

Analítica de Datos

Más aspectos sobre Regresión Lineal

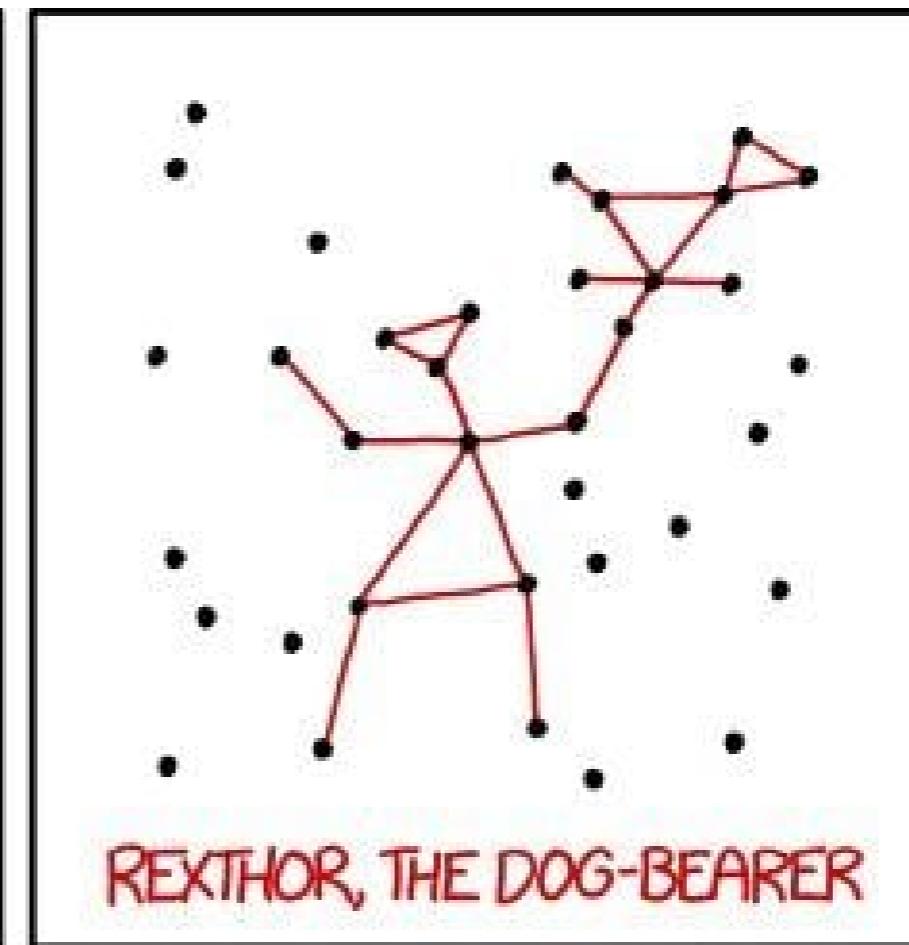
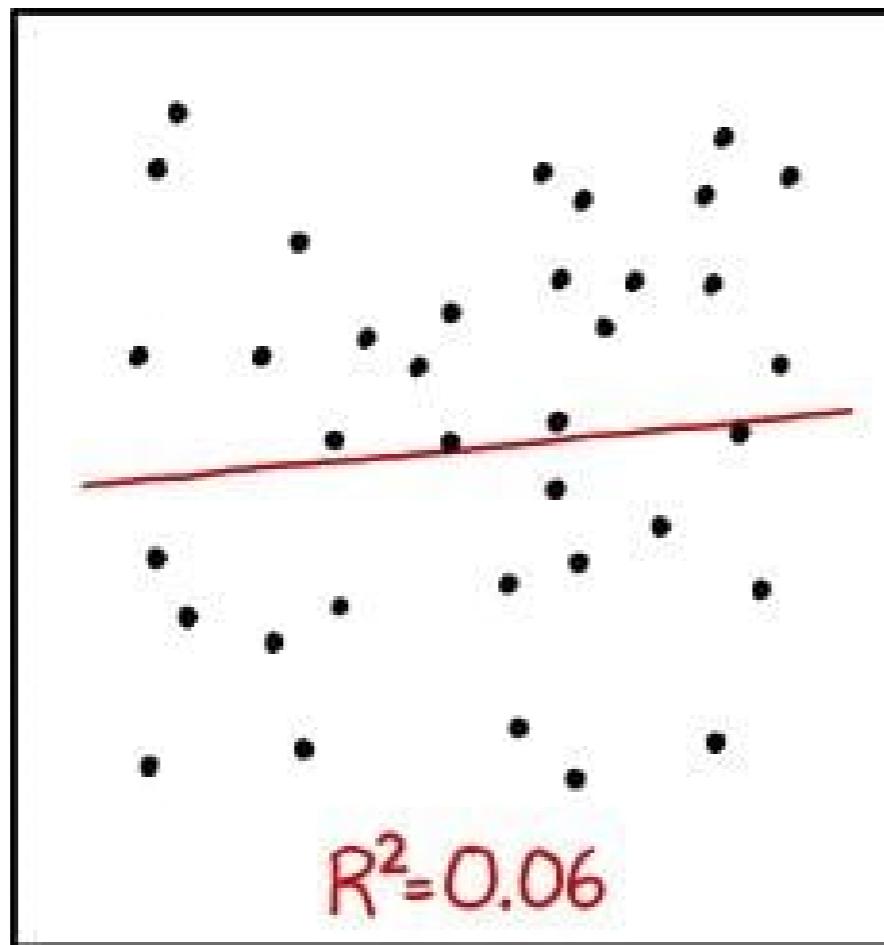
Carlos Cardona Andrade

Tabla de contenido

1. Condiciones del modelo
2. Consideraciones para una Regresión
3. Transformación Logarítmica

Condiciones del modelo

Diagnóstico general de una regresión lineal



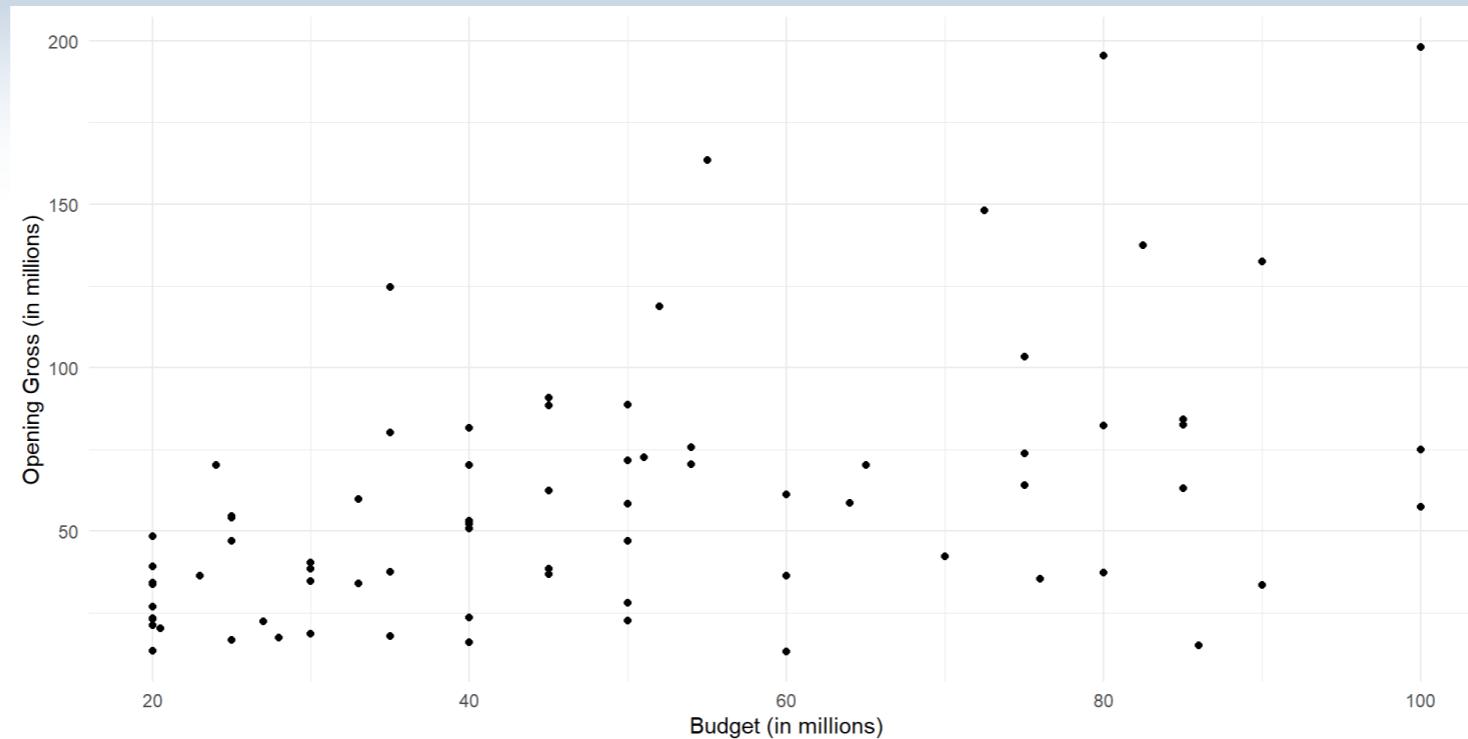
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

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

US Gross vs Budget

```
1 ggplot(data = hollywood, aes(x = budget, y = us_gross)) +  
2   geom_point() +  
3   labs(  
4     x = "Budget (in millions)",  
5     y = "Opening Gross (in millions)"  
6   ) +  
7   theme_minimal()
```



US Gross vs Budget

$$\widehat{\text{US Total Gross}} = 18 + 0.84 \times \text{Budget}$$

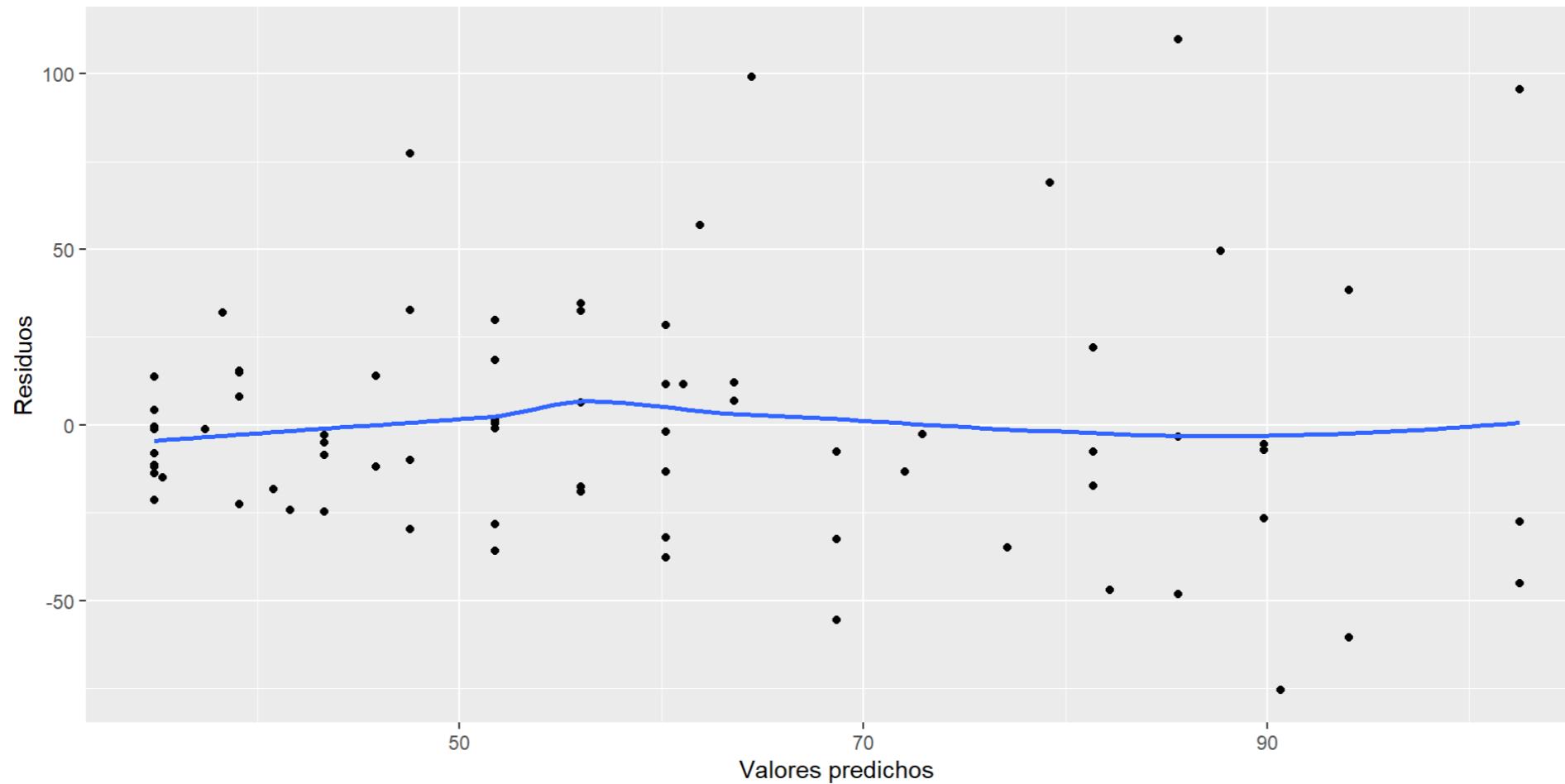
```
1 hollywood_model <- lm(us_gross ~ budget, data=hollywood)
2 tidy(hollywood_model, conf.int = TRUE)

# A tibble: 2 × 7
  term      estimate std.error statistic   p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) 18.0      9.44     1.90    0.0609    -0.843    36.8
2 budget       0.845     0.173     4.89  0.00000590     0.500     1.19
```

Error vs Residuo

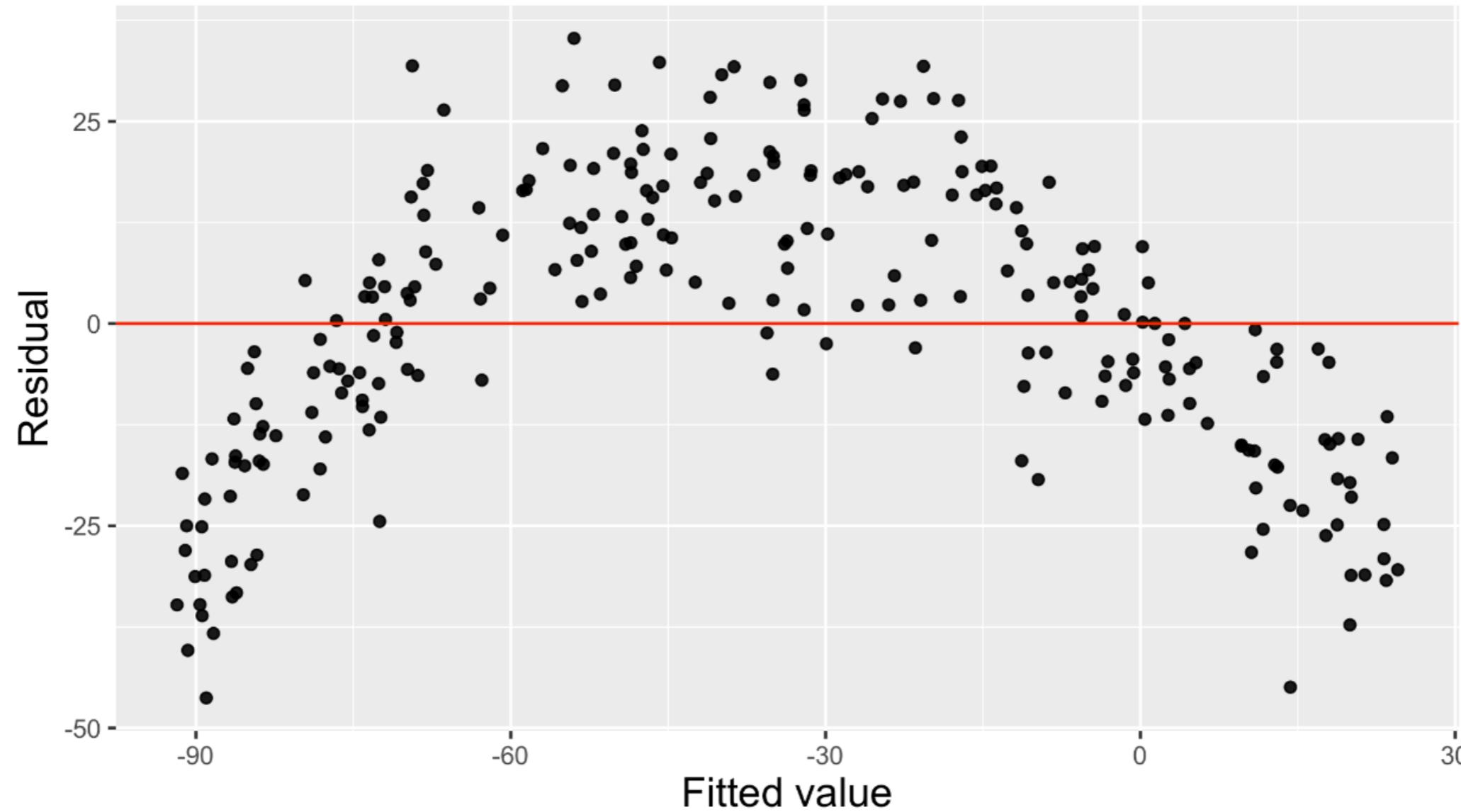
NUNCA vamos a conocer pero podemos usar los residuos como un estimado de los errores.

¿Son lineales los residuos?



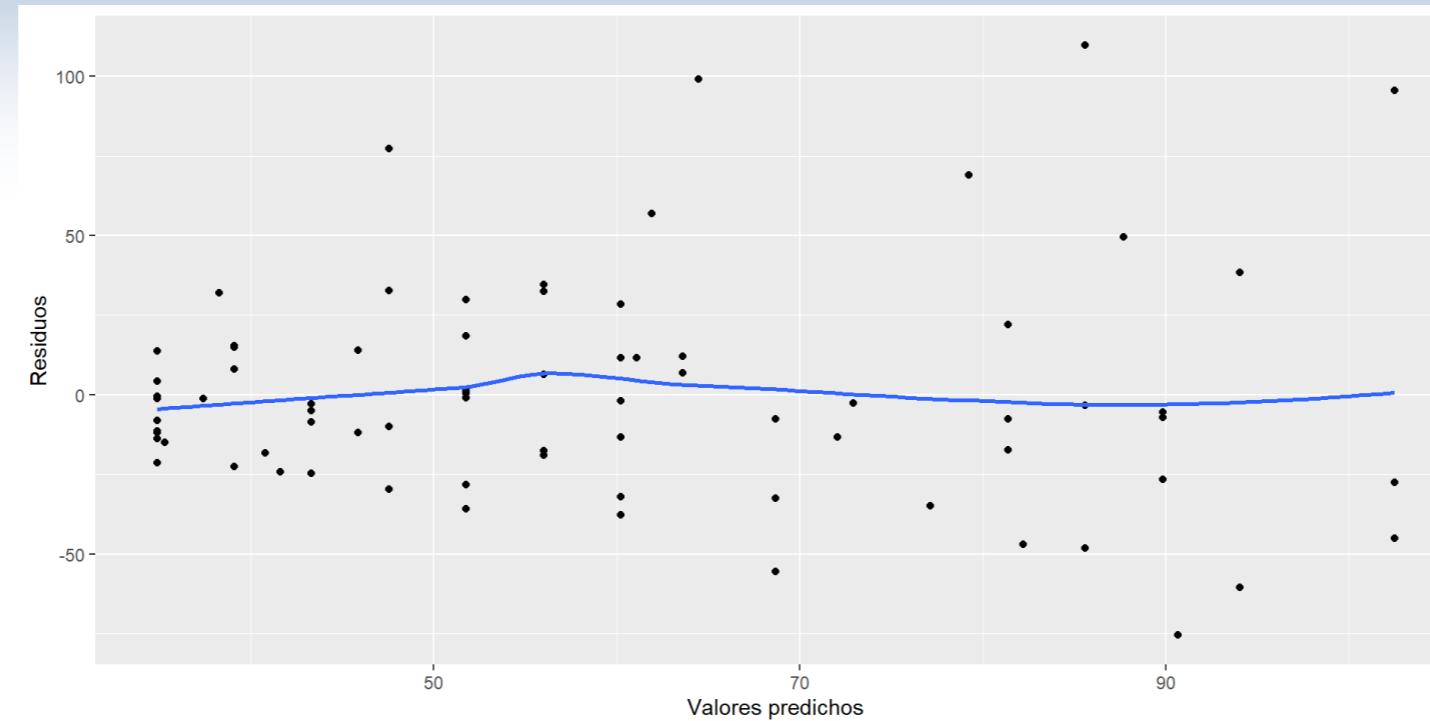
Los residuos no siguen un patrón o estructura clara. Parecen aleatoriamente distribuidos

X Claro patrón en los residuos

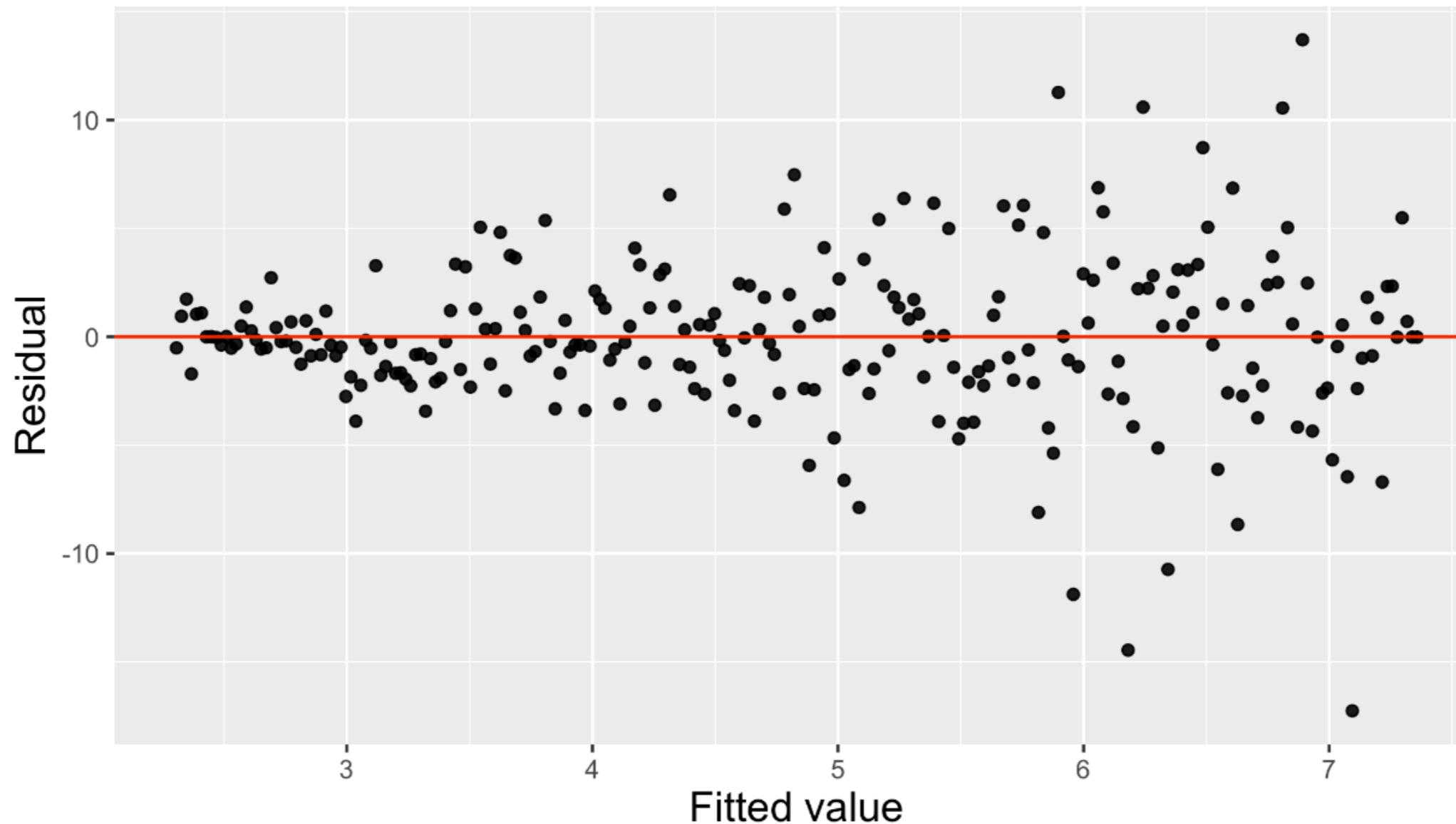


Evaluando varianza constante

```
1 augment(hollywood_model) %>%
2   ggplot(aes(x = .fitted, y = .resid)) +
3   geom_point() +
4   geom_smooth(se = FALSE) +
5   labs(x = "Valores predichos", y = "Residuos")
```

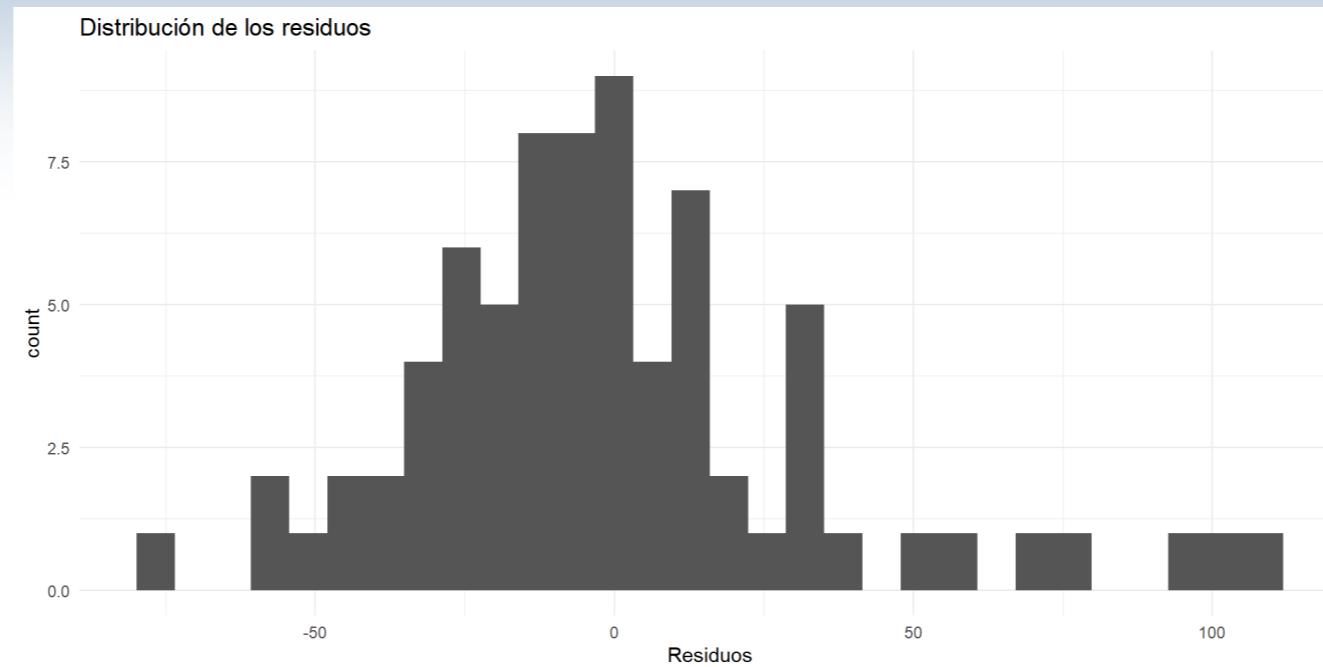


X La varianza no es constante



Condiciones del modelo

```
1 augment(hollywood_model) %>%
2   ggplot(aes(x = .resid)) +
3   geom_histogram() +
4   labs(title = "Distribución de los residuos",
5        x = "Residuos") +
6   theme_minimal()
```



- ✓ La distribución de los errores se parece a una distribución normal

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 residuos versus el orden en que se recolectaron los datos.
-  Basado en la información disponible, el error de una película no nos dice nada sobre el error de otra película.

En la práctica..

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

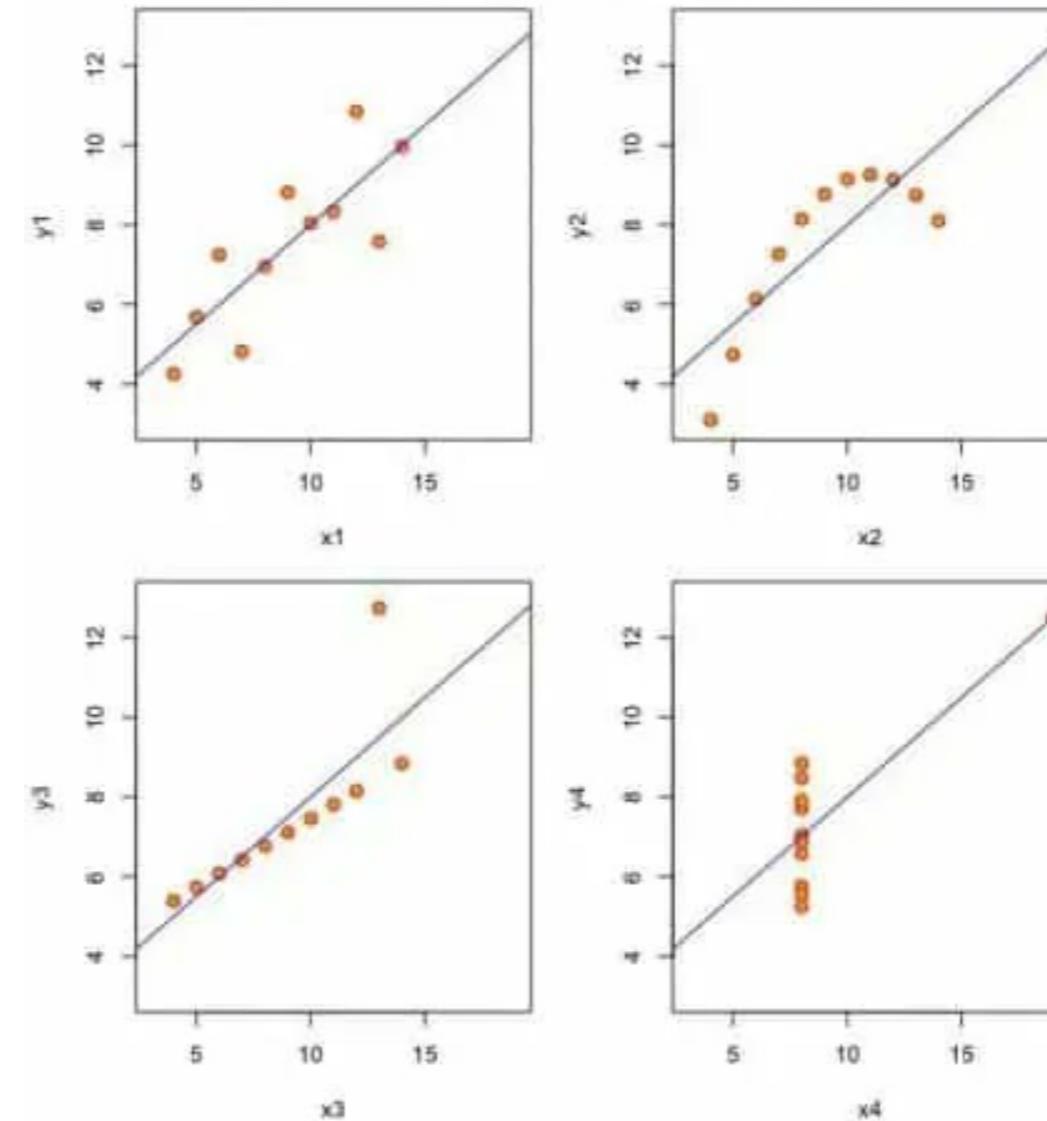
1. Se deba proponer un modelo diferente.
2. Las conclusiones extraídas del modelo deban usarse con precaución.

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

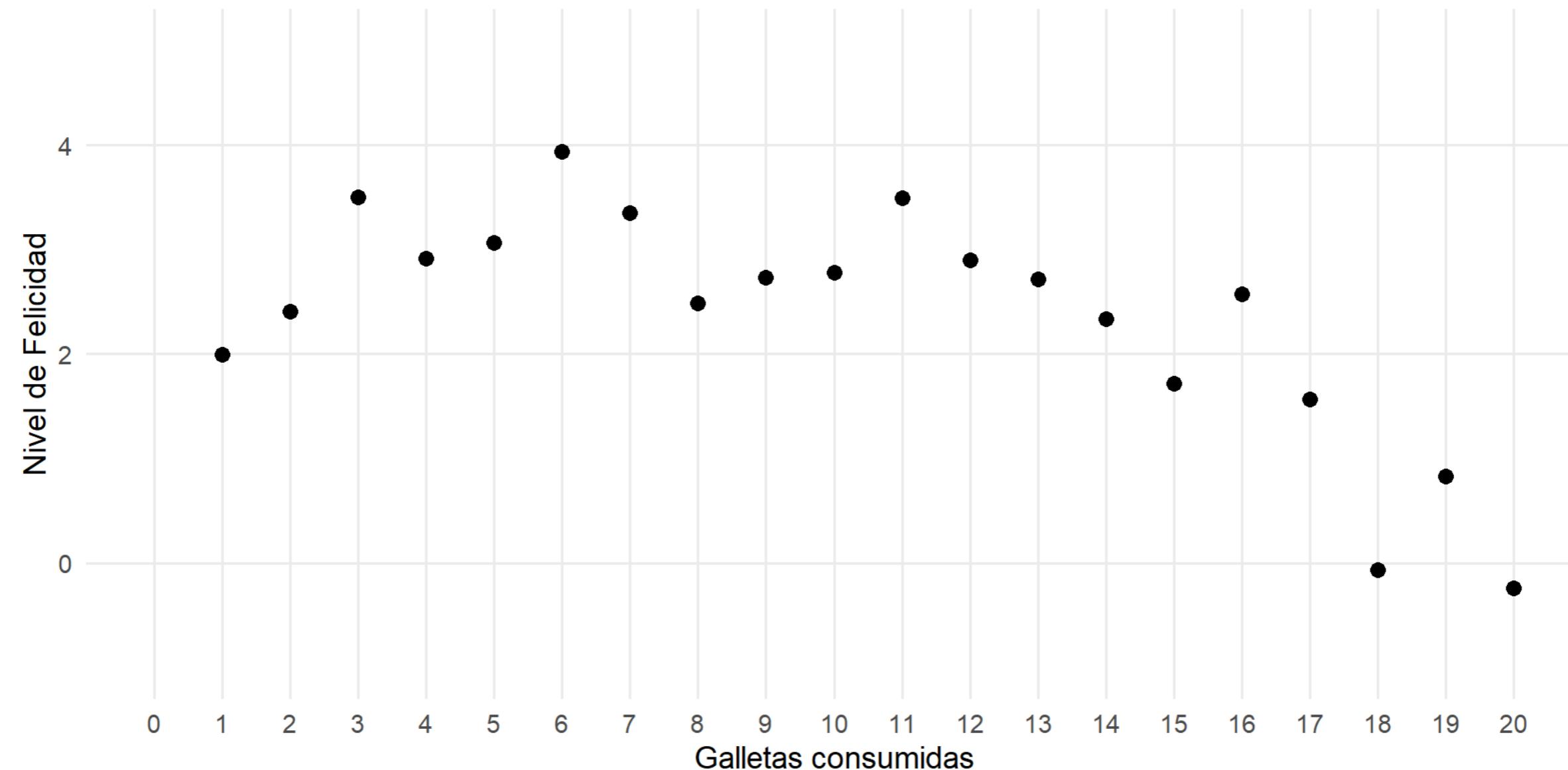
Consideraciones adicionales para una Regresión

Cuarteto de Anscombe

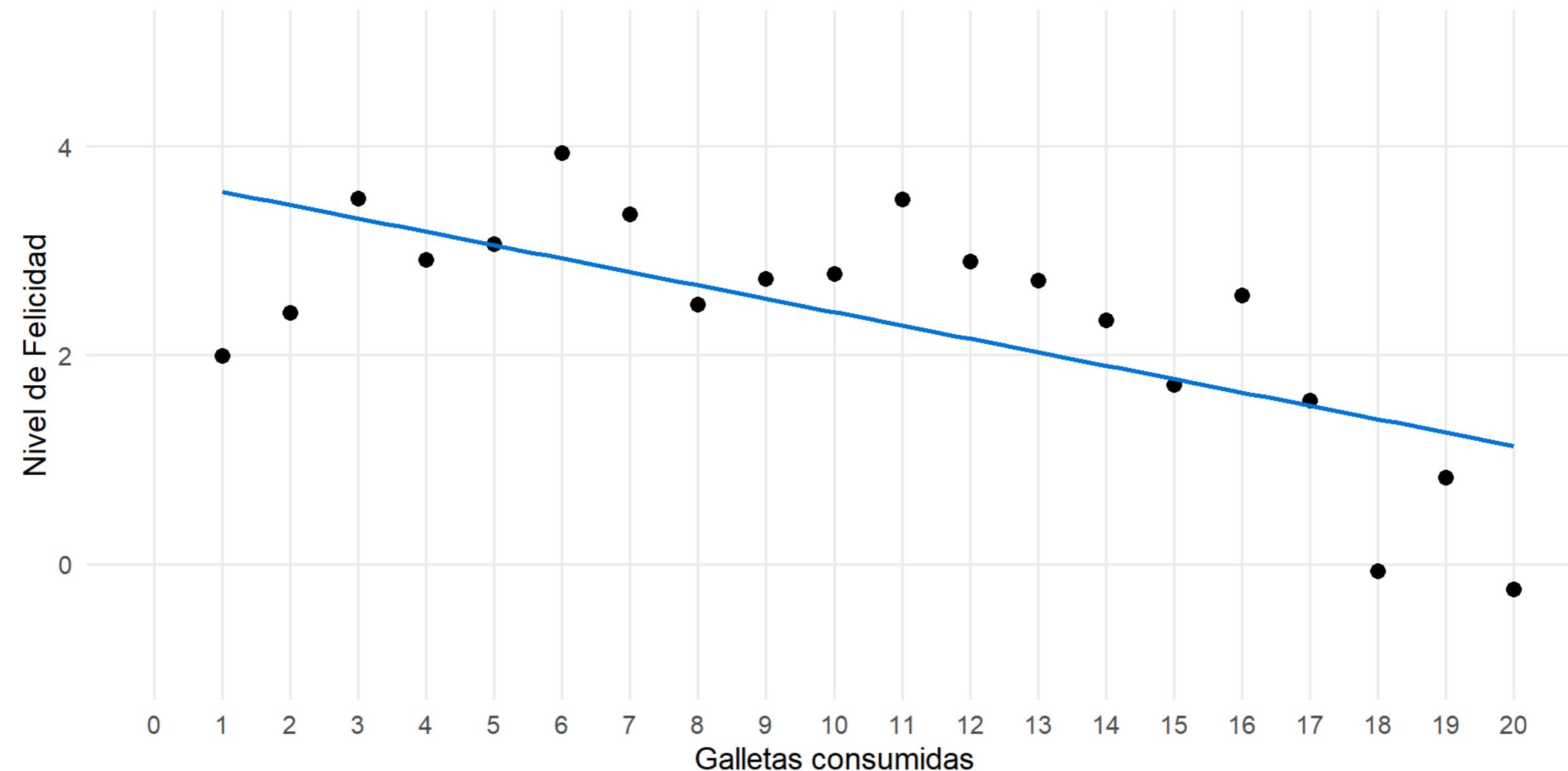
Anscombe's 4 Regression data sets



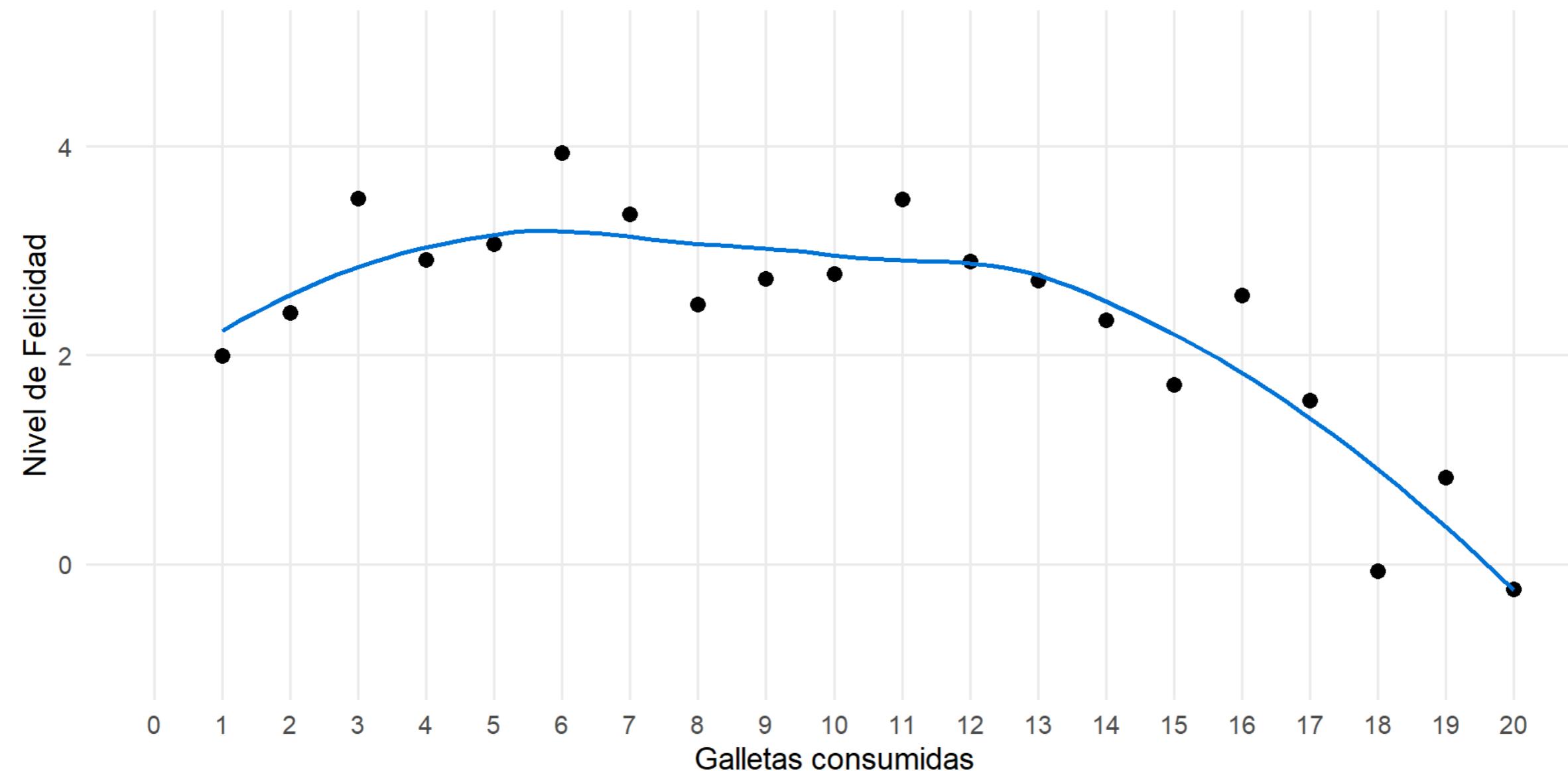
Relaciones no lineales



Relaciones no lineales



Relaciones no lineales



Relaciones no lineales

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

```
# A tibble: 3 x 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
```

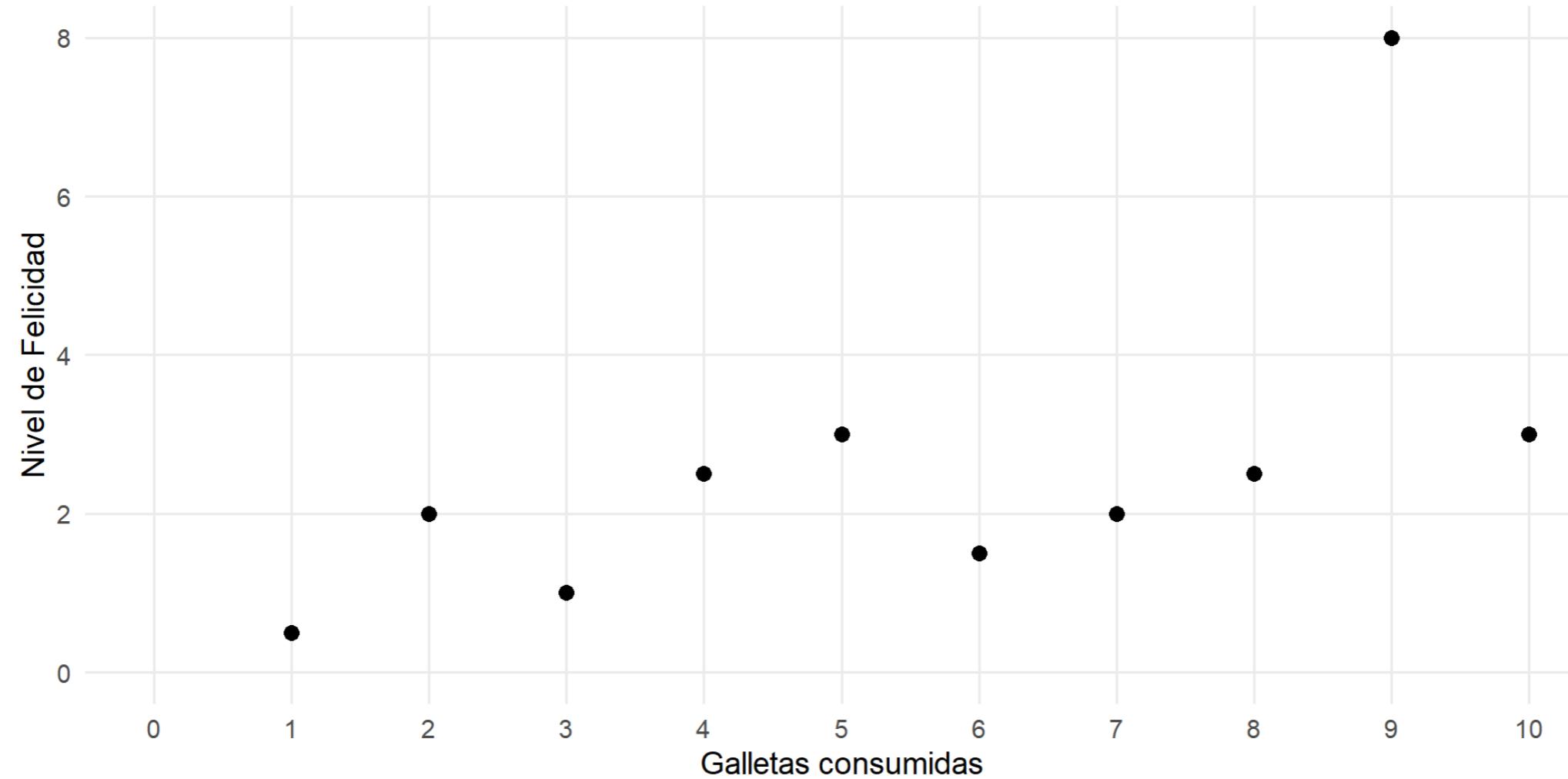
```
1 library(knitr)
2 tidy(modelo_felicidad, conf.int = TRUE) %>%
3   knitr::kable(format = "html") %>%
4   kableExtra::kable_styling(font_size = 30)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.0019260	0.3777083	5.300191	0.0000588	1.2050311	2.7988208
galletas	0.3337378	0.0828373	4.028835	0.0008711	0.1589664	0.5085091
I(galletas^2)	-0.0219886	0.0038316	-5.738727	0.0000241	-0.0300726	-0.0139046

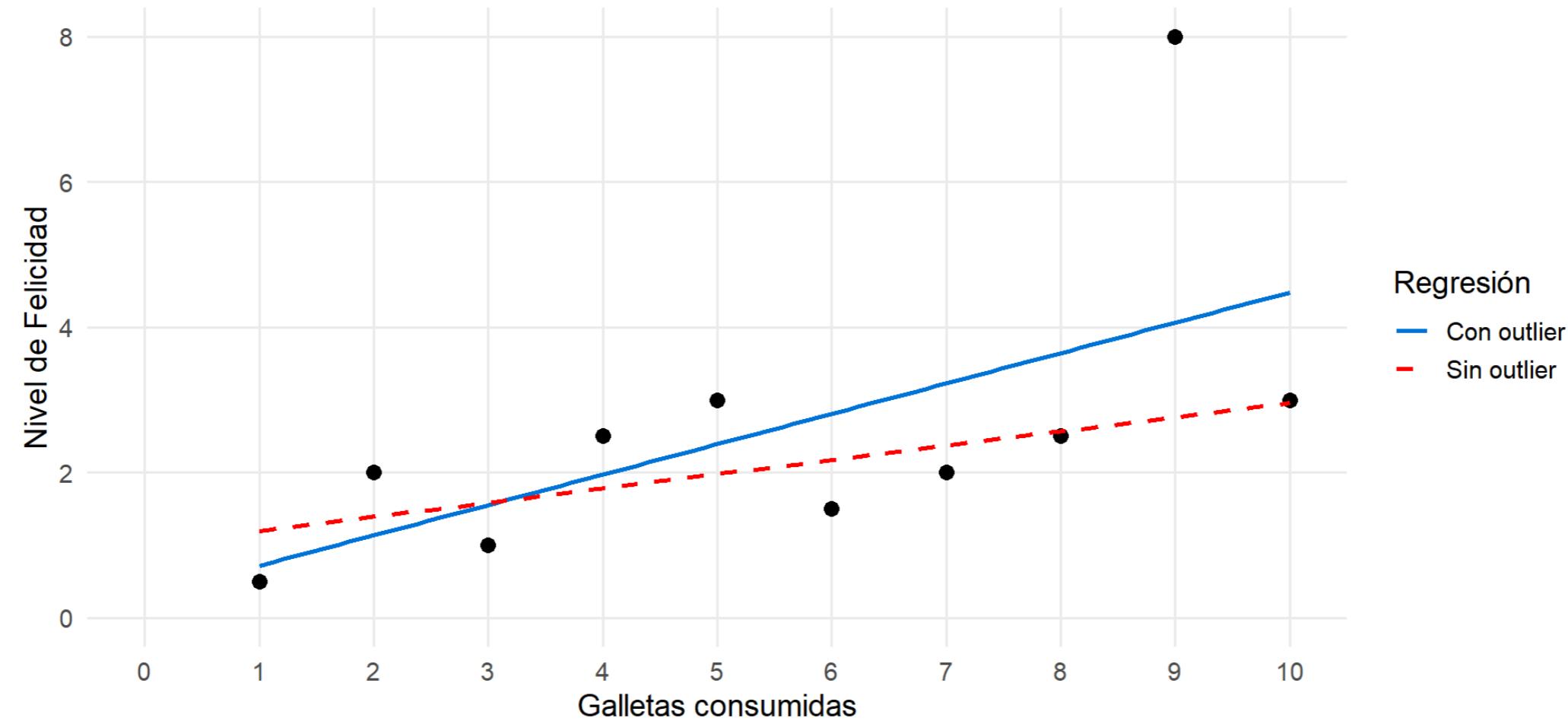
¿Existe la relación cuadrática que vemos en la gráfica?

Evaluamos con una prueba de hipótesis:

¿Qué hacer con una observación atípica (outlier)?



¿Qué hacer con una observación atípica (outlier)?



¿Cómo identificar los outliers?

No existen reglas estrictas. Depende del conocimiento del área y la comprensión de la recopilación de datos para identificar valores atípicos, inusuales e imposibles.

- Pero tenemos algunas herramientas:
 1. Histogramas
 2. Diagramas de caja
 3. Análisis de los residuos

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

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

Efecto interacción - Hollywood

El productor de Hollywood piensa que las críticas afectan menos el recaudo en US de las comedias que del resto de películas. Para eso estimamos:

```
1 hollywood <- hollywood %>%
2   mutate(comedy = ifelse(genre == "Comedy", 1, 0))
3 model_part9<- lm(us_gross ~ opening_gross + budget + mpaa_d
4                     + comedy * critics_opinion , data=hollywood)
5 tidy(model_part9, conf.int = TRUE)
```

A tibble: 7 x 7

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-36.3	8.47	-4.28	5.90e- 5	-53.2	-19.4
2	opening_gross	2.78	0.194	14.3	3.08e-22	2.39	3.16
3	budget	0.259	0.0951	2.72	8.24e- 3	0.0691	0.449
4	mpaa_d	-9.79	5.41	-1.81	7.46e- 2	-20.6	1.00
5	comedy	16.7	14.4	1.16	2.51e- 1	-12.1	45.4
6	critics_opinion	0.684	0.161	4.24	6.99e- 5	0.362	1.01
7	comedy:critics_opini~	-0.228	0.296	-0.772	4.43e- 1	-0.818	0.362

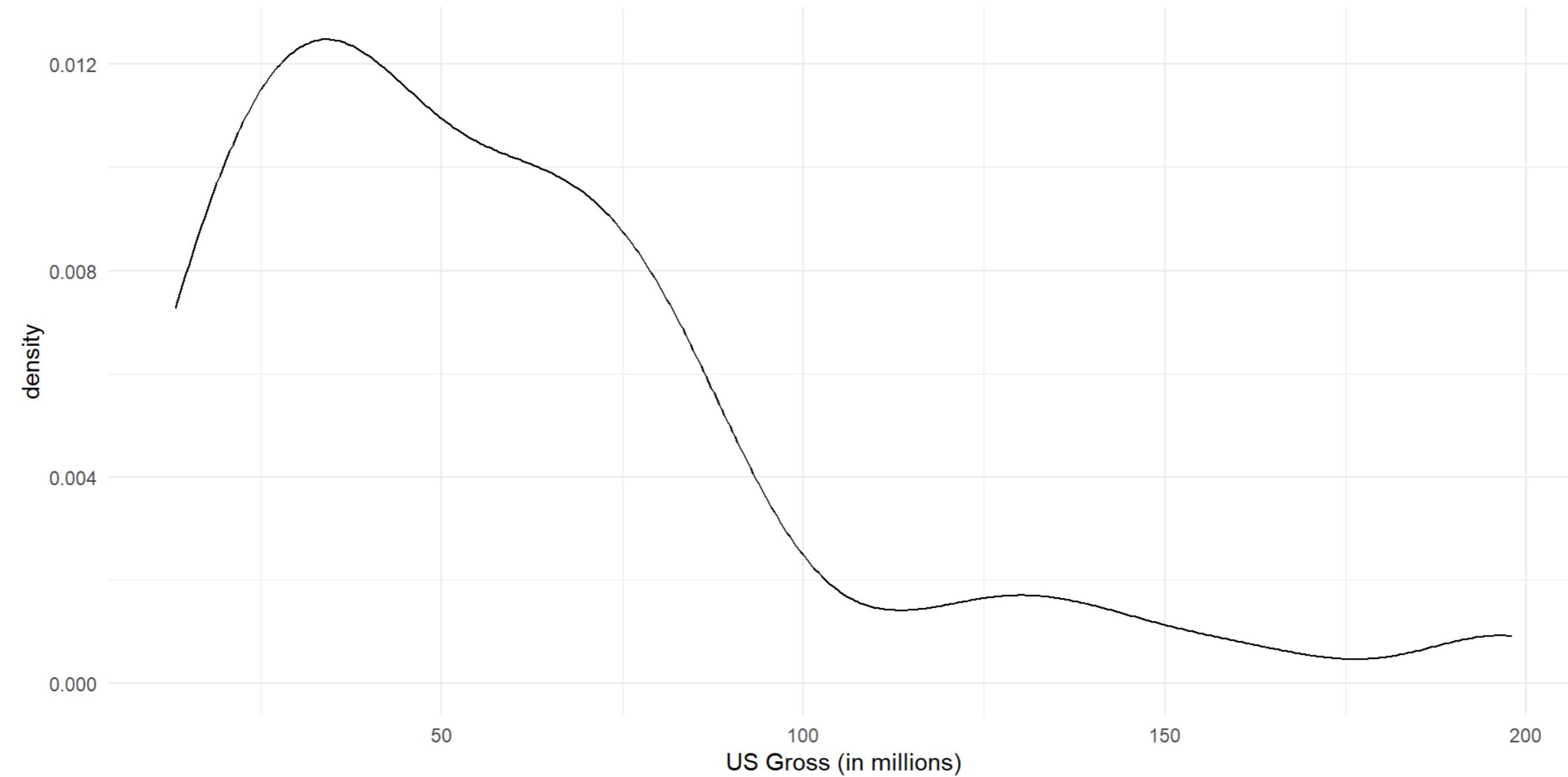
Transformación Logarítmica

Transformación logarítmica

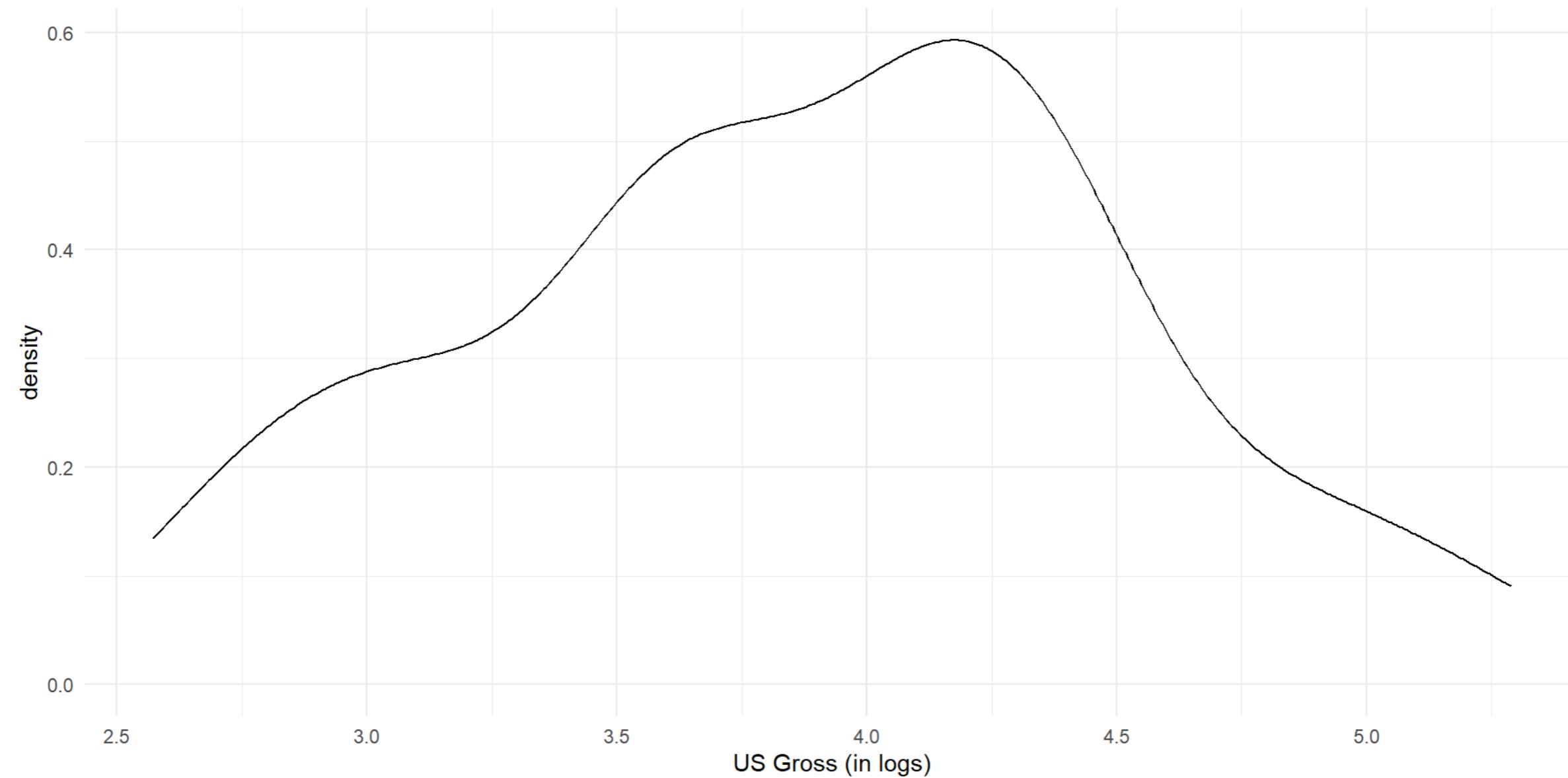
Podemos considerar una transformación logarítmica de las variables de nuestro modelo (tanto variable dependiente e independientes) cuando:

- 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).

Transformación logarítmica



Transformación logarítmica



Transformación logarítmica

```
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)  5.11      4.50     1.13  2.60e- 1    -3.87     14.1
2 opening_gross 3.12      0.218    14.3   7.07e-23     2.69     3.56
```

```
1 hollywood <- hollywood %>%
2   mutate(log_opening_gross=log(opening_gross),
3         log_us_gross=log(us_gross))
4
5 hollywood_model_logs <- lm(log_us_gross ~ log_opening_gross, data=hollywood)
6 tidy(hollywood_model_logs, 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)  1.25      0.181     6.91  1.50e- 9    0.892     1.61
2 log_opening_gross 0.977    0.0658    14.8   9.60e-24    0.845     1.11
```

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 x_1 , está asociado a un cambio de en β_1 .

