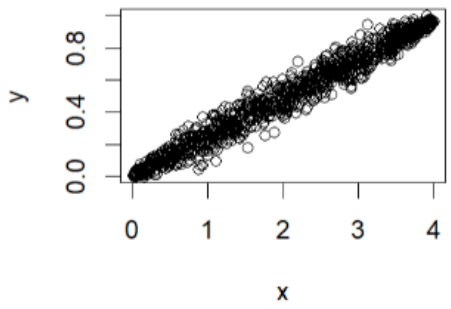
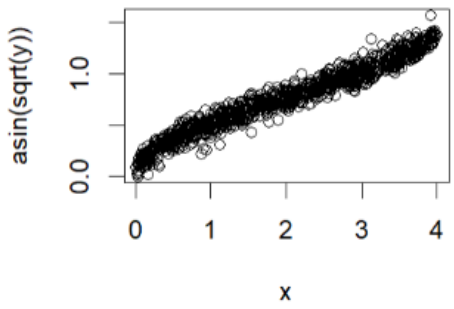
EJEMPLO 4

ORIGINAL (y) TRANSFORMADO: $\text{ASIN}(\sqrt{y}) = \sin^{-1}(\sqrt{y})$






```
> lm(y~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm(y ~ x))
W = 0.99238, p-value = 5.089e-05
```

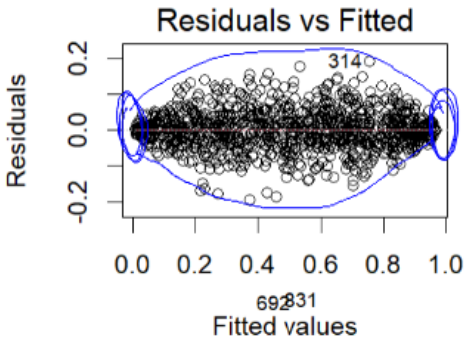
$< \alpha \rightarrow$ se rech. H_0 ✗ Normalidad

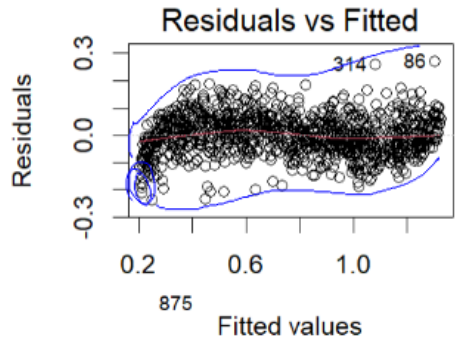
```
> lm(asin(sqrt(y))~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm(asin(sqrt(y)) ~ x))
W = 0.99792, p-value = 0.2492
```

$> \alpha \rightarrow$ No se rech. H_0 ✓ Normalidad






```
> lm(y~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.645588, Df = 1, p = 0.19956
Hubo 12 avisos (use warnings() para verlos)
> lm(asin(sqrt(y))~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.654319, Df = 1, p = 0.030976
```

$> \alpha$ ↗ sí cumple

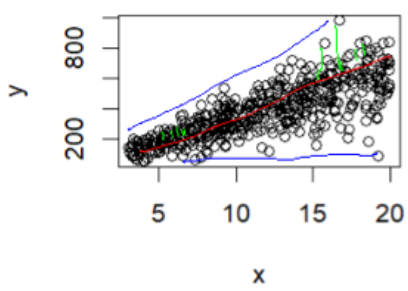
$< \alpha$ ↗ NO cumple

otra opción: otra transf.

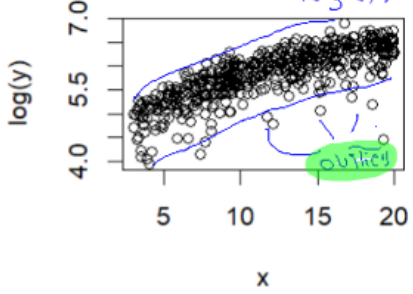
($y \in (0,1) \rightarrow$ Regresión Beta)

EJEMPLO 5

original



transformado log(y)




```
> lm(y~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm(y ~ x))
W = 0.97856, p-value = 1.017e-06
```

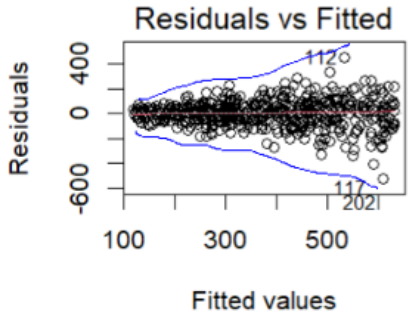
$< \alpha \rightarrow$ No normalidad

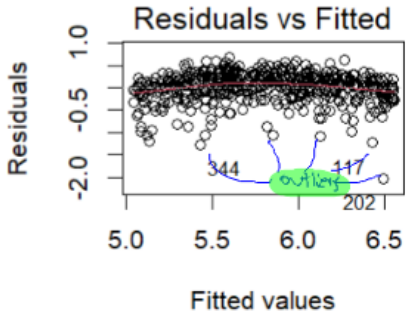
```
> lm(log(y)~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm(log(y) ~ x))
W = 0.92222, p-value = 2.127e-15
```

$< \alpha \rightarrow$ No normalidad






```
> lm(y~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 99.64723, Df = 1, p = < 2.22e-16
```

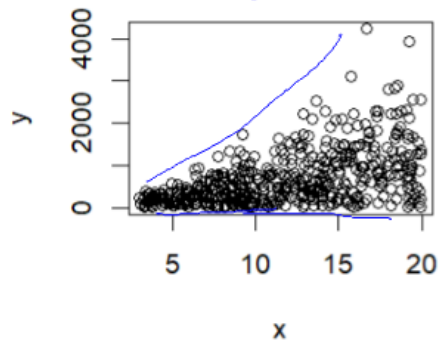
$< \alpha$ No homocedasticidad

```
> lm(log(y)~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.385331, Df = 1, p = 0.53476
```

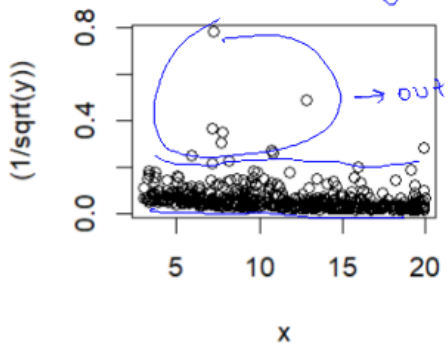
$> \alpha$ Sí homocedasticidad

EJEMPLO 6

original



transformado: $y' = \frac{1}{\sqrt{y}}$




```
> lm(y~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm(y ~ x))
W = 0.93334, p-value = 3.819e-14
```

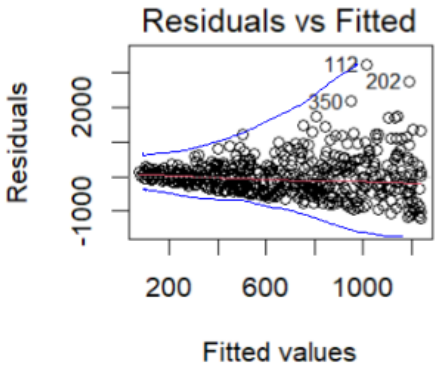
$< \alpha$

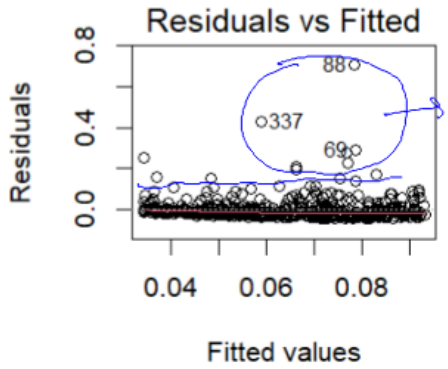
```
> lm((1/sqrt(y))~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm((1/sqrt(y)) ~ x))
W = 0.54536, p-value < 2.2e-16
```

$< \alpha$





→ outliers

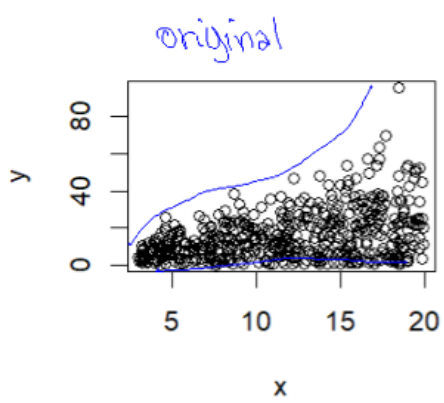
```
> lm(y~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 151.1686, Df = 1, p = < 2.22e-16
> lm((1/sqrt(y))~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 39.69094, Df = 1, p = 2.975e-10
```

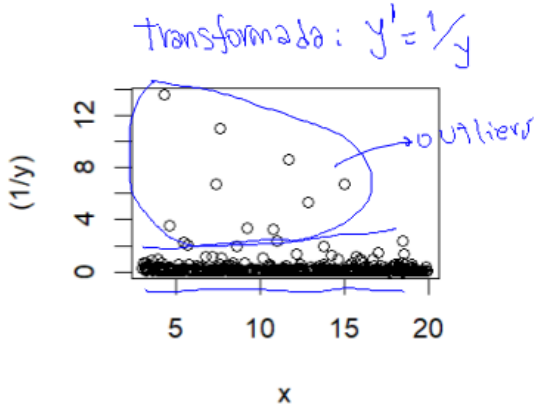
$< \alpha$

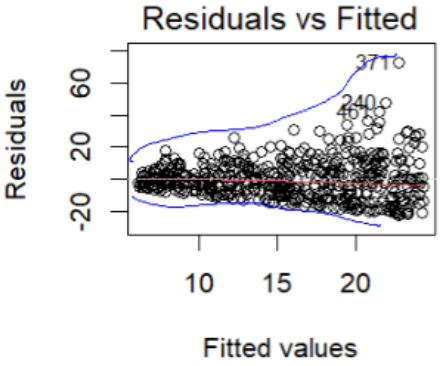
$< \alpha$

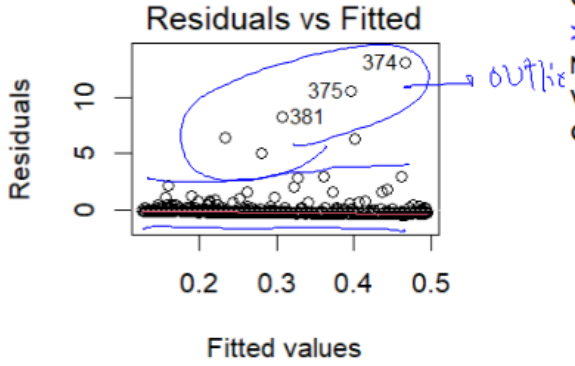
↓ aumentó

EJEMPLO 7










```
> lm(y~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm(y ~ x))
W = 0.95254, p-value = 1.371e-11
```
$< \alpha$

```
> lm((1/y)~x) |> residuals() |> shapiro.test()

        Shapiro-Wilk normality test

data:  residuals(lm((1/y) ~ x))
W = 0.26127, p-value < 2.2e-16
```
$< \alpha$

```
> lm(y~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 108.0548, Df = 1, p = < 2.22e-16
```
$< \alpha$

```
> lm((1/y)~x) |> car::ncvTest()
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 134.7769, Df = 1, p = < 2.22e-16
```
$< \alpha$

Si no se logra resolver la heteroscedasticidad

$E(\hat{\beta}) = \beta$

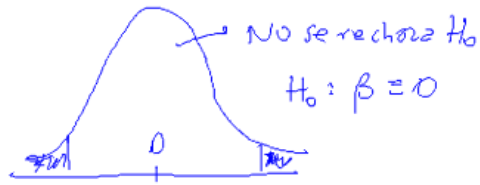
Si no se logra estabilizar la varianza, los estimadores seguirán siendo insesgados pero no tendrán la propiedad de mínima varianza, lo que significa que los errores estándar serán ¿más pequeños o más grandes? ¿Cómo afecta eso en sus intervalos de confianza y pruebas de hipótesis asociadas?

↳ más ancho

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.73054    0.88750  -0.823  0.42064
Educacion    0.15695    0.06238   2.516  0.02102 *
SexoM        0.82211    0.48017   1.712  0.10315
Edad         0.10463    0.03004   3.482  0.00249 **
X4          -0.04538    0.04043  -1.122  0.27573
---
```

↑ ↑ ↑

$t_{calc} = \frac{\hat{\beta} - \beta}{S_{\hat{\beta}}} \Rightarrow t_{calc} \downarrow$



Cuando la variable respuesta es transformada, los valores predichos se encontrarán en la escala de la nueva variable. Es común convertir dichos valores a la escala original, sin embargo esa transformación inversa genera la estimación de la mediana de la distribución de la variable endógena en vez de la media. En cuanto a los límites de los intervalos de confianza para estimar o predecir un valor individual, estos pueden ser directamente transformados ya que se tratan de percentiles, los cuales no se ven afectados por transformaciones. Sin embargo, no es posible asegurar que los intervalos obtenidos son los más cortos posibles.

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log(Y) = \hat{\beta}_0^* + \hat{\beta}_1^* x_1 + \hat{\beta}_2^* x_2 \quad \ldots \rightarrow \text{predecir } \log(y)$$

$$* \; e^{\log(y)} = y$$

TRANSFORMACIÓN BOX COX

```
lm((y**0.25-1)/0.25~x) |> aov() |> tidy()
```

```
library(forecast)
lm(BoxCox(y, 0.25)~x) |> aov() |> tidy()
```

```
lm(BoxCox(y, 0.25)~x) |> aov() |> tidy()
```

```
# A tibble: 2 x 6
  term         df  sumsq  meansq statistic    p.value
  <chr>     <dbl>  <dbl>   <dbl>     <dbl>      <dbl>
1 x             1  1508. 1508.        541.  9.35e-114
2 Residuals  4998 13927.    2.79       NA   NA
```

13 927

```
lm(BoxCox(y, 0.50)~x) |> aov() |> tidy()
```

```
# A tibble: 2 x 6
  term         df  sumsq  meansq statistic    p.value
  <chr>     <dbl>  <dbl>   <dbl>     <dbl>      <dbl>
1 x             1  5595. 5595.        657.  3.17e-136
2 Residuals  4998 42568.    8.52       NA   NA
```

42 568

```
lm(BoxCox(y, 0.73)~x) |> aov() |> tidy()
```

```
# A tibble: 2 x 6
  term         df   sumsq  meansq statistic    p.value
  <chr>     <dbl>   <dbl>   <dbl>     <dbl>      <dbl>
1 x             1  20063. 20063.        727.  1.53e-149
2 Residuals  4998 138002.    27.6       NA   NA
```

138 002

Para elegir el valor óptimo de λ se busca aquel que maximiza la (log) verosimilitud al aplicar la transformación:

$$y' = \frac{y^{\lambda} - 1}{\lambda \hat{Y}^{\lambda - 1}} \qquad \hat{Y} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(y_i)\right)$$

Esto último permite que la variabilidad (a través de las sumas de cuadrados) de los modelos con diferentes λ sean comparables, lo cual se puede apreciar en estos ejemplos:

```
lambda = 0.25
exp(1/length(y)*sum(log(y))) -> yp
(y**lambda-1)/(lambda*yp**(lambda-1)) -> ynueva1
lm(ynueva1~x) |> aov() |> tidy()
```

```
> boxcox(lm(y~x),lambda=seq(0.35,0.45,0.01),plotit = FALSE) |>
+   rbind.data.frame()
      x         y
1  0.35 -1538.267
2  0.36 -1537.991
3  0.37 -1537.794
4  0.38 -1537.675
5  0.39 -1537.633
6  0.40 -1537.668
7  0.41 -1537.778
8  0.42 -1537.963
9  0.43 -1538.222
10 0.44 -1538.553
11 0.45 -1538.957
```

→ la mayor logVero

λ logverosimilitud

$\lambda = 0.39$

```
> boxcox
otit = F
+   rbin
```

Mínimos Cuadrados Generalizados

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad , \quad V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} = \sigma^2 \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}_{n \times n}$$

$$V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} = \begin{pmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \sigma^2 \end{pmatrix}_{n \times n}$$

$$\operatorname{diag}(\sigma^2 I) = \underbrace{(\sigma^2 \sigma^2 \dots \sigma^2)}_{\text{homogeneidad de varianzas}}$$

$\Downarrow$

$\sigma^2 \mathbf{V}$ → existen las covarianzas $\neq 0$ (no es la matriz identidad)

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\underbrace{K^{-1} Y}_{Z} = \underbrace{K^{-1} X}_{B} \beta + \underbrace{K^{-1} \varepsilon}_{g}$$

$$Z = B\beta + g$$

$$\hat{\beta} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{z} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad \mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \qquad \mathbf{B} = \mathbf{K}^{-1}\mathbf{X}, \qquad \mathbf{g} = \mathbf{K}^{-1}\varepsilon$$

$$(X' K^{-1'} K^{-1} X)^{-1} X' K^{-1'} K^{-1} y$$

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} y$$

Modelo tradicional (mínimos cuadrados ordinarios): $\hat{\beta} = (X'X)^{-1}X'y$

Mínimos cuadrados generalizados: $\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} y$

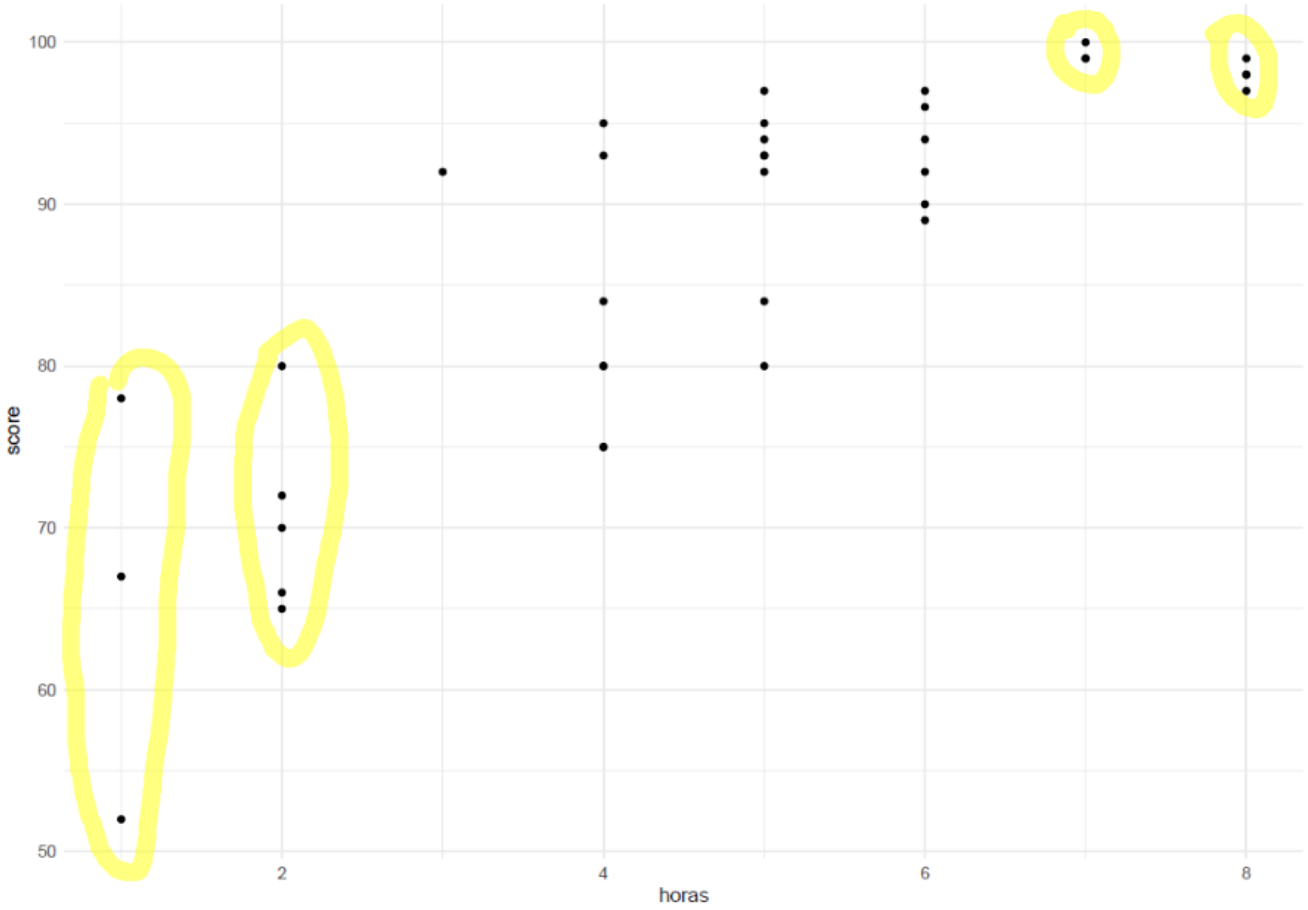
$V \neq I$

Mínimos Cuadrados Ponderados

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad V(\boldsymbol{\varepsilon}) = \sigma^2 V = \sigma^2 \begin{pmatrix} \frac{1}{\omega_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\omega_m} \end{pmatrix} = \begin{pmatrix} \sigma^2/\omega_1 & 0 & \cdots & 0 \\ 0 & \sigma^2/\omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2/\omega_n \end{pmatrix}$$

$$\operatorname{diag}(\sigma^2 V) = \left(\frac{\sigma^2}{\omega_1} \quad \frac{\sigma^2}{\omega_2} \quad \cdots \quad \frac{\sigma^2}{\omega_n} \right)$$

Varianzas distintas = heterocedasticidad



X ↑ Variab.Y ↓

$$V(\epsilon_i) = \frac{\sigma^2}{x_i}.$$

Primera propuesta: A mayor peso, menor variabilidad de los errores, es decir $V(\epsilon_i) = \frac{\sigma^2}{x_i}$. De ahí que $V_i = \frac{1}{x_i}$ y $\omega_i = x_i$.

```
datos2$horas -> peso1
lm(score ~ horas, data = datos2, weights=peso1) -> modelo1
```

Segunda propuesta: La inversa al cuadrado de los valores ajustados resultantes de la regresión de los residuales absolutos en función de los valores ajustados de la regresión por mínimos cuadrados ordinarios. (sin pesos)

```
1/lm(abs(modelo0$residuals) ~ modelo0$fitted.values)$fitted.values**2 -> peso3
```

```
lm(score ~ horas, data = datos2, weights=peso3) -> modelo3
modelo3 |> residuals() |> shapiro.test()
```

```
    Shapiro-Wilk normality test

data:  residuals(modelo3)
W = 0.97271, p-value = 0.4865
```

$> \alpha$ ✓ Normalidad de errores

```
modelo3 |> car::ncvTest()
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.592739, Df = 1, p = 0.20694
```

$> \alpha$ ✓ Homocedasticidad de errores