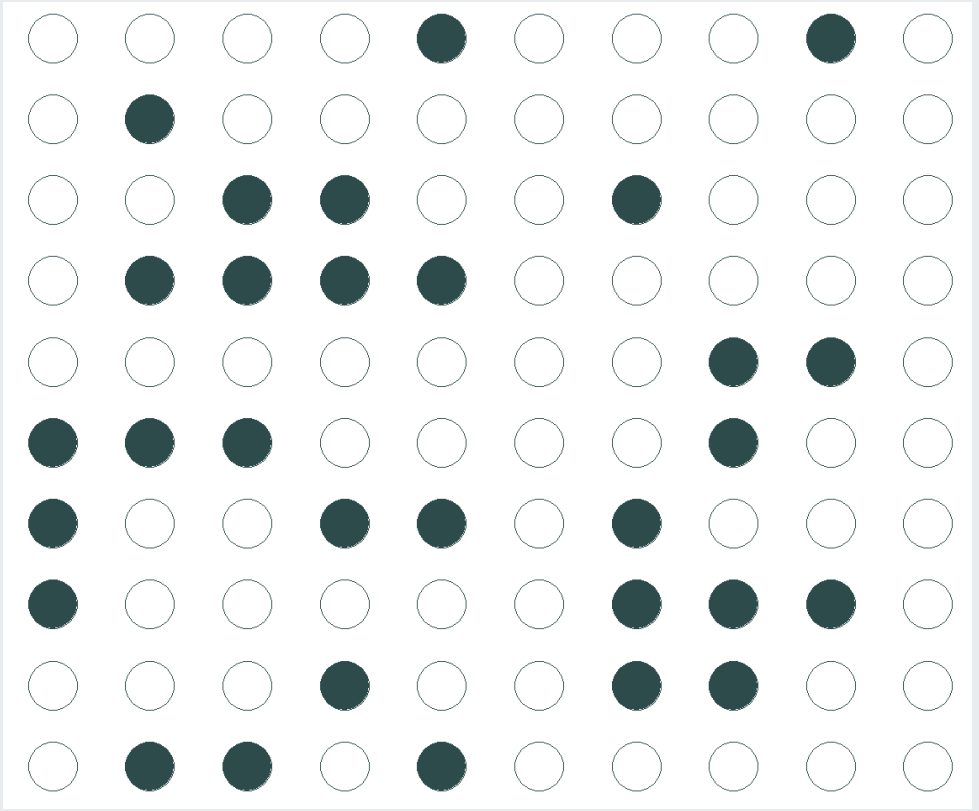


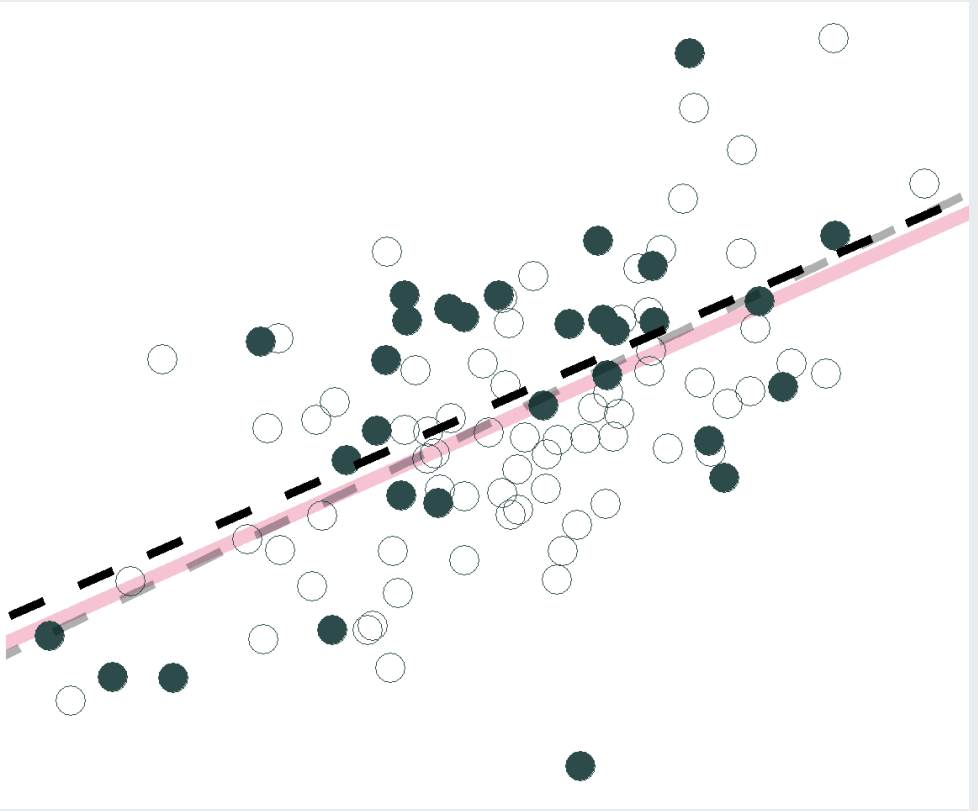
Regresión para la Muestra 1



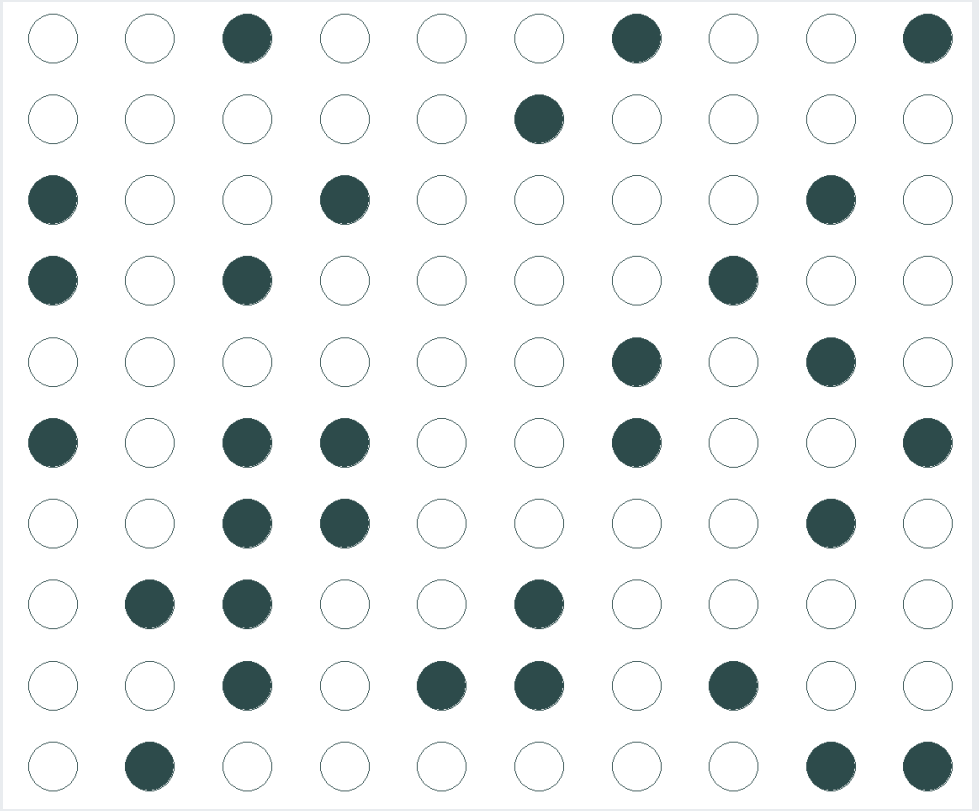


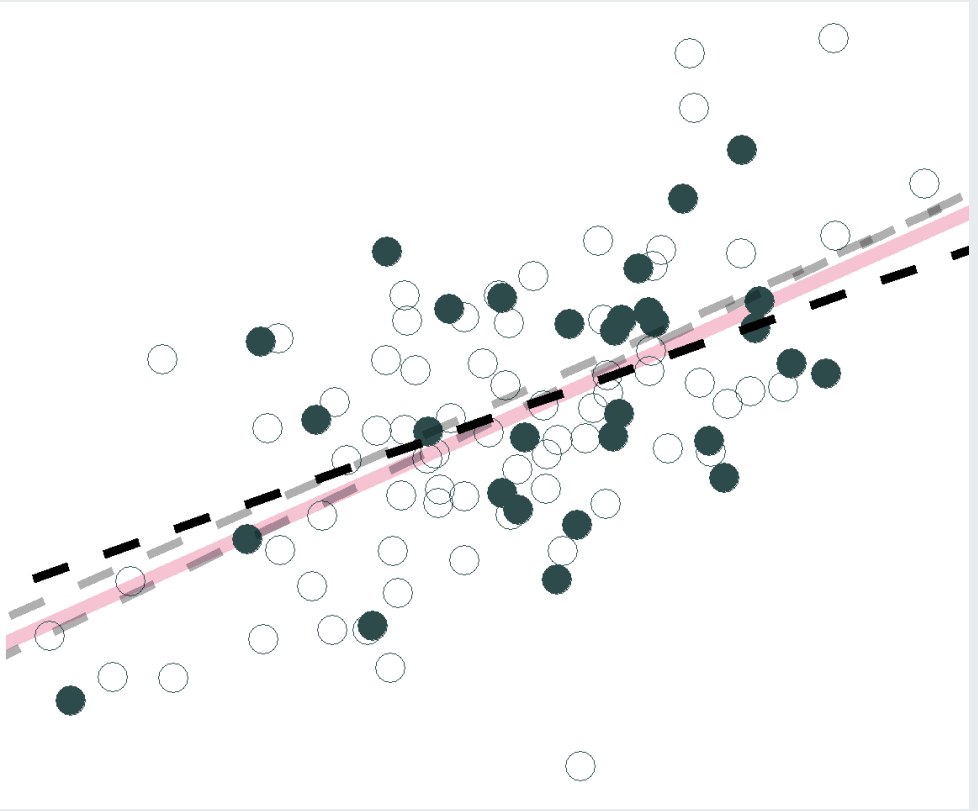
Regresión para la Muestra 2





Regresión para la Muestra 3





La incertidumbre importa!

Como no sabemos si tenemos una buena o mala muestra, y queremos:

usamos para acercarnos al verdadero efecto .

Por eso, utilizamos intervalos de confianza y pruebas de hipótesis para cuantificar la incertidumbre en nuestras estimaciones

Condiciones del modelo

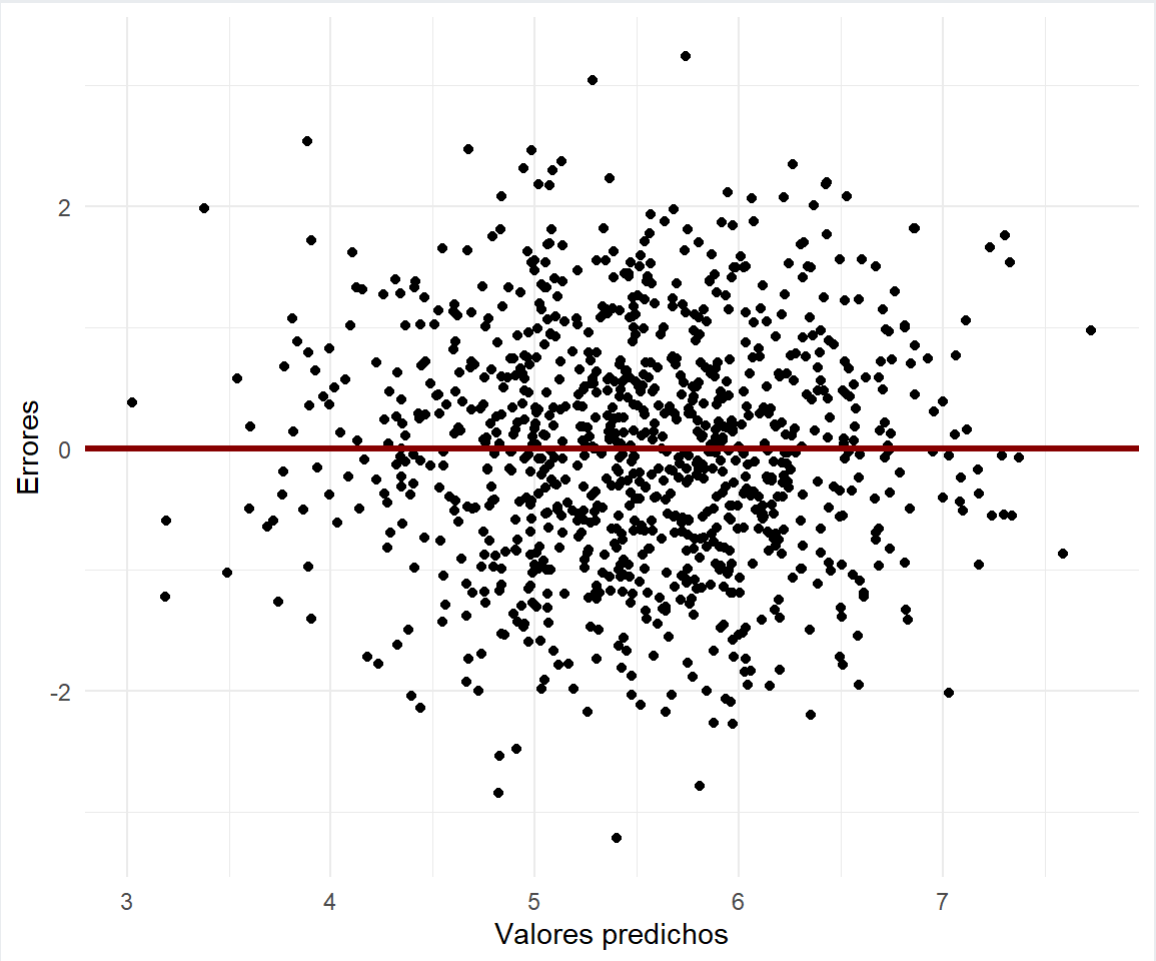
Condiciones del modelo

1. Linealidad: hay una relación lineal entre la variable dependiente y la variable explicativa
2. Varianza Constante: la variabilidad de los errores es igual para todos los valores de la variable explicativa
3. Normalidad: los errores siguen una distribución normal
4. Independencia: los errores son independientes entre ellos

¿Cómo graficar la variación de los errores en R?

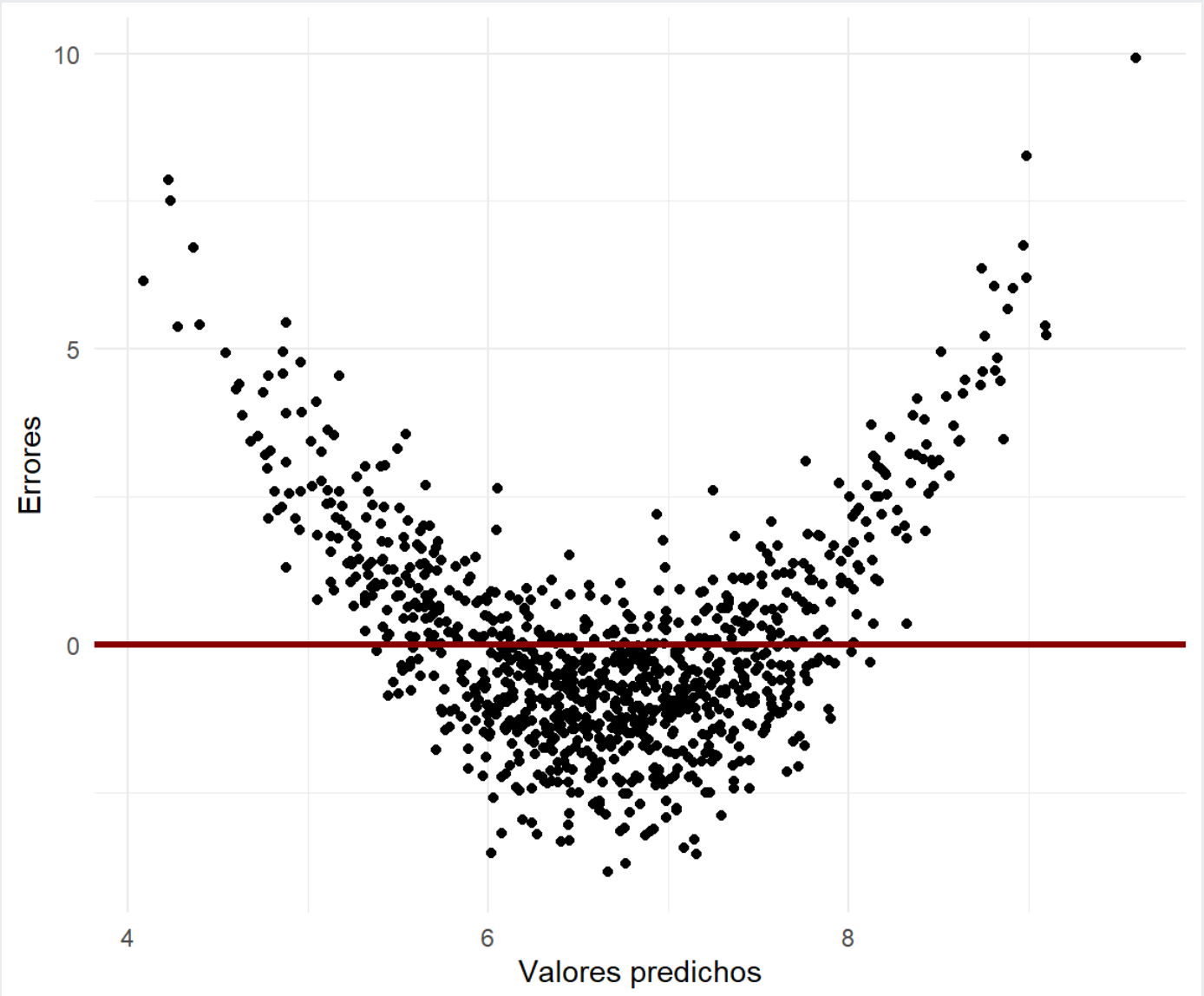
```r
ols_model <- lm(y ~ x, data=DATA)

augment(ols_model) |>
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  labs(x = "Valores predichos", y = "Errores") +
  theme_minimal()
```

¿Son lineales los Errores?

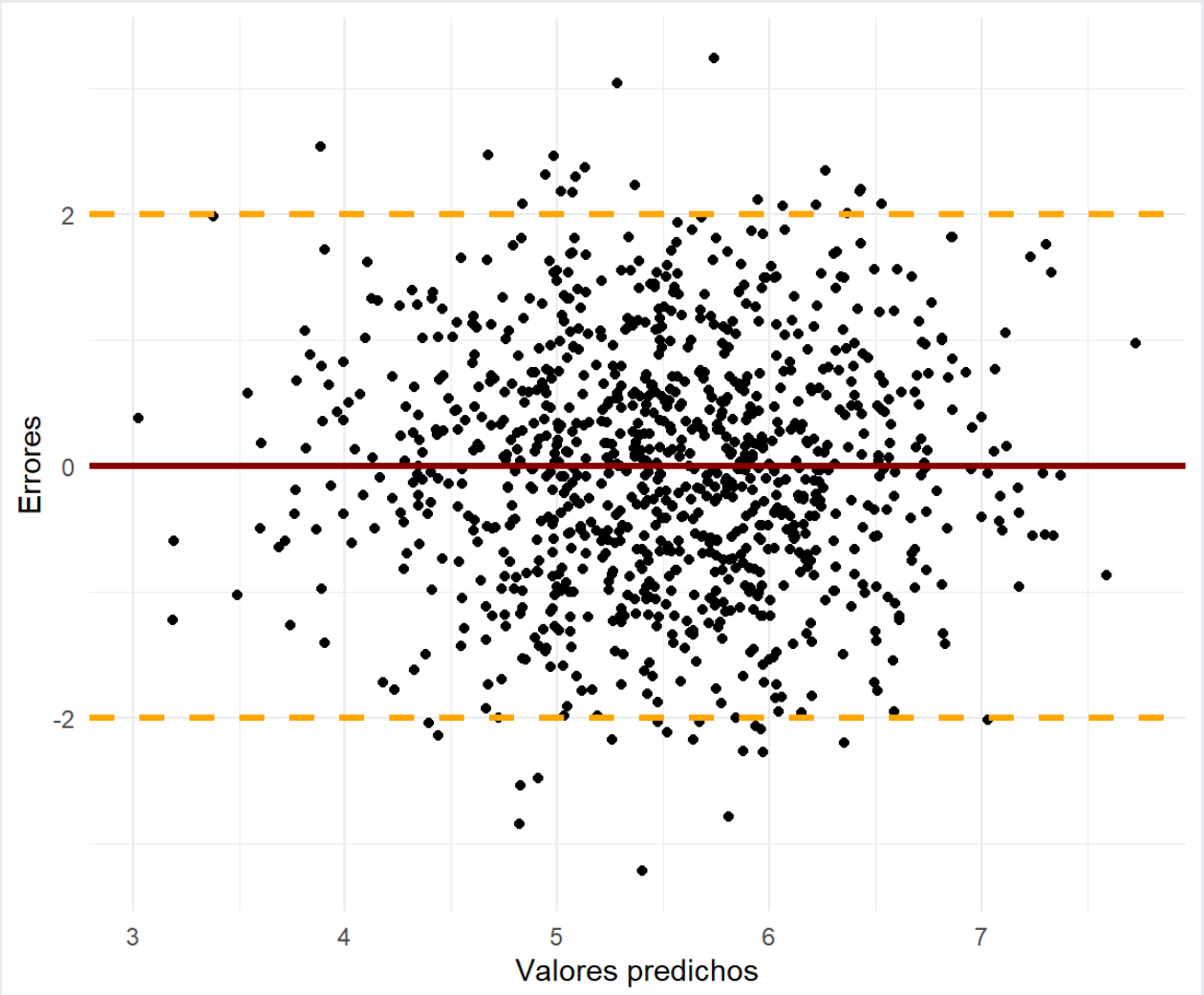


✅Los errores no siguen un patrón o estructura clara. Parecen aleatoriamente distribuidos

✗ Claro patrón en los Errores

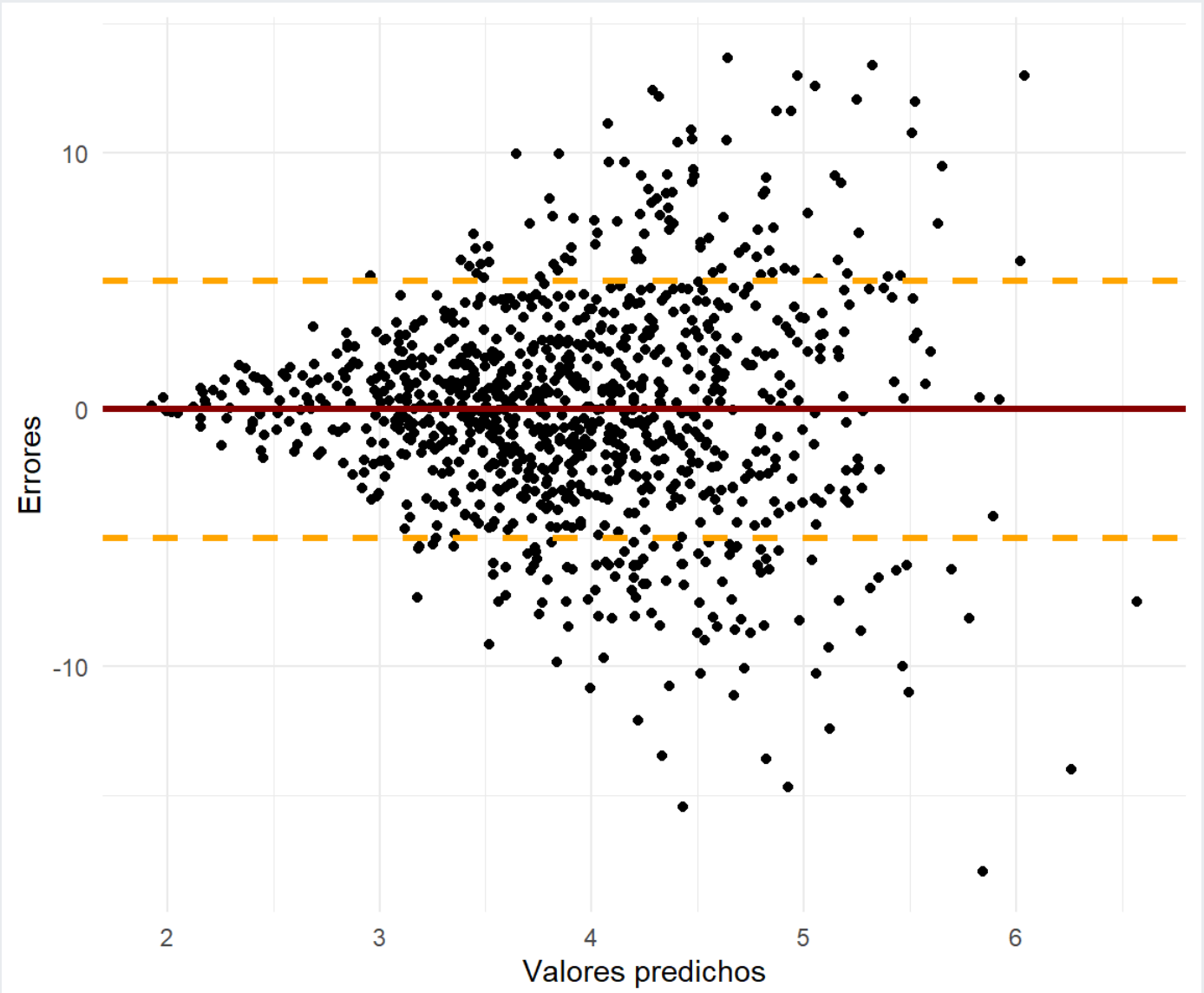


¿La varianza de los errores es constante?

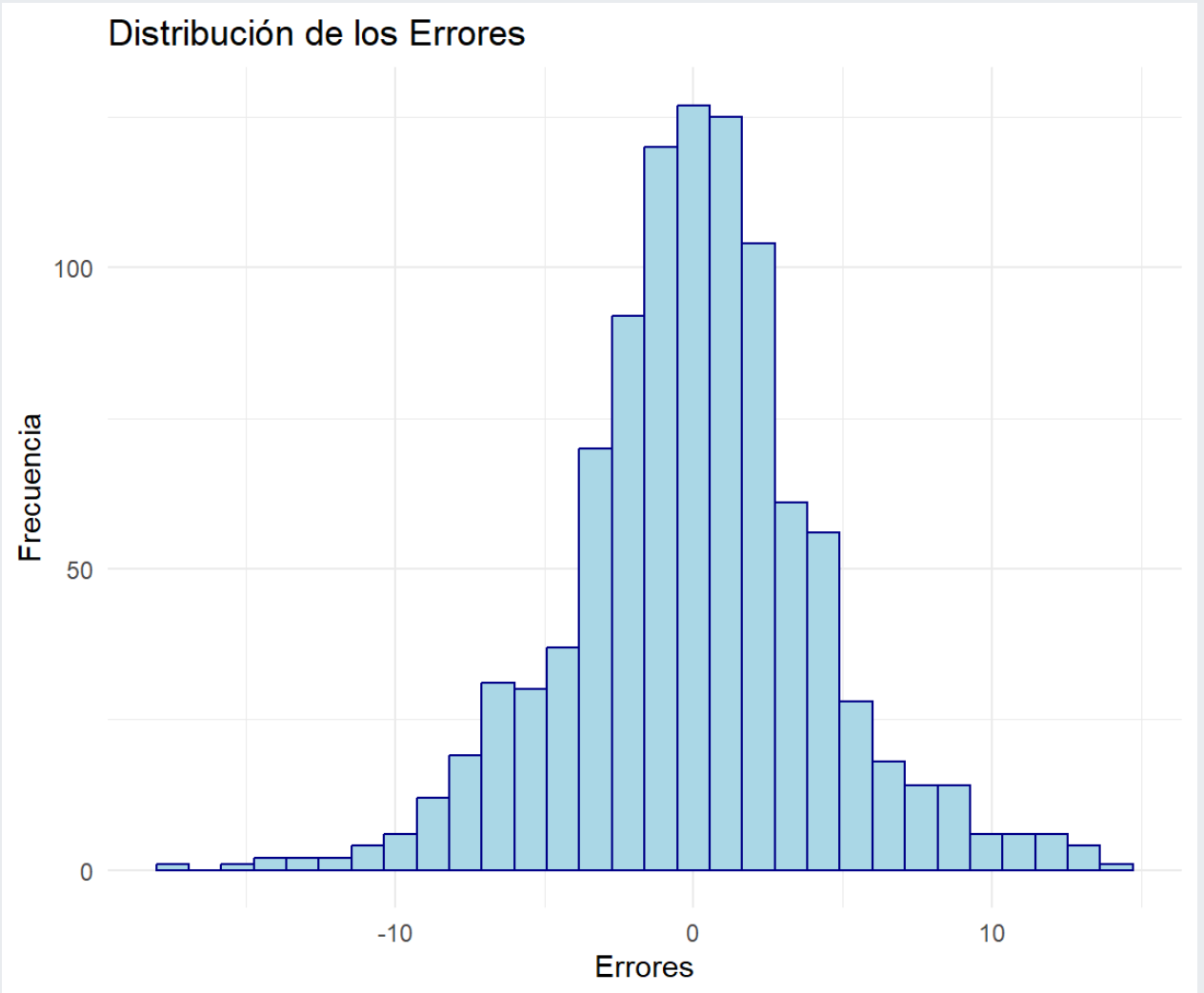


✅ La dispersión vertical de los errores es relativamente constante en la gráfica

La varianza no es constante

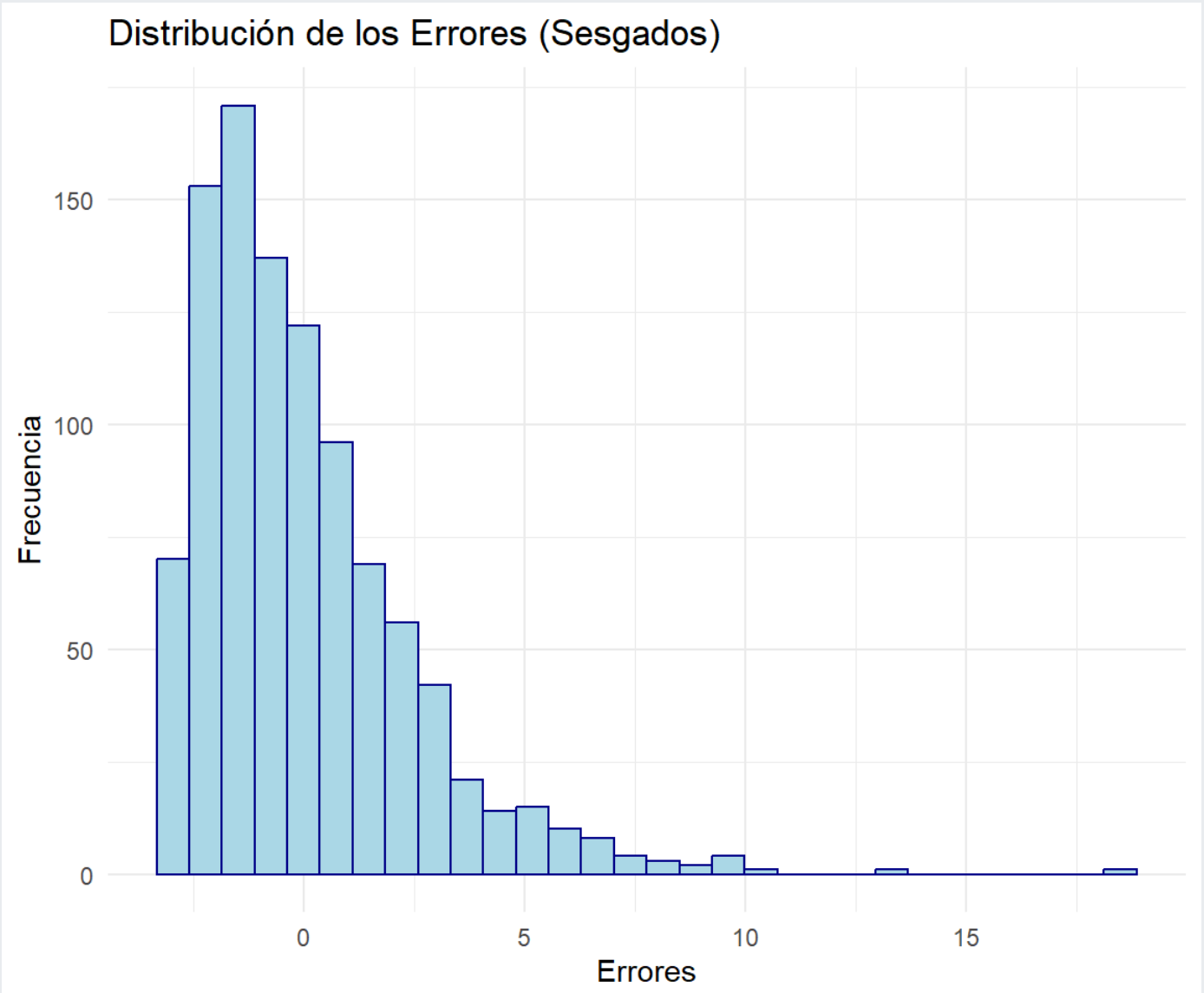


¿Los errores se distribuyen normalmente?



✅ La distribución de los errores se parece a una distribución normal

✗ Los errores no se distribuyen como una normal



¿Cómo graficar la distribución de los errores en R?

```r
augment(ols_model) |>
  ggplot(aes(x = .resid)) +
  geom_histogram(fill = "lightblue", color = "darkblue", bins = 30) +
  labs(title = "Distribución de los Errores",
       x = "Errores",
       y = "Frecuencia") +
  theme_minimal()
```

Independencia

- Podemos verificar el supuesto de independencia a menudo basándonos en el contexto de los datos y en cómo se recolectaron las observaciones.
- Si los datos se recolectaron en un orden particular, examina un diagrama de dispersión de los errores versus el orden en que se recolectaron los datos.

En la práctica..

Al verificar las condiciones del modelo, preguntense si alguna desviación de estas condiciones es tan grande que:

1. Se deba proponer un modelo diferente.
2. Las conclusiones extraídas del modelo deban usarse con precaución.

Si no es así, las condiciones se cumplen suficientemente y podemos proceder con el modelo actual.

¿Qué tan bueno es el modelo?

El R cuadrado

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

- 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 hollywood_model <- lm(us_gross ~ opening_gross + budget + sequel, data=holl
2 glance(hollywood_model)
```

```
# 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     0.790         0.782 18824686.      89.3 4.91e-24     3 -1361. 2731.
2743.
# ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

según la primera columna, este modelo de regresión:

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

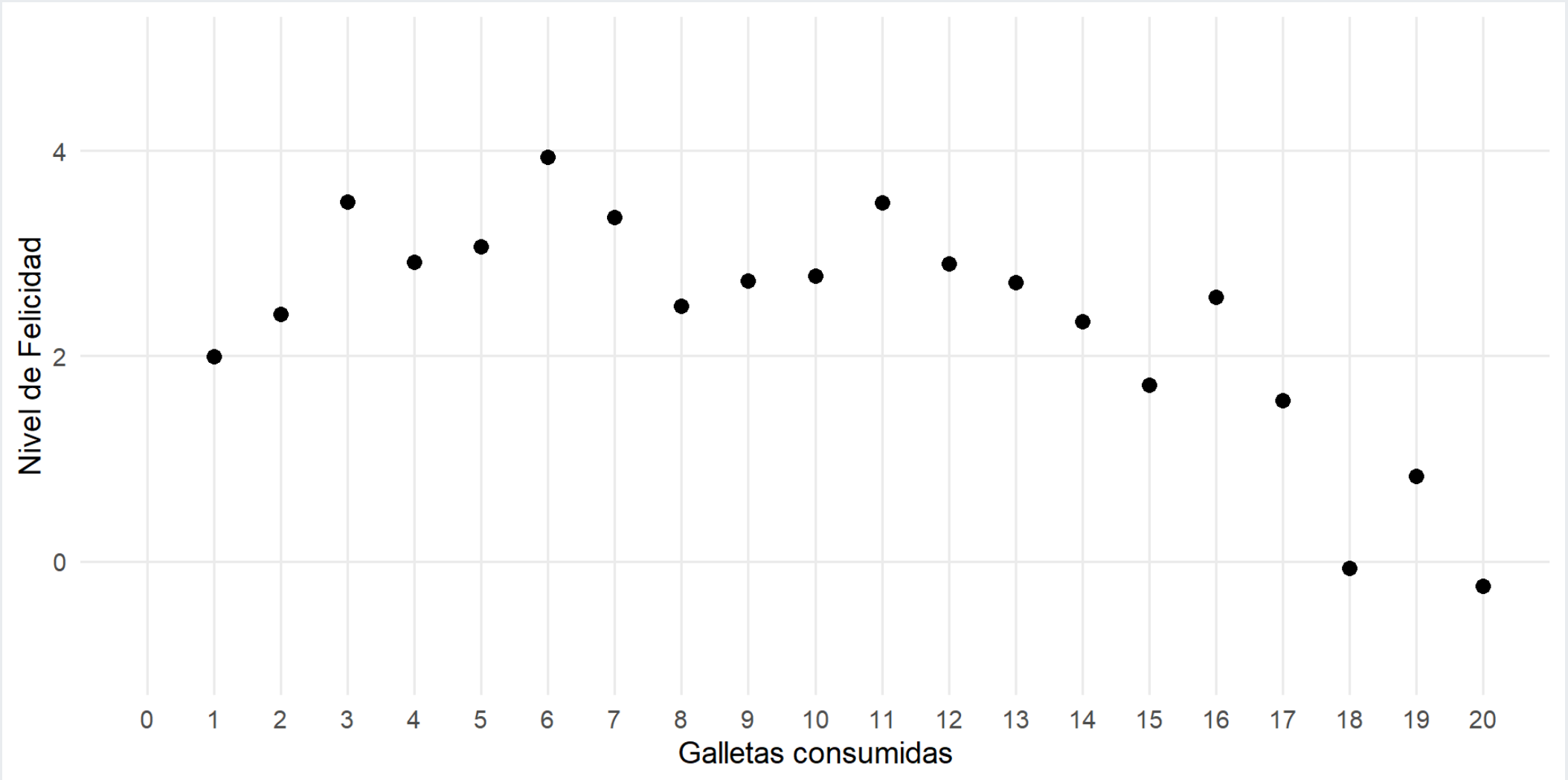
💪 Ejercicio 1

1. Usando los datos `hollywood_data.xlsx`, evalúen si los errores del siquiente modelo de regresión son lineales, su varianza es constante y si se distribuyen normalmente:

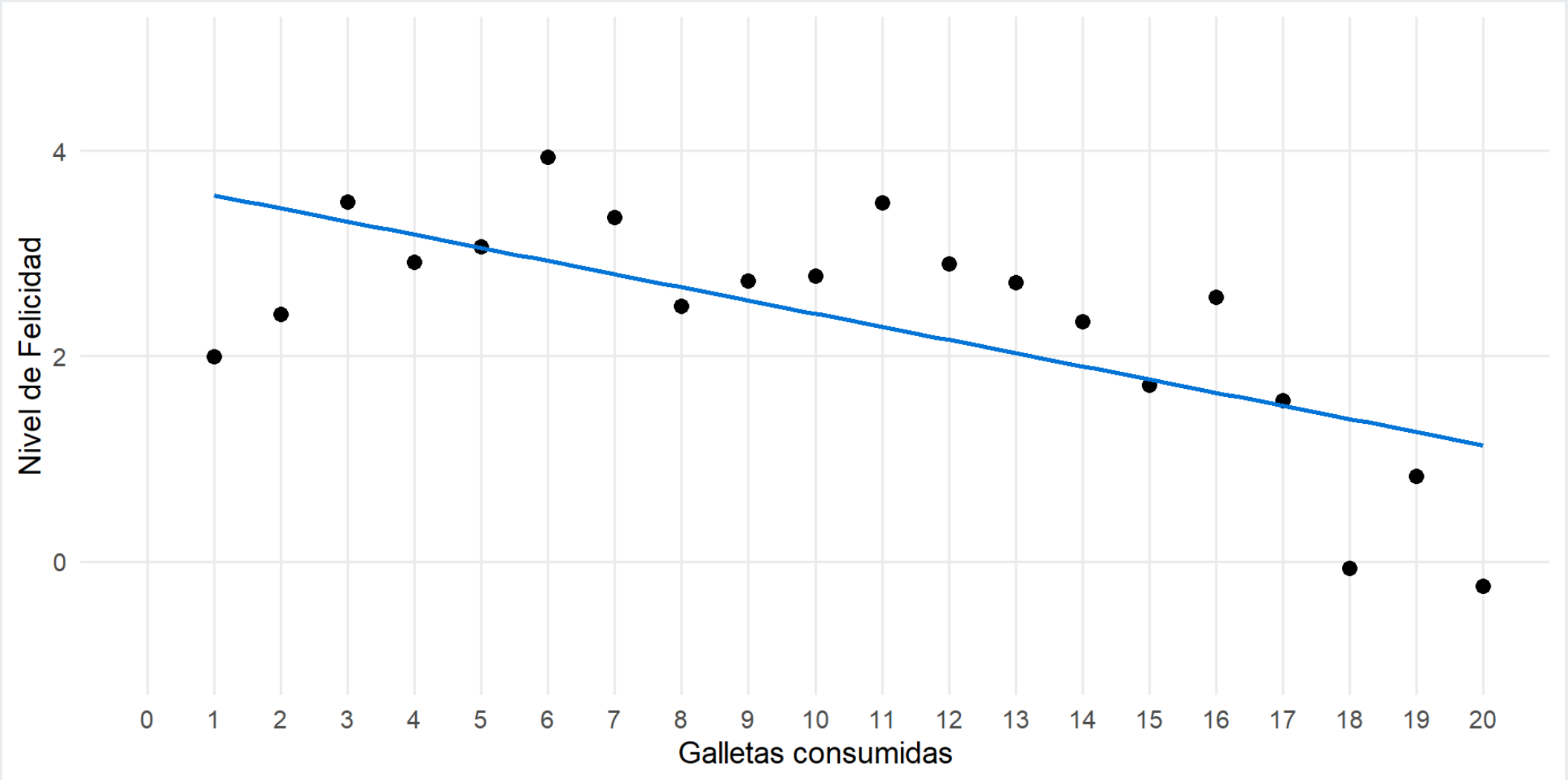
2. ¿Cuál es el del modelo en el punto 1?
3. Comparen el del punto 2 con el de la diapositiva anterior. ¿Cuál explica más la varianza del recaudo total en US?¿Por qué creen que es así?

Consideraciones adicionales para una Regresión

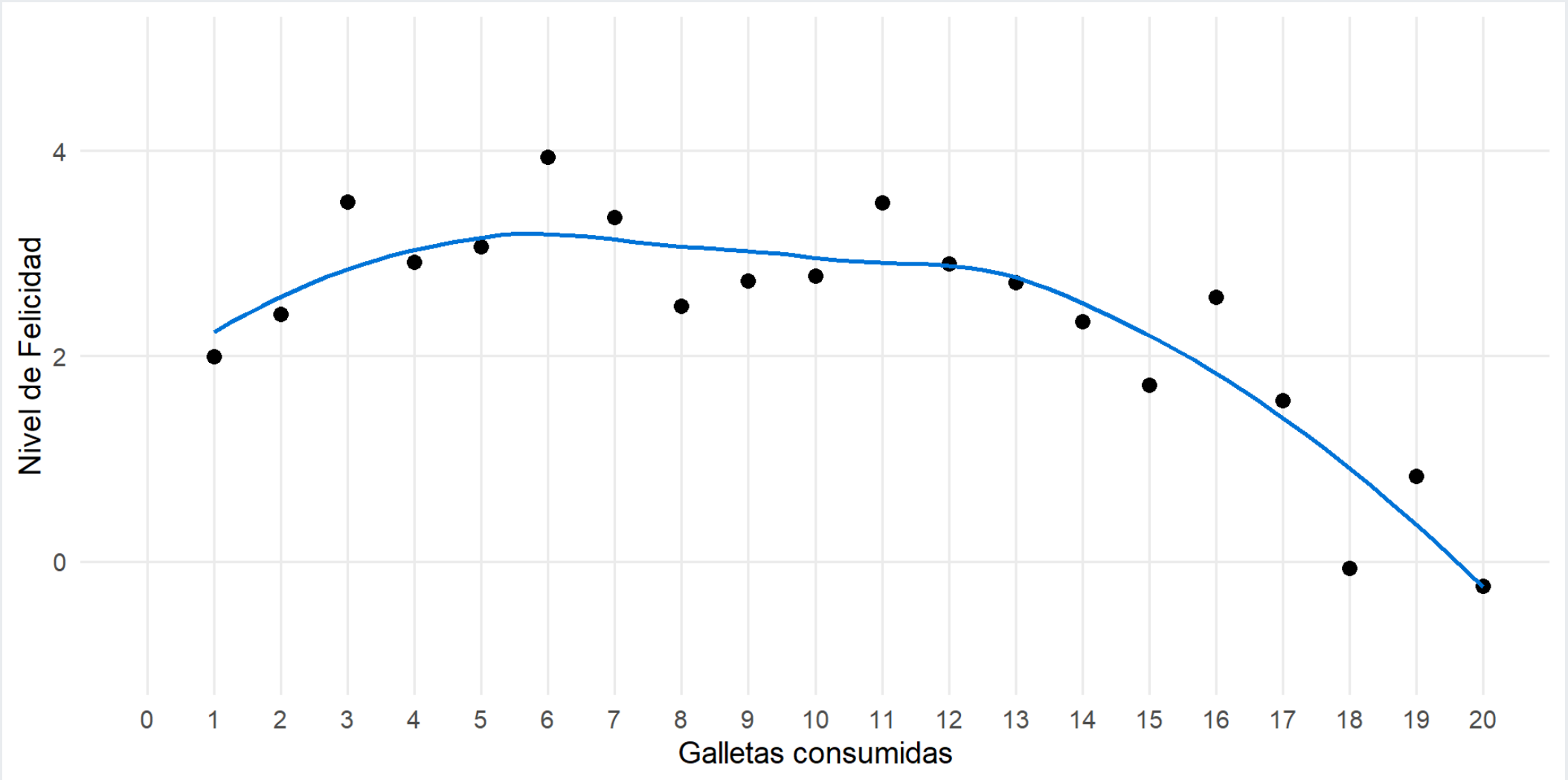
Relaciones no lineales



Relaciones no lineales



Relaciones no lineales



Relaciones no lineales en R

Para agregar el término cuadrático a la regresión, se incluye el término `I(x^2)` en el código:

```
modelo_felicidad <- lm(felicidad ~ galletas + I(galletas^2), data = galleta
tidy(modelo_felicidad, conf.int = TRUE)
```

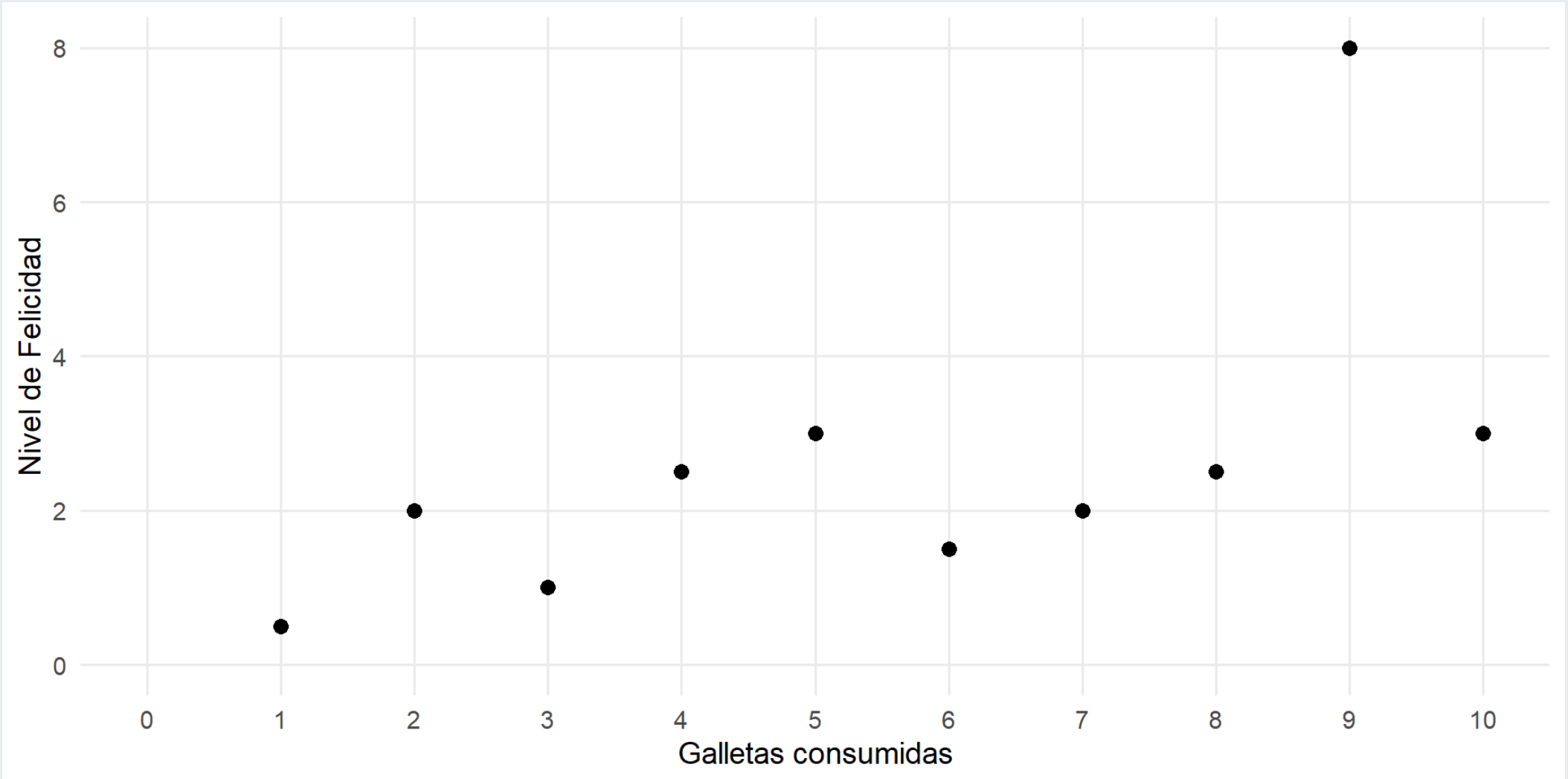
```
# A tibble: 3 × 7
  term           estimate std.error statistic   p.value conf.low conf.high
  <chr>             <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)      2.00     0.378         5.30 0.0000588   1.21      2.80
2 galletas         0.334    0.0828        4.03 0.000871    0.159     0.509
3 I(galletas^2)   -0.0220   0.00383      -5.74 0.0000241  -0.0301   -0.0139
```

Nuestro modelo estimado entonces es:

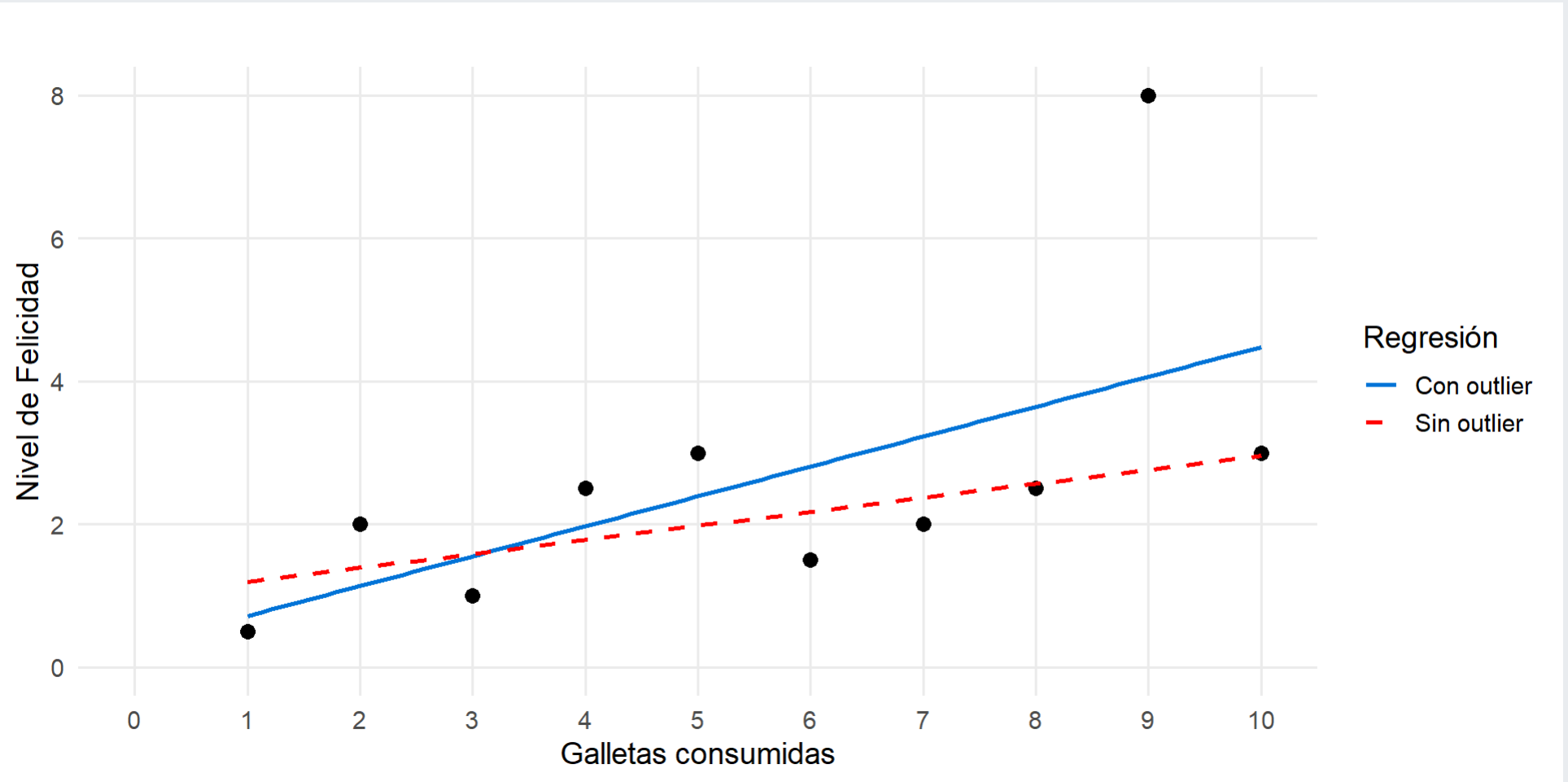
¿Existe la relación cuadrática que vemos en la gráfica?

Evaluamos con una prueba de hipótesis:

¿Qué hacer con una observación atípica (outlier)?



¿Qué hacer con una observación atípica (outlier)?



Como se menciono antes, es importante analizar la causa de los valores atípicos antes de eliminarlos, ya que pueden contener información valiosa.

Efecto interacción

Hasta ahora hemos asumido que el efecto de una variable es independiente de otra:

El efecto del incremento en una unidad de en , es siempre e independiente de

Por ejemplo:

Efecto interacción

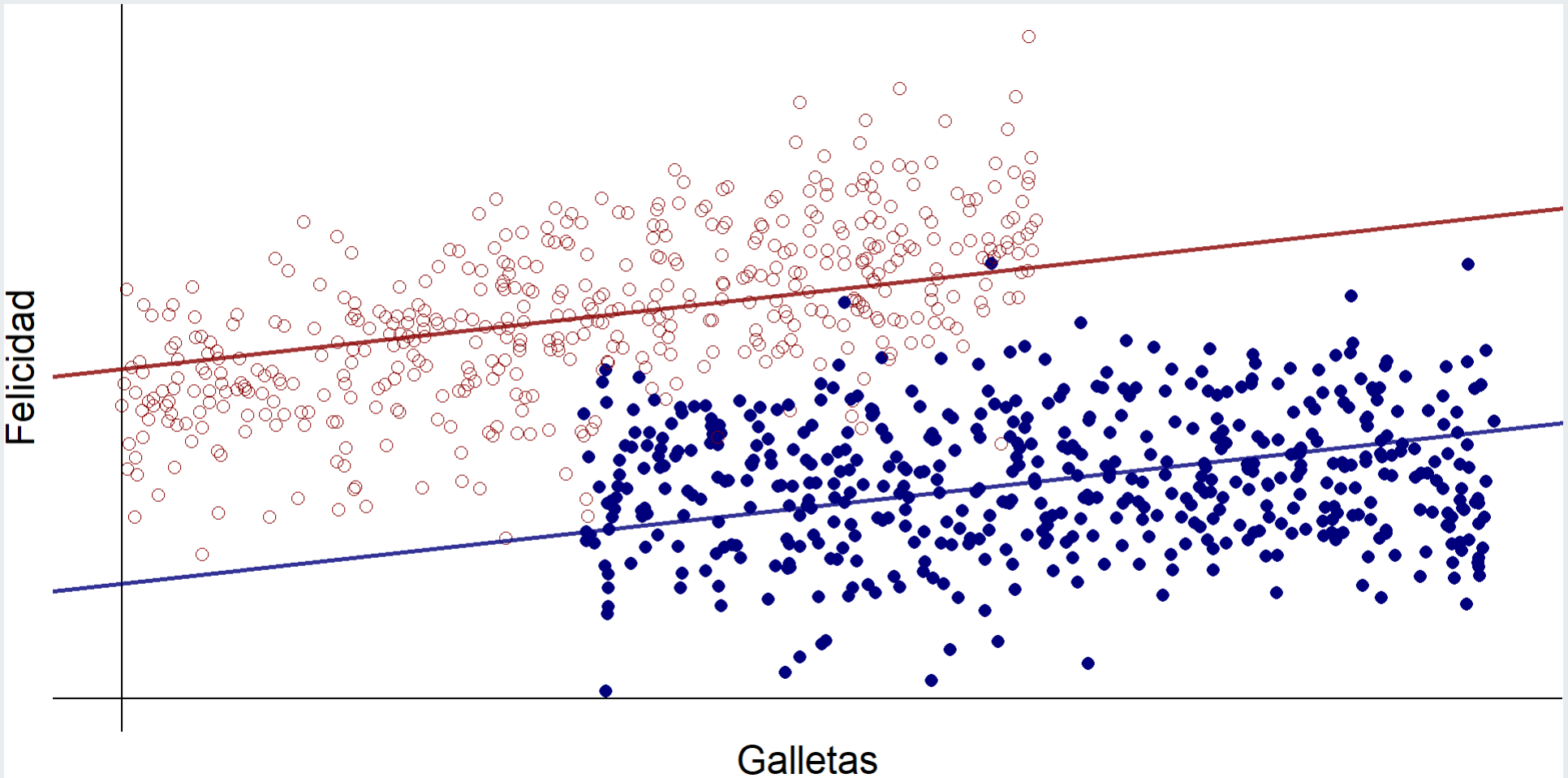
Si pensamos que el efecto de depende del valor de , entonces debemos agregar una tercera variable de interacción al modelo:

El termino de interacción es

Por ejemplo:

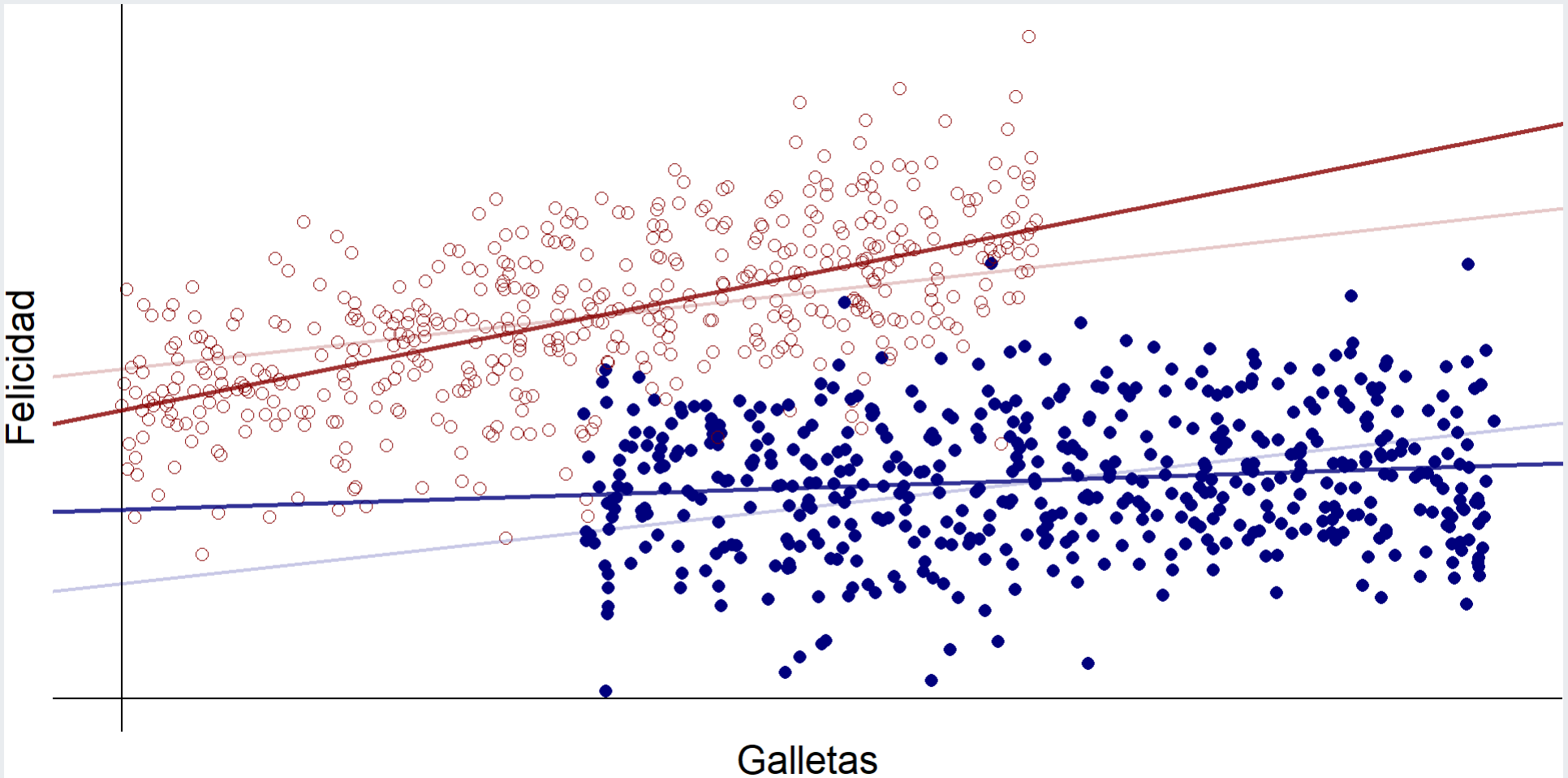
Interacciones - Análisis Visual

El modelo donde las galletas tienen el mismo efecto para profesores y estudiantes



Interacciones - Análisis Visual

El modelo donde el efecto de las galletas puede diferir para profesores y estudiantes



💪 Ejercicio 2

1. Usando los datos `hollywood_data.xlsx`, creen una variable `comedy` que sea igual a 1 si el género de la película es “Comedy” y 0 en caso contrario.
2. Algunos productores de Hollywood creen que los retornos al recaudo el día del estreno son más altos para las películas de comedia. Comprueben esto Estimando la siguiente regresión:

¿Tienen razón los productores de Hollywood? (Pista: debe ser positivo y estadísticamente diferente de 0)

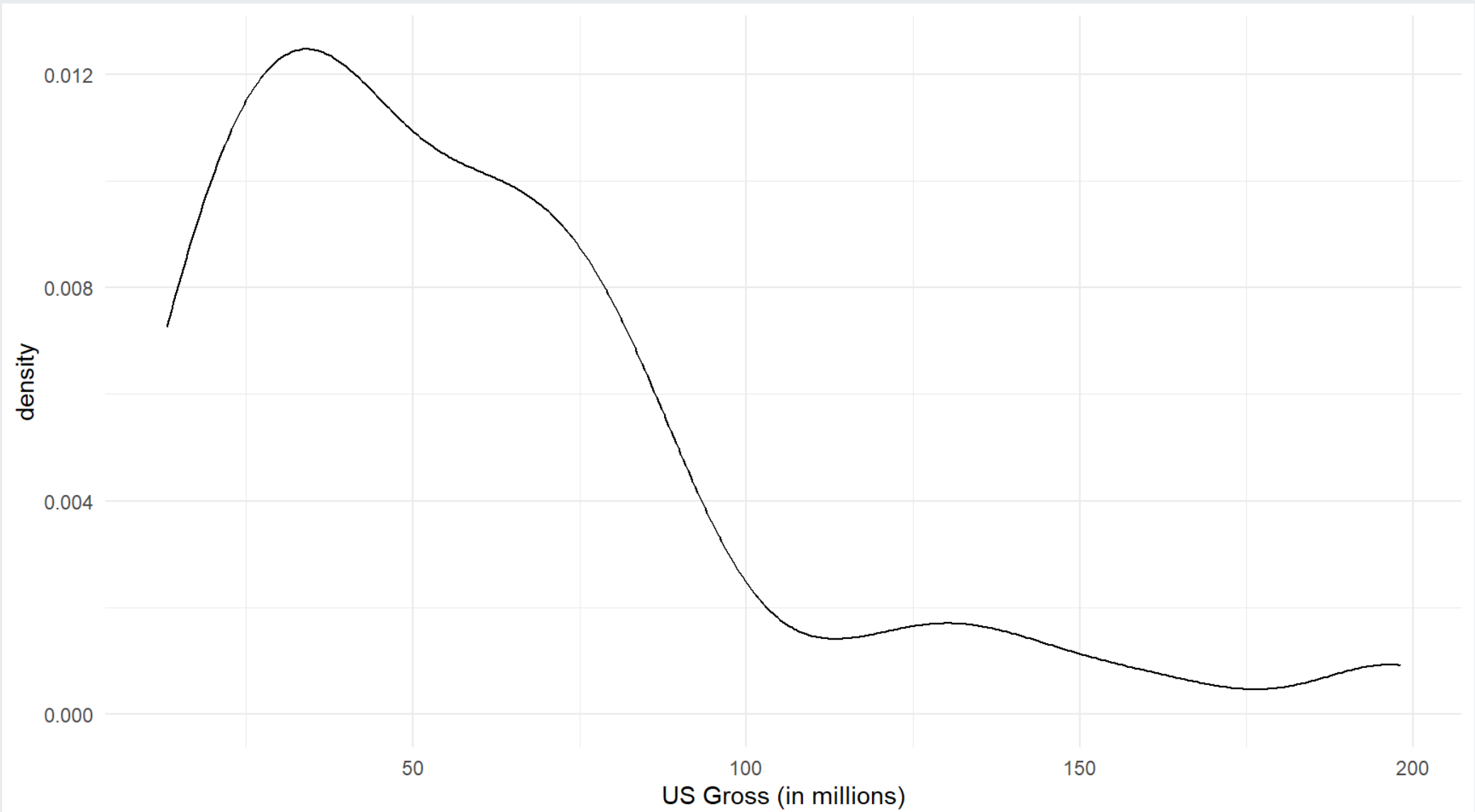
Transformación Logarítmica

Transformación logarítmica

Podemos considerar una transformación logarítmica de las variables de nuestro modelo (tanto variable dependiente e independientes) cuando:

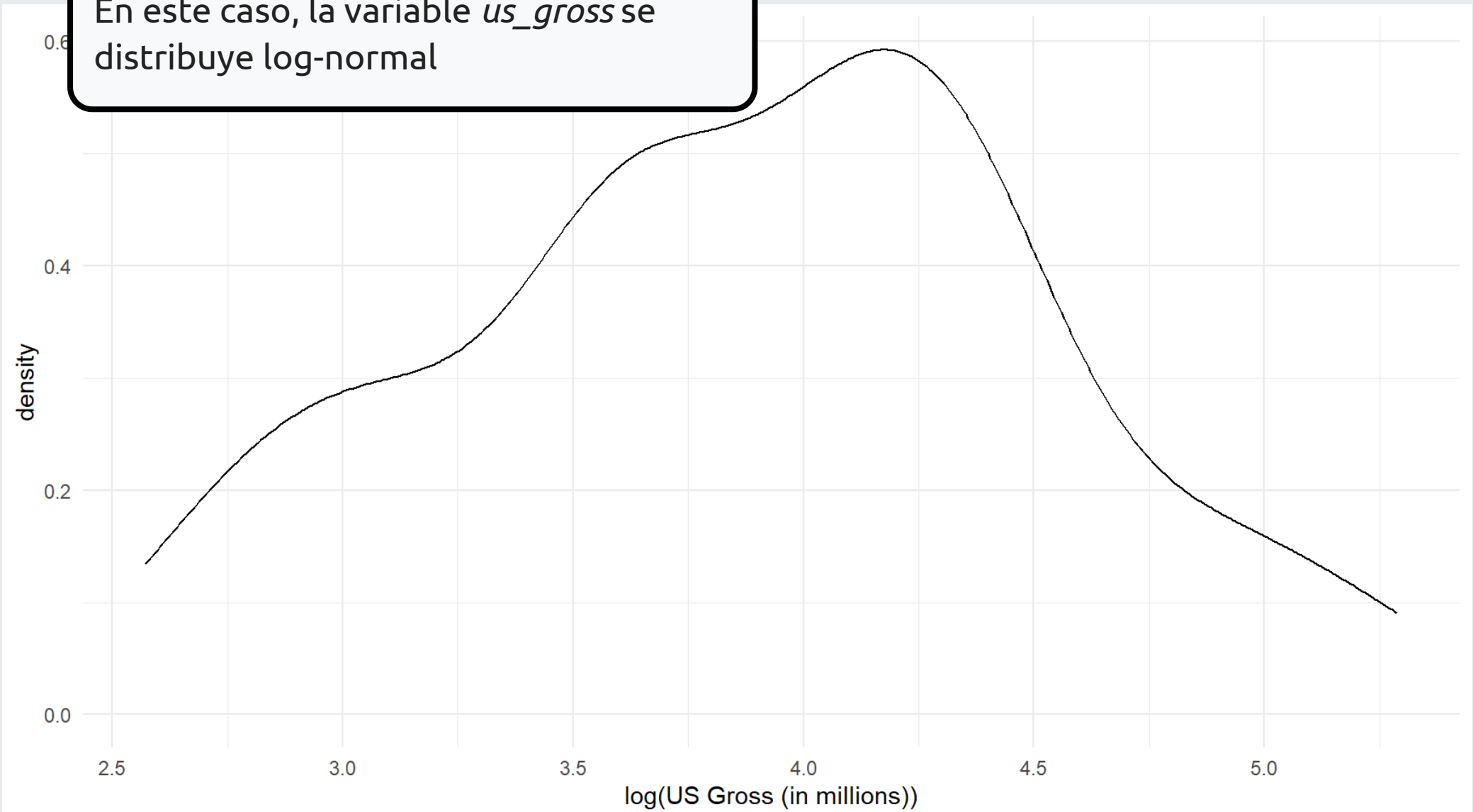
- Exista una relación no-lineal (exponencial) entre la variable dependiente y la explicativa.
- Alguna de las variables tenga una distribución sesgada (muy distinta a la normal).

¿Es la distribución del recaudo normal?



¿Es la distribución del logaritmo del recaudo normal?

En este caso, la variable *us_gross* se distribuye log-normal



Transformación logarítmica

```
  1  hollywood_model <- lm(us_gross ~ opening_gross, data=hollywood)
  2  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
```

La regresión lineal estimada es:

Transformación logarítmica

¿Cómo cambian los resultados si estimamos una regresión con ambias variables como logarítmos?

```
hollywood <- hollywood |>
  mutate(log_opening_gross=log(opening_gross),
         log_us_gross=log(us_gross))

hollywood_model_logs <- lm(log_us_gross ~ log_opening_gross, data=hollywood
tidy(hollywood_model_logs, 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)          1.58     1.09        1.45 1.51e- 1   -0.590      3.74
2 log_opening_gross    0.977    0.0658     14.8  9.60e-24    0.845      1.11
```

La regresión lineal estimada es:

Regresión Nivel-Nivel

Interpretación: un cambio de una unidad en , está asociado a un cambio en unidades en .

Regresión Nivel-Log

Interpretación: un aumento de 1% en , es asociado a un cambio en unidades a .

Regresión Log-Nivel

Interpretación: un incremento de una unidad en , está asociado a un cambio de en .

Regresión Log-Log

Interpretación: un aumento de 1% en , está asociado a un cambio de en .