Análisis e Interpretación de Datos

Dra. Mariana Edith Miranda Varela

7 julio 2025

Regresión lineal simple

Dataset

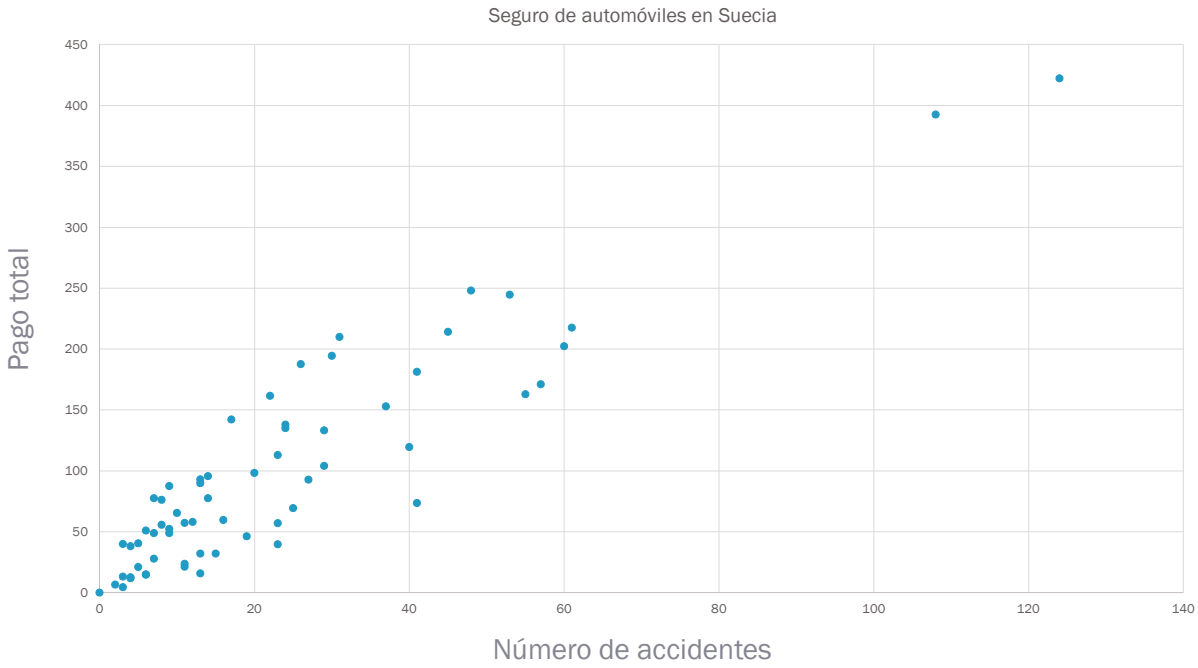
Seguro de automóviles en Suecia

X = número de accidentes

Y = pago total de todos los accidentes en miles de coronas suecas en Suecia

Referencias:
https://data.world/anujonthemove/auto-insurance-in-sweden
https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/slr06.html

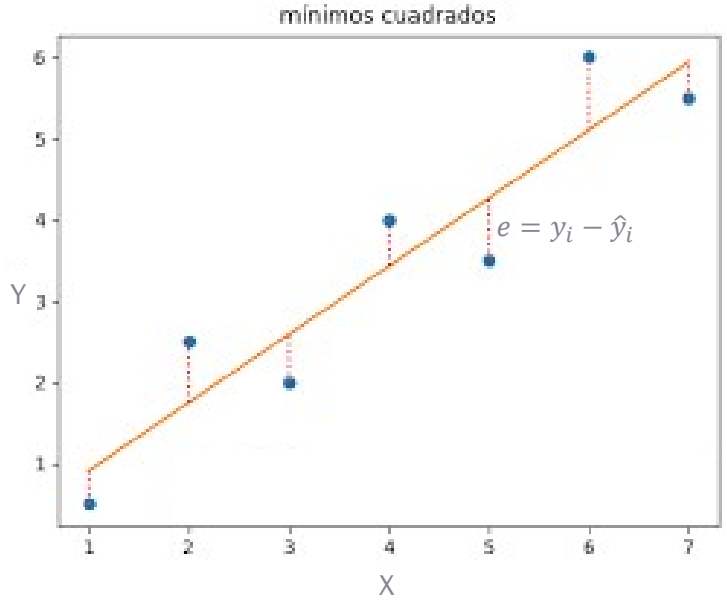
Gráfico de dispersión



Coeficiente de relación

0.91287824

Recta de mínimos cuadrados



$$\hat{y} = a + bx$$

$$\mathrm{b} = \mathrm{r}\frac{s_x}{s_y}$$

$$a = \bar{y} - b\bar{x}$$

Fuente: http://blog.espol.edu.ec/analisisnumerico/minimos-cuadrados-con-python/

Homocedasticidad

$$E(e_i^2) = \sigma_i^2$$

El término de error tiene la misma varianza en cada observación

Supuesto de homocedasticidad

Los residuos se distribuyen de manera homogénea para todos los valores de la variable predicha

Ausencia de homocedasticidad

- Omisión de variables importantes en el modelo

Comando lm

Variable dependiente

Variable independiente

```
> reg_S <- lm(Y ~ X, data = df_Suecia)
> summary(reg_S)

Call:
lm(formula = Y ~ X, data = df_Suecia)

Residuals:
    Min      1Q  Median      3Q     Max
-86.561 -24.051  -0.347  23.432  83.977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9945     6.3678    3.14   0.0026 **
X             3.4138     0.1955   17.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.94 on 61 degrees of freedom
Multiple R-squared:  0.8333,    Adjusted R-squared:  0.8306
F-statistic:   305 on 1 and 61 DF,  p-value: < 2.2e-16
```

$$\hat{y} = 19.9945 + 3.4138x$$

P-values para los coeficientes

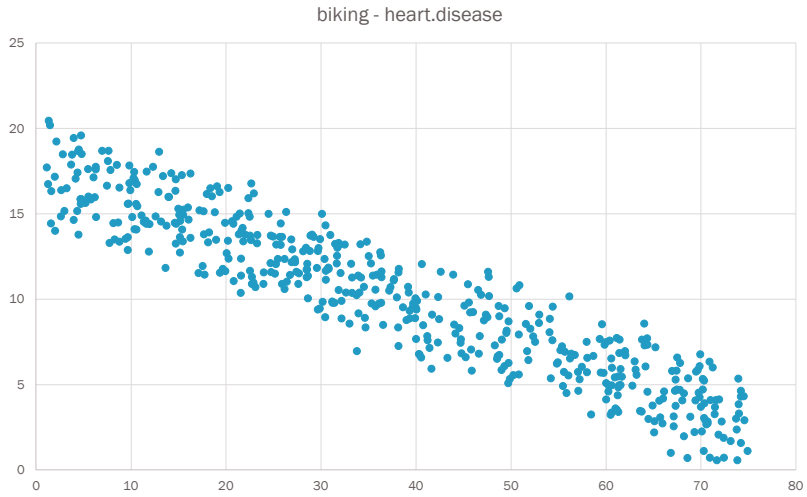
Coeficiente de determinación mide la bondad del ajuste de la recta a los datos

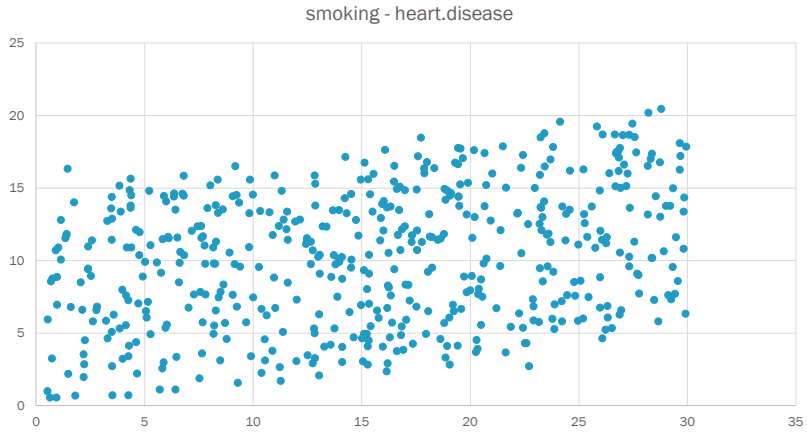
Regresión lineal múltiple

Enfermedades del corazón

- Variables independientes
  - biking -> **-0.935455474**
  - smoking -> **0.309130979**

- Variable dependiente
  - heart.disease





Otros tipos de regresión

- **Modelo lineal generalizado**
  - Gráfico de dispersión no muestra una relación lineal
  - Transformar variables
  - Función **glm**
- **Regresión de mínimos cuadrados recortados**
  - Se emplea en datos que tienen outliers
  - Función **ltsReg**

Gráficos de residuos

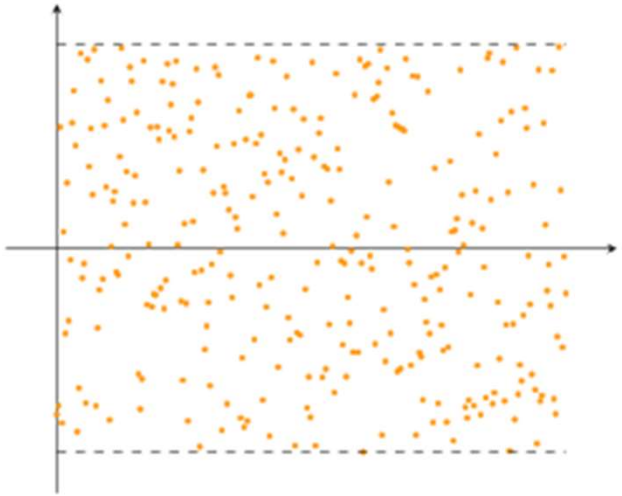
Gráfico de residuos

Permiten evaluar lo siguiente:

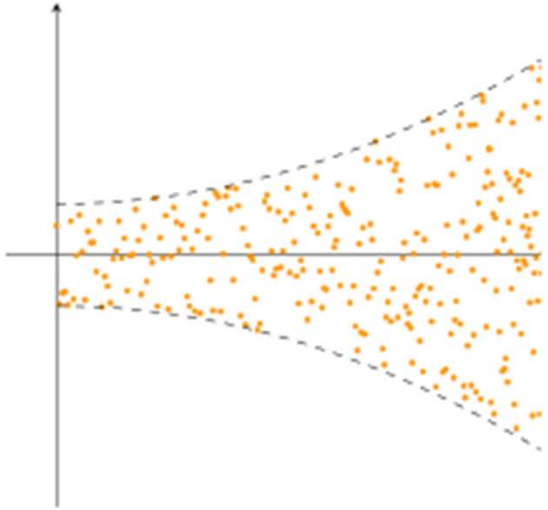
- Si el modelo debe ser no lineal en un lugar uno lineal
- Si la varianza es constante, esto es, los residuos se distribuyen al azar alrededor del cero.
- Si existen datos atípicos (outliers)

Homocedasticidad vs heterocedasticidad

Ejemplo de Homocedasticidad



Ejemplo de heterocedasticidad

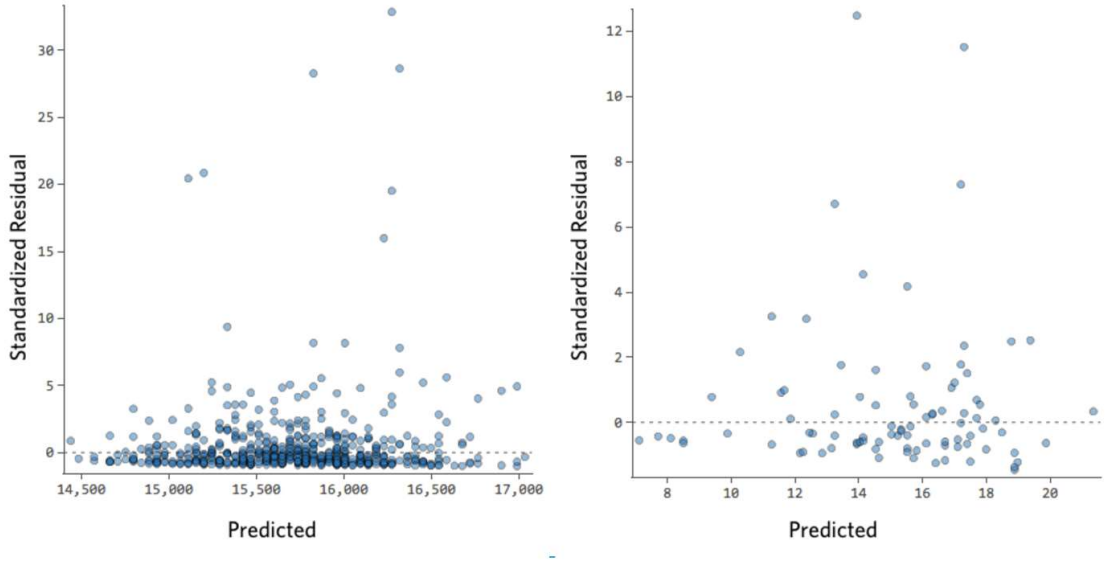


$$\hat{y}_i = a + bx$$

$$donde \;\; b\, x_i = 0$$

FUENTE: https://ecab-estadistica.medium.com/homocedasticidad-vs-heterocedasticidad-34b36cefacf9#:~:text=En%20concreto%2C%20un%20modelo%20lineal,contrario%2C%20el%20modelo%20ser%C3%A1%20heteroced%C3%A1stico.

Desbalance en el eje Y

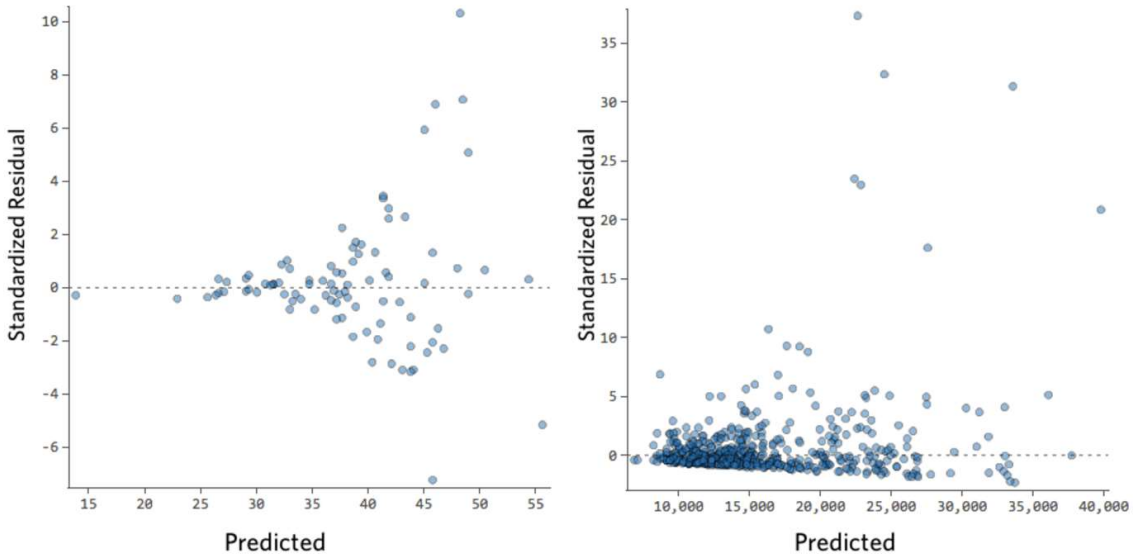


SOLUCIÓN

- Transformar los datos, específicamente la variable dependiente (de respuesta)
- Existe la posibilidad de que al modelo le falte una variable

FUENTE: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Heterocedasticidad



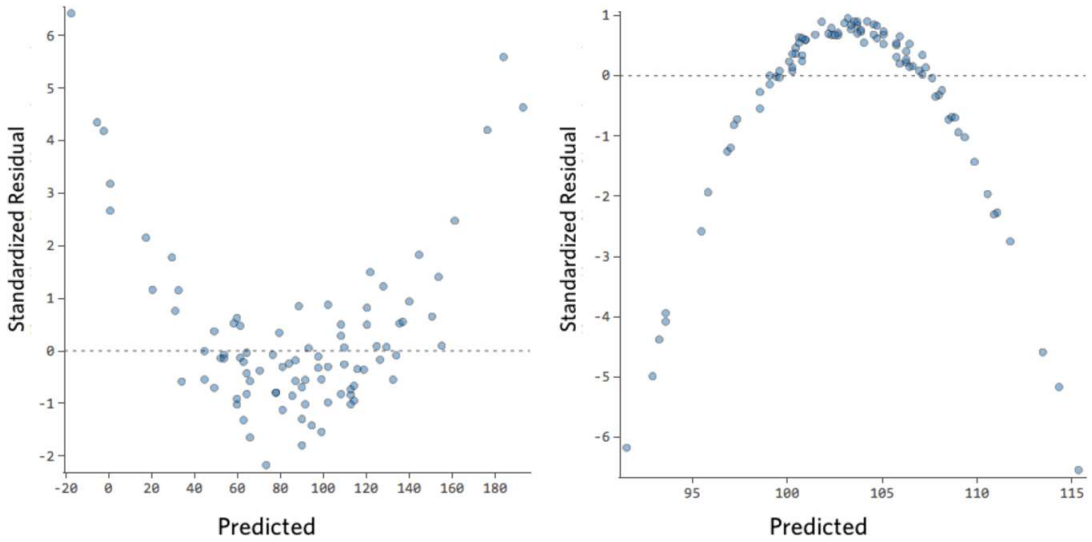
La heterocedasticidad se presenta cuando los residuos aumentan a medida que la predicción cambia de pequeña a grande, forma de embudo o abanico.

SOLUCIÓN

- Transformar una variable
- La heterocedasticidad indica que al modelo le falta una variable

FUENTE: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

No lineal



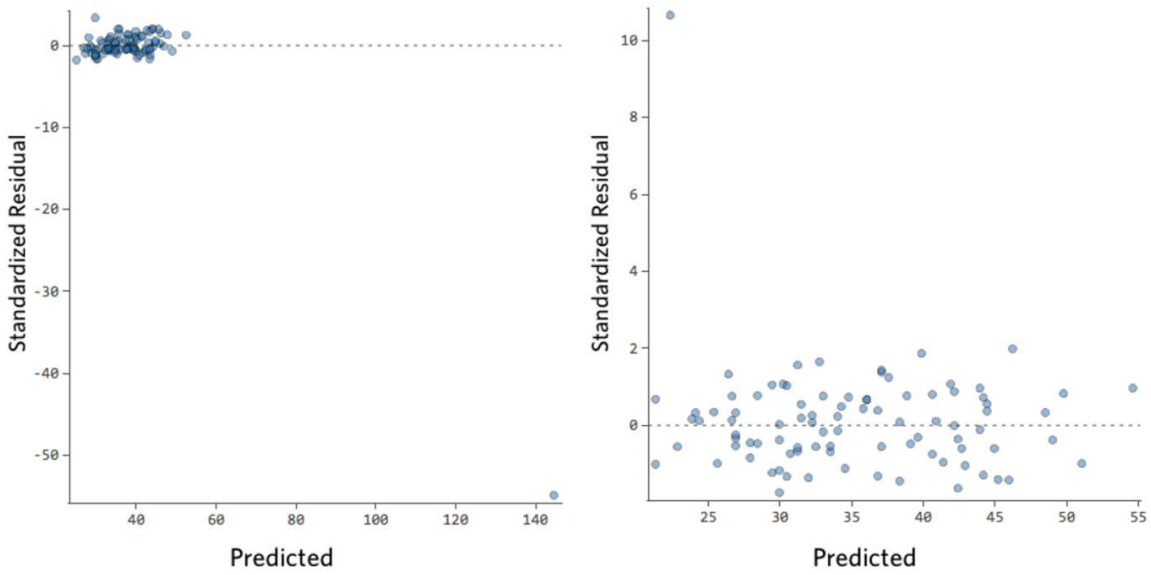
Los residuos tienen un patrón curvilíneo

SOLUCIÓN

- Transformar una variable
- Modelo de regresión no lineal
- Existe la posibilidad de que al modelo le falte una variable

FUENTE: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Outliers



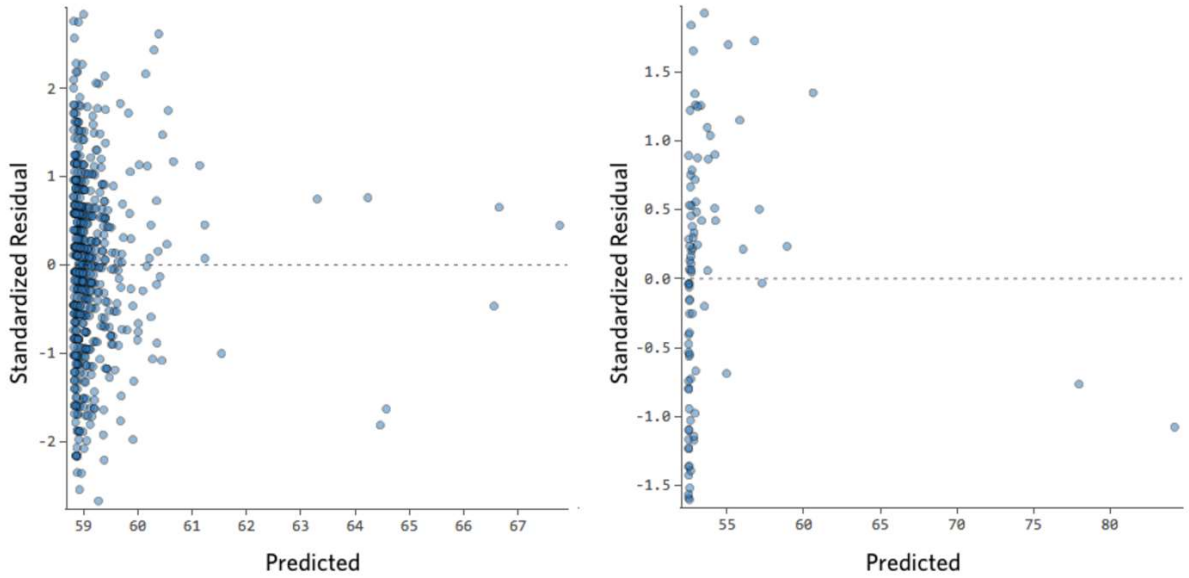
Un punto lejos del cero

SOLUCIÓN

- Borrar los datos outliers.
- Transformar la variable que tenga distribución asimétrica
- Evaluar el impacto del dato atípico en caso de que este correcto

FUENTE: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Desbalance en el eje X



SOLUCIÓN

- Transformar los datos, normalmente una variable independiente (explicativa)
- Existe la posibilidad de que al modelo le falte una variable

FUENTE: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Próxima sesión

Semana 7
21 al 25 de julio

muchas gracias.