Diagnóstico

En general, en presencia de heteroscedasticidad se suelen tomar una de las siguientes medidas: utilizar pesos o transformar las variables.

En muchas ocasiones la transformación de la variable dependiente o una de las independientes puede ser mucha utilidad. En general, las transformaciones son usadas para estabilizar varianzas, simplificar modelos u obtener normalidad.

Detección de Heteroscedasticidad

En algunos casos el reconocer la naturaleza de la variable dependiente puede prevenirnos sobre la heterogeneidad de varianzas.

De hecho, si la variable de respuesta fuese una variable de conteo de tipo Poisson, tendríamos que $\sigma_i^2 \simeq E(Y_i)$ y por lo tanto no podríamos esperar que

se cumpla el supuesto de homoscedasticidad.

Aún cuando no conozcamos la distribución de Y_i podemos tener una idea aproximada del comportamiento de su varianza.

Ejemplos

Mercado inmobiliario: Variación de precio de venta de casas.

σ constante:

$$\begin{aligned} 50{,}000\$ &\longleftrightarrow 100{,}000\$ \\ 1{,}000{,}000\$ &\longleftrightarrow 1{,}050{,}000\$ \end{aligned}$$

σ No constante:

$$\begin{aligned} 50{,}000\$ &\longleftrightarrow 60{,}000\$ \\ 1{,}000{,}000\$ &\longleftrightarrow 1{,}200{,}000\$ \end{aligned}$$

Ejemplo Los siguientes datos representan el tiempo de viaje (y) al centro de una ciudad y la distancia recorrida (x).

Distancia (en millas)	*.5*	*1*	*1.5*	*2*	*3*	*4*	*5*	*6*	*8*	*10*
Tiempo viajado (en minutos)	*15*	*15.1*	*16.5*	*19.9*	*27.7*	*29.7*	*26.7*	*35.9*	*42*	*49.4*

Supongamos $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

No parece razonable suponer que la varianza sea constante.

De hecho, si la longitud de una cuadra en esta ciudad es d, el viaje de x millas comprende x/d cuadras. Luego, y puede ser expresada como

$$y = z_1 + z_2 + \ldots + z_{x/d}\,,$$

donde $z_j, j = 1, \ldots, x/d$ es el tiempo en recorrer la j–ésima cuadra.

Si suponemos que las z_j son v.a. independientes y con la misma varianza ten-

dríamos que:

$$\begin{aligned} v(y) &= v(z_1) + v(z_2) + \ldots + v(z_{x/d}) = (x/d)v(z_j) \\ &\propto x\sigma^2 \,. \end{aligned}$$

Usando el método de mínimos cuadrados ponderados con pesos $w_i = 1/x_i$ obtenemos los estimadores $\widehat{\beta}_0 = 12{,}561$ y $\widehat{\beta}_1 = 3{,}714$.

Por lo tanto:

$$\begin{aligned} \text{si } y_i \text{ v.a. conteo} &\longrightarrow \sigma_i^2 \simeq E(Y_i) \\ \text{si } y_i = \frac{m_i}{n_i} &\longrightarrow \sigma_i^2 = \frac{E(Y_i)(1 - E(Y_i))}{n_i} \\ \text{si } y_i = \sum_{i=1}^{n_i} \frac{z_{ij}}{n_i} &\longrightarrow \sigma_i^2 = \frac{\sigma^2}{n_i} \quad \text{si } z_{ij} \text{ homoscedásticos} \end{aligned}$$

¿Cómo diagnosticar?

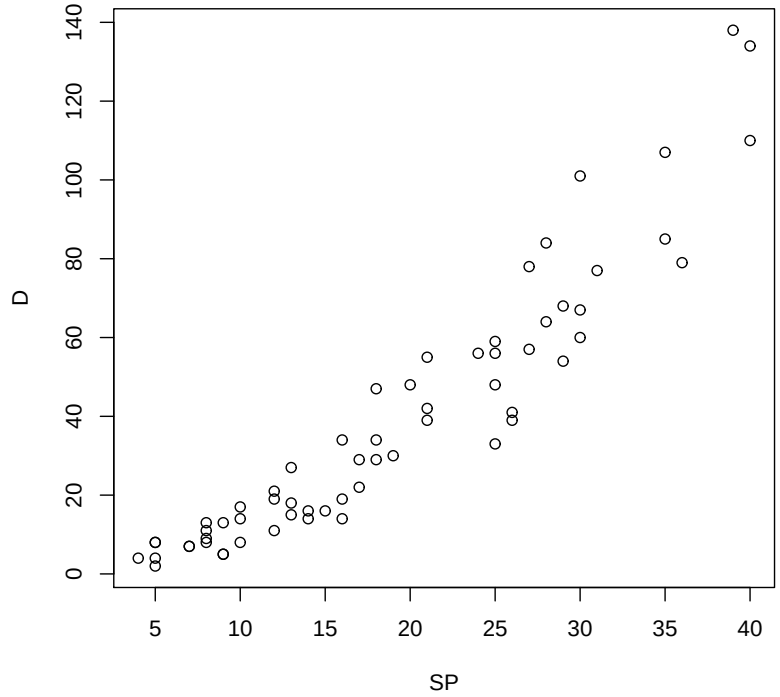
Hemos recomendado el gráfico de r_i vs. $\hat{Y}_i$ para detectar heteroscedasticidad, ¿cómo podemos ver en él la relación entre $V(Y_i)$ y $E(Y_i)$?

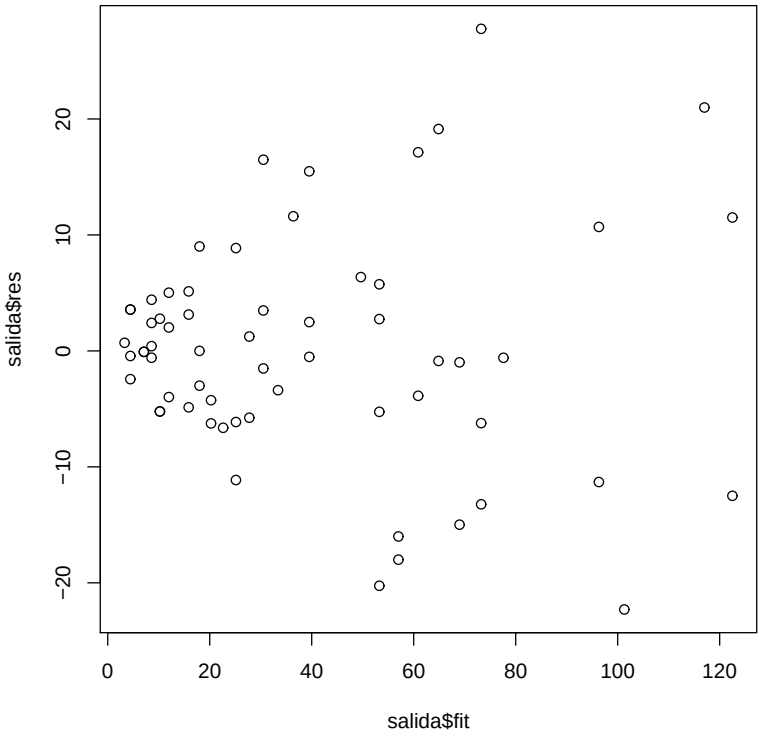
Consideremos el siguiente ejemplo.

El siguiente plot corresponde a datos de velocidad (SP) y distancia de frenado en un automóvil (D).

En este gráfico se puede ver cierta heteroscedaticidad que es mucho más evidente en el gráfico de r_i vs. $\hat{Y}_i$ obtenidos después de ajustar a los datos el modelo

$$D_i = \beta_1 SP_i + \beta_2 SP_i^2 + \epsilon_i \, .$$


D
SP
0
20
40
60
80
100
120
140
5
10
15
20
25
30
35
40


salida$res
20
10
0
−10
−20
0
20
40
60
80
100
120
salida$fit

¿Cómo podemos determinar la relación entre $V(Y_i)$ y $E(Y_i)$?

Un procedimiento es obtener estimadores de $E(Y_i)$ y de $V(Y_i)$ por regiones y tratar de establecer que relación hay entre ellas.

Para ello, se recomienda dividir el rango de $\hat{Y}_i$ en tres regiones de manera de hacer un compromiso entre que las regiones tengan igual tamaño e igual cantidad de puntos cada una.

En el ejemplo de velocidad, estas regiones podrían estar delimitadas por los valores **25** y **72**.

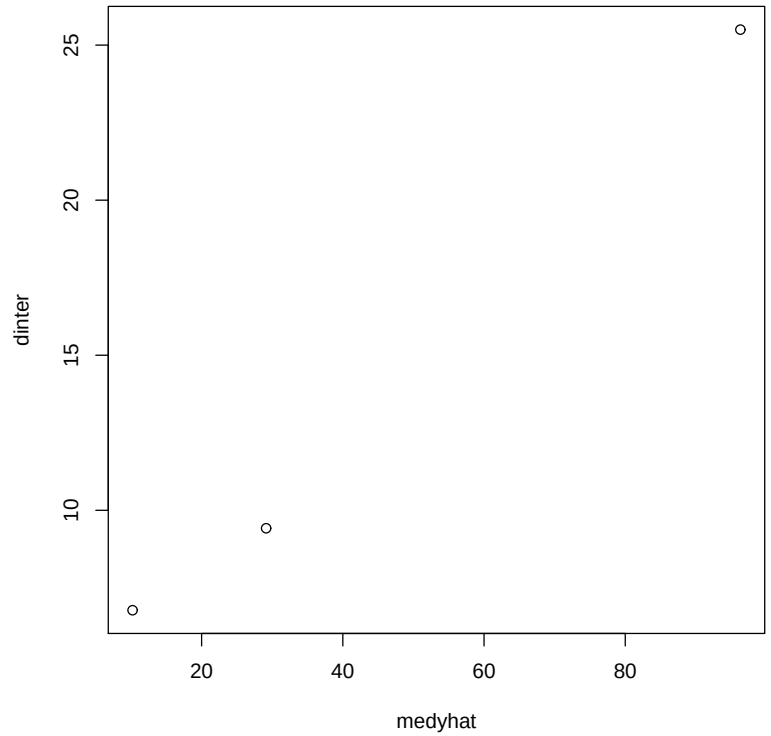
Luego calcularíamos

- la mediana de cada región: $\hat{Y}^{(1)}$, $\hat{Y}^{(2)}$ y $\hat{Y}^{(3)}$
- la distancia intercuartil de cada una: $d^{(1)}$, $d^{(2)}$ y $d^{(3)}$
- graficamos $\hat{Y}^{(i)}$ vs. $d^{(i)}$

En nuestro ejemplo, obtendríamos
$(\hat{Y}^{(1)}, \hat{Y}^{(2)}, \hat{Y}^{(3)}) = (10{,}22315; 29{,}13797; 96{,}30877)$

$(d^{(1)}, d^{(2)}, d^{(3)}) = (6{,}778; 9{,}421; 25{,}500)$

graficando, obtenemos



Este gráfico sugiere que $\sqrt{V(Y_i)} \cong \alpha E(Y_i)$ y por lo tanto

$$V(Y_i) \propto E^2(Y_i)$$

Recordemos que cuando $\Sigma_{\mathbf{Y}} \neq \sigma^2 \mathbf{I}$ aplicábamos el método de mínimos cuadrados generalizados o ponderados:

Supongamos que $\Sigma_{\mathbf{Y}} = \sigma^2 \mathbf{V}$, donde $\mathbf{V} \in \Re^{n \times n}$ es una matriz definida positiva de constantes. Podemos entonces escribir: $\mathbf{V} = \mathbf{K}\mathbf{K}'$ con $\mathbf{K}$ una matriz invertible.

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \mathbf{K}^{-1}\mathbf{Y} &= \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^{-1}\boldsymbol{\epsilon} \end{aligned}$$

Por lo tanto, tenemos un nuevo problema transformado es:

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

que satisface las condiciones de Ω.

Hallar el estimador de mínimos cuadrados, $\tilde{\boldsymbol{\beta}}$, en el problema transformado equivale a:

$$\min_{\mathbf{b}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\mathbf{b}\|^2 = \min_{\mathbf{b}} (\mathbf{Y} - \mathbf{X}\mathbf{b})' \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{b})$$

Para el modelo transformado, los residuos serían

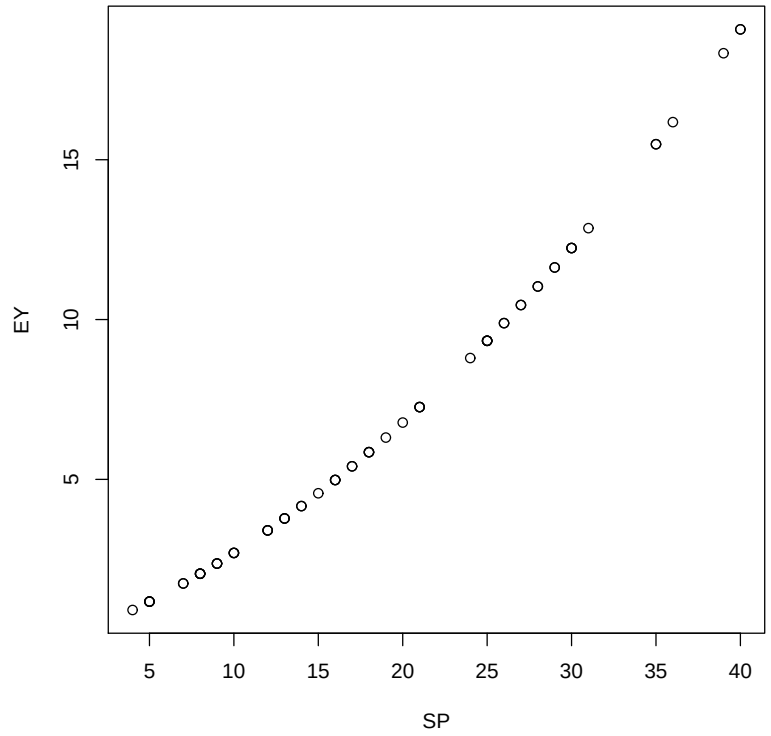
$$\begin{aligned} \tilde{\mathbf{r}} &= \tilde{\mathbf{Y}} - \widehat{\tilde{\mathbf{Y}}} \\ &= \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} \\ &= \mathbf{K}^{-1}\mathbf{Y} - \mathbf{K}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} \\ &= \mathbf{K}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \end{aligned}$$

Volviendo al ejemplo de velocidad

Si ajustamos nuevamente los datos usando pesos.

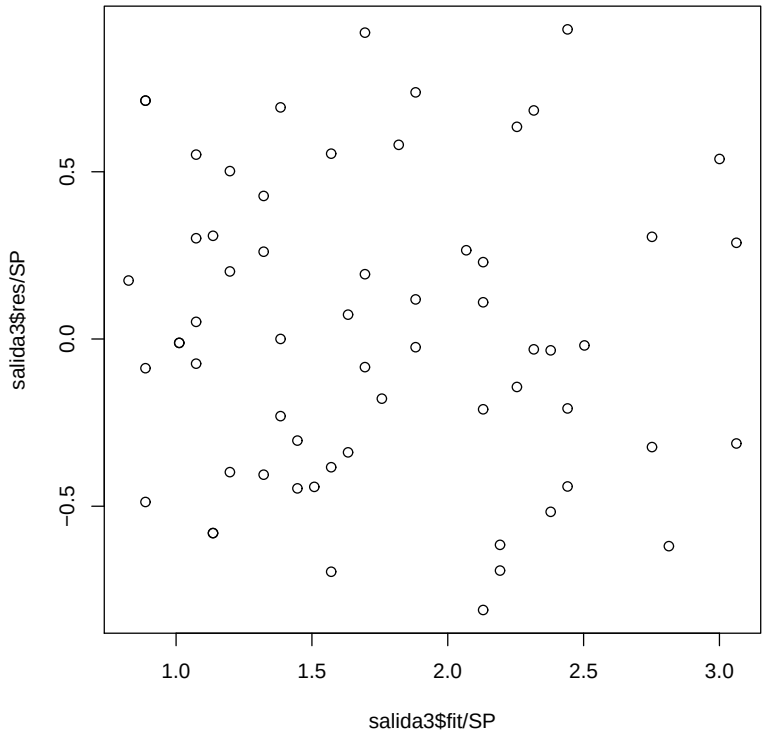
```
speed<-read.table("C:/Users/Ana/ModeloLineal/datos/Speed.txt", header=T)
attach(speed)
```

```
plot(SP,D)

SP2=SP*SP
salida<- lm(D~SP+SP2-1)

summary(salida)

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
SP  0.576599   0.200804   2.871  0.00564 **
SP2 0.062145   0.006904   9.001 9.83e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.852 on 60 degrees of freedom
Multiple R-squared: 0.9644,     Adjusted R-squared: 0.9632
F-statistic: 813.5 on 2 and 60 DF,  p-value: < 2.2e-16

plot(salida$fit,salida$res)

EY=0.200804*SP+0.006904*SP2

plot(SP,EY)
```



Como vemos en el gráfico, $E(Y_i)$ sería proporcional a SP, por lo tanto de acuerdo a lo visto $V(Y_i) \propto E^2(Y_i)$ y en consecuencia usaríamos pesos de la forma $1/SP^2$.

```
salida3<- lm(D~SP+SP2,weight=peso)
summary(salida3)

Call:
lm(formula = D ~ SP + SP2, weights = peso)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79915 -0.32983 -0.02599  0.27541  0.92972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50605    2.03544   0.740    0.462
SP           0.41968    0.34326   1.223    0.226
SP2          0.06557    0.01057   6.205  5.9e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4514 on 59 degrees of freedom
Multiple R-squared: 0.9131,     Adjusted R-squared: 0.9101
F-statistic: 309.8 on 2 and 59 DF,  p-value: < 2.2e-16

plot(salida3$fit/SP,salida3$res/SP)
```



Transformaciones Estabilizadoras de la Varianza

Podríamos preguntarnos si alguna transformación podría estabilizar la varianza.

Supongamos que f es continua, con seunda derivada f'' finita, entonces veremos que

$$V(f(Y_i)) \cong (f'(\mu_i))^2 \, \sigma_i^2(\mu_i) \quad \text{donde } \mu_i = E(Y_i)$$

Por lo tanto, para que $V(f(Y_i)) =$ cte , necesitamos que

$$(f'(\mu_i))^2 \cong \frac{c}{\sigma_i^2(\mu_i)} \quad \text{donde } c = \text{ cte.}$$

Una función f con esta propiedad se llama *función estabilizadora de la varianza.*

Por ejemplo:

- y_i v.a. de conteo tipo Poisson $\longrightarrow f(\mu_i) = \sqrt{\mu_i}$
- $y_i = \frac{m_i}{n_i}$ (v.a. tipo Binomial) $\longrightarrow f'(\mu_i) = c n_i^{1/2} / \sqrt{\mu_i(1-\mu_i)}$

  $\longrightarrow f(\mu_i) = 2c n_i^{1/2} arcsen(\sqrt{\mu_i})$

Otra forma de elegir una transformación

Cuando tenemos una sola variable independiente el diagrama de dispersión de las observaciones puede sugerirnos inmediatamente que es necesaria una transformación de los datos y cuál elegir.

Si no es así, Tukey y Mosteller (1977) sugieren la siguiente estrategia:

- Dividimos el rango de las $x's$ en tres regiones haciendo un compromiso entre el número de observaciones en cada región y un tamaño homogéneo de las mismas.
- En cada región calculamos la mediana de las $x's$ y de las correspondientes $y's$.
- Hallamos la pendiente de la recta de los dos primeros puntos y de los dos últimos.
- Si las pendientes son iguales entonces los puntos están sobre una recta. Si no, el punto del medio estará por debajo de los otros dos (convexo) o más arriba de los otros dos (cóncavo).

- Transformamos a x o a y usando el cuadro que se encuentra más abajo.

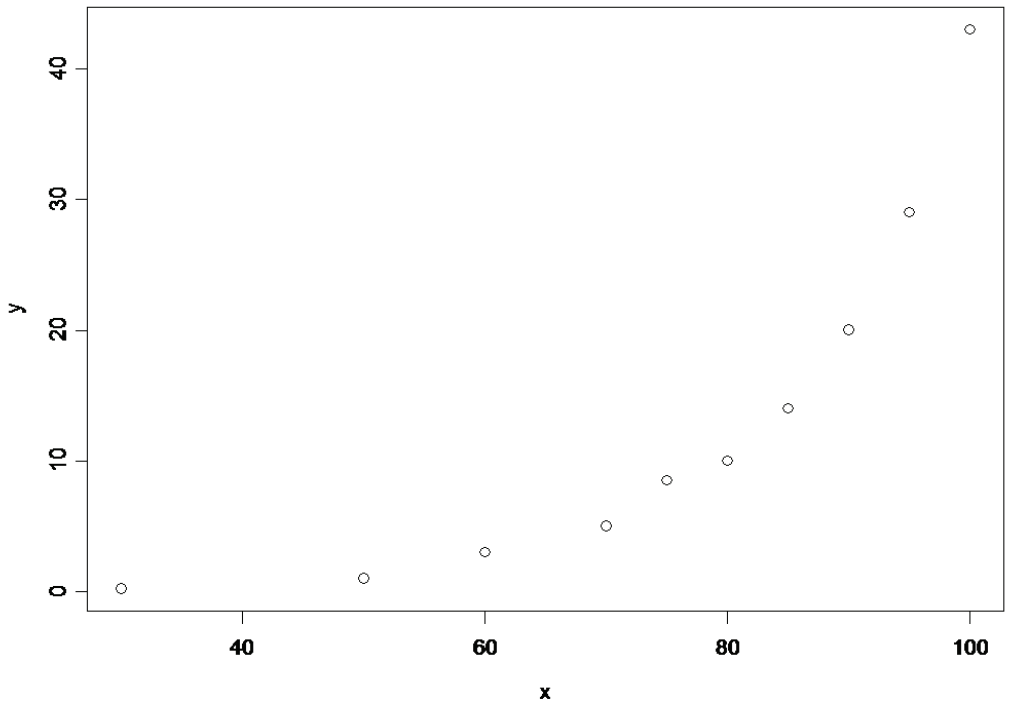
Podemos aplicar la transformación elegida a los tres puntos y verificar si las pendientes dan iguales. En este caso habríamos encontrado una transformación adecuada.

Cuadro de Transformaciones

.		.
$-1/y^2$		.
$-1/y$		x^5
$-1/y^{1/2}$		x^4
$log(y)$	$\Uparrow$	x^3
$y^{1/2}$	convexo	x^2
y	$\Longleftrightarrow$	x
y^2	$\Downarrow$	$x^{1/2}$
y^3	cóncavo	$\log(x)$
y^4		$-1/x^{1/2}$
y^5		$-1/x$
.		.

Ejemplo Stevens (1956) pidió a un conjunto de individuos que comparasen notas de varios decibeles contra un standard (80 decibeles) y que les asignaran un rango de sonoridad, donde el rango standard era 10. Obtuvo los siguientes datos

Estímulo (x)	30	50	60	70	75	80	85	90	95	100
Respuesta mediana (y)	0.2	1.0	3.0	5.0	8.5	10.0	14.0	20	29	43



Consideramos: (50,1), (77.5,9.25) y (95,29).

(50,1), (77.5,9.25) y (95,29)

pendiente $\frac{y_2-y_1}{x_2-x_1} \Rightarrow$

- entre los dos primeros es $\frac{8,25}{27,5} = ,3$
- entre los dos últimos $\frac{19,75}{17,5} = 1,13$.

Transformamos ay. Comenzando con escala descendente transformaríamos con $\sqrt{y}$. Aplicamos esta transformación a la segunda coordenada de los tres puntos y al recalcular las pendientes obtenemos

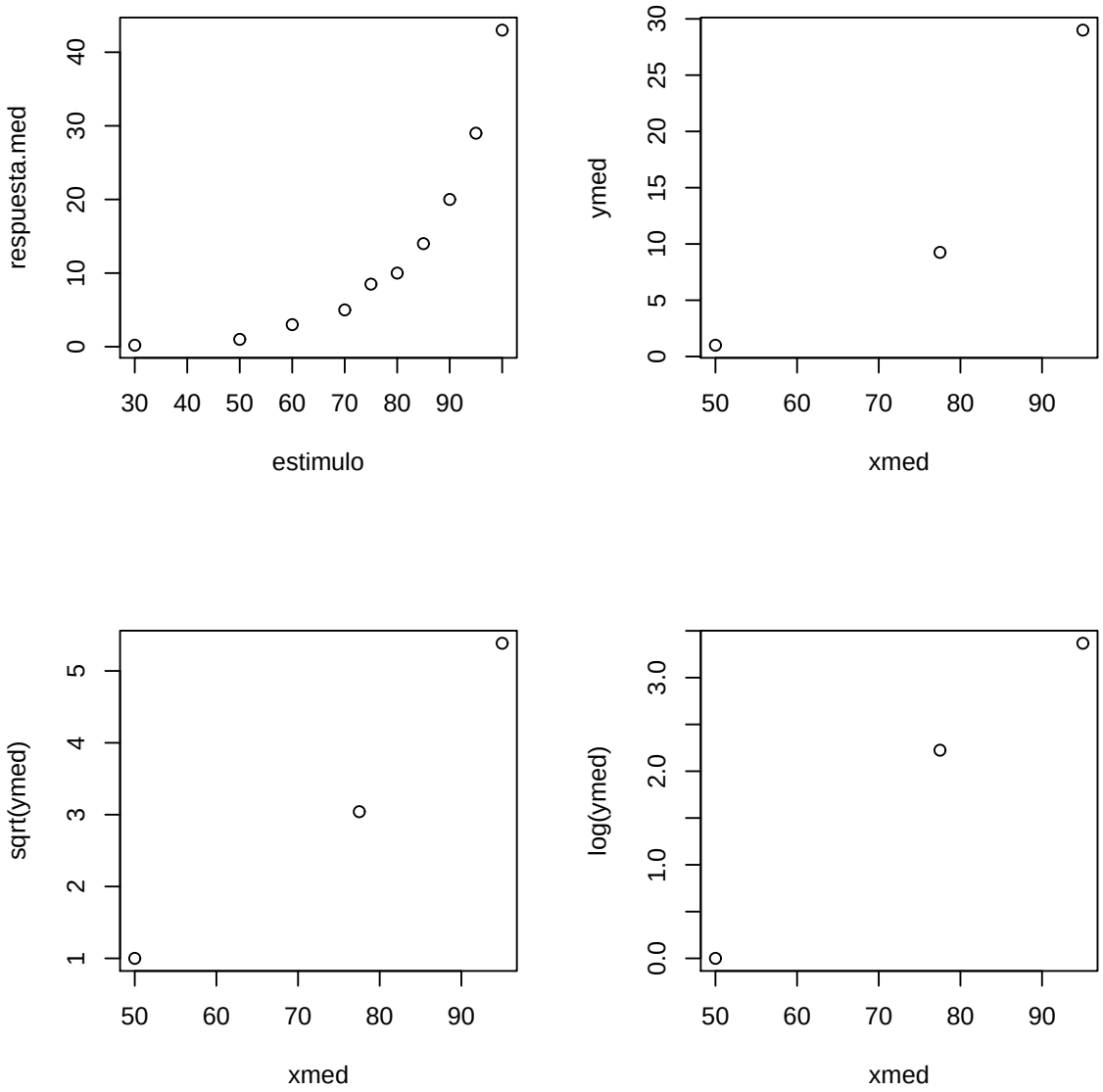
- entre los dos primeros es $\frac{2,04}{27,5} = 0,074$
- entre los dos últimos $\frac{2,35}{27,5} = 0,134$

Podríamos probar con la transformación que sigue en la escala descendente, es decir $-1/y^{1/2}$. Las nuevas pendientes son: 0.025 y 0.0082
$\Rightarrow$ estaríamos empeorando.

De acuerdo con este análisis, nos quedaríamos con la transformación logaritmo.

```
## Datos de Stevens

estimulo<- c( 30,50,60,70,75,80,85,90,95,100)
respuesta.med<- c( 0.2,1.0,3.0,5.0,8.5,10.0,14.0,20,29,43)

xmed<- c(50,77.5,95)
ymed<- c(1,9.25,29)

par(mfrow=c(2,2))
plot(estimulo,respuesta.med)
plot(xmed,ymed)
plot(xmed,sqrt(ymed))
plot(xmed,log(ymed))
```

Nota ¿Qué ocurre si la variable de respuesta toma valores negativos? En ese caso se suma una constante c a Y antes de aplicarle una potencia. Dolby (1963) propuso un método gráfico para elegir la constante c.


respuesta.med
0
10
20
30
40
30
40
50
60
70
80
90
estimulo
ymed
0
5
10
15
20
25
30
50
60
70
80
90
xmed
sqrt(ymed)
1
2
3
4
5
50
60
70
80
90
xmed
log(ymed)
0.0
1.0
2.0
3.0
50
60
70
80
90
xmed

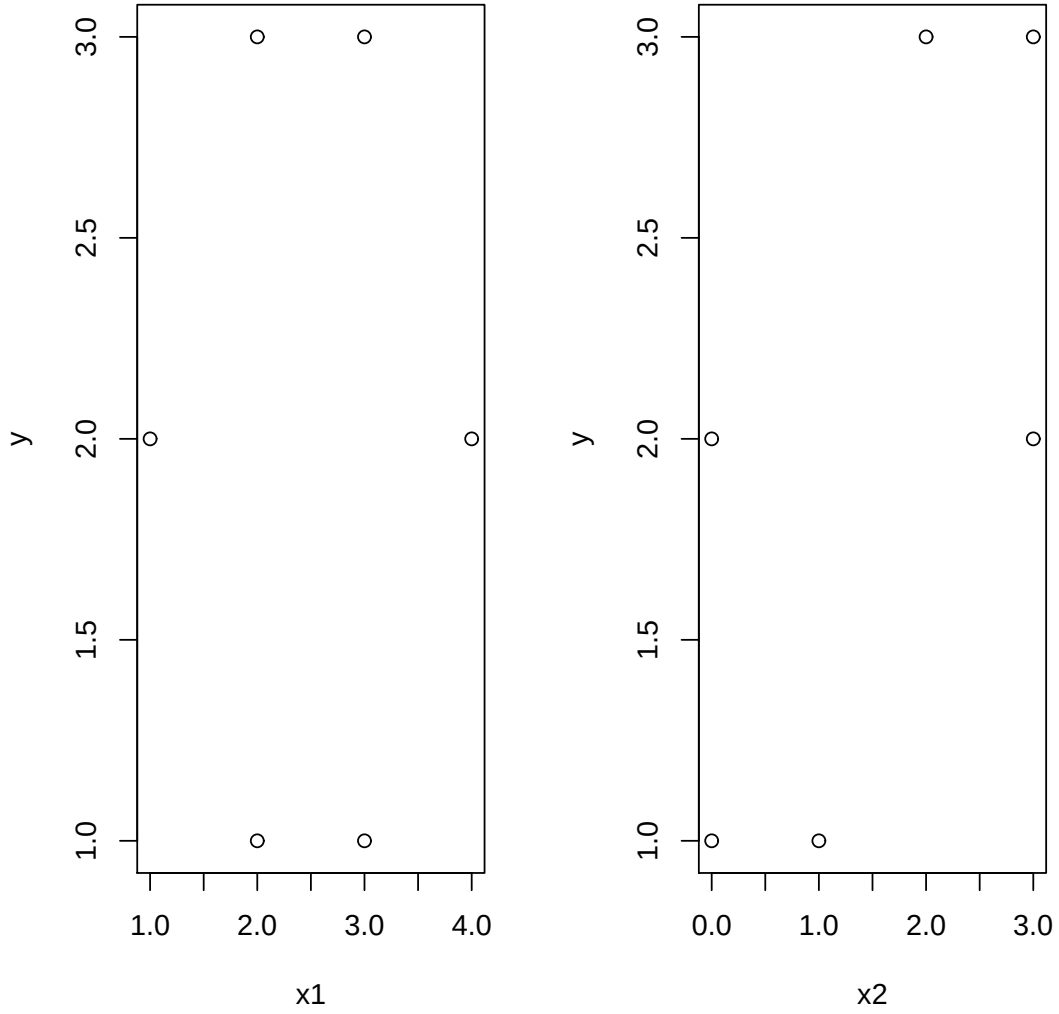
Cuando hay 2 o más variables explicativas

El principal problema que se nos presenta en este caso es los gráficos de Y vs. cada una de las covariables X_j pueden ser no informativos.

Así por ejemplo , si consideramos los puntos

```
y<- c(2,1,3,1,3,2)
x1<- c(1,2,2,3,3,4)
x2<- c(3,1,3,0,2,0)

par(mfrow=c(1,2))
plot(x1,y)
plot(x2,y)
```

el gráfico que obtenemos es:


y
3.0
2.5
2.0
1.5
1.0
1.0
2.0
3.0
4.0
x1
y
3.0
2.5
2.0
1.5
1.0
0.0
1.0
2.0
3.0
x2

No presenta ninguna estructura cuando graficamos Y vs. X_1 (aparece un hexágono), aún cuando lo puntos yacen sobre el plano: $Y = -3 + X_1 + X_2$

Wood (1973) propuso el siguiente método.

Supongamos que ajustamos el modelo

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

y obtenemos los estimadores $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$, entonces los residuos serán:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik})$$

Luego:

$$Y_i - \hat{\beta}_0 - \sum_{j \neq m}^{k} \hat{\beta}_j x_{ij} = e_i + \hat{\beta}_m x_{im} \quad \text{residuos parciales}$$

Por lo tanto, $e_i + \hat{\beta}_m x_{im}$ es Y_i sin el efecto de todas las otras covariables, de manera que graficando $e_i + \hat{\beta}_m x_{im}$ vs. X_{im} obtenemos un gráfico en el que el efecto de las otras covariables ha sido removido.

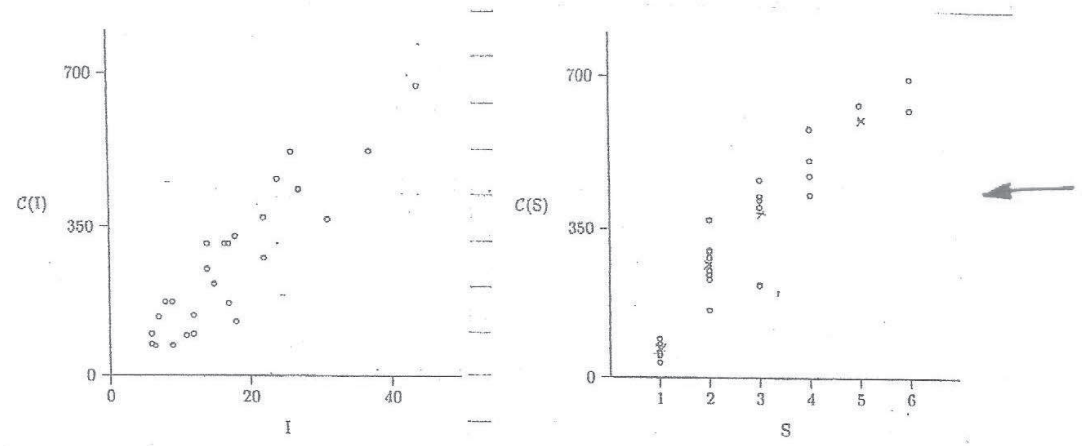
En estos gráficos de residuos parciales podemos aplicar las técnicas para una sola variable independiente.

Una desventaja de este método es que si dos covariables estuvieran muy correlacionadas, podría ocurrir que los $\beta's$ no estuvieran bien estimados y por lo tanto los plots de residuos parciales podrían ser confusos.

Ejemplo

R	I	S	R	I	S	R	I	S
99	6.5	1	670	26.0	2	400	18.0	3
125	11.0	1	820	44.0	2	640	22.0	3
200	17.0	1	325	8.0	2	468	6.0	4
550	37.0	1	366	15.0	2	668	22.0	4
100	9.0	1	325	9.0	2	850	24.0	4
250	6.0	2	411	7.0	3	825	27.0	4
400	14.0	2	580	14.0	3	950	27.0	5
475	18.0	2	580	17.0	3	1000	31.0	6
250	12.0	2	580	16.5	3	780	12.0	6

EXHIBIT 9.8: Data on Monthly Rent (R dollars), Annual Income (I × 1000 dollars) and Household Size (S).



(1,60), (2,280), (3,410), (5,580) ← 4 puntos

curva cóncava ⇒ consideramos $\sqrt{S}$. Si alinea

computamos las 3 pendientes quedan: 5.1, 4.3 y 3.

Tomando logaritmo obtenemos: 3.2, 3.2 y 3.7

⇒ $\sqrt{S}$ más adecuada.

Diagnóstico en ANOVA

En el caso de Anova 1 Factor computando los residuos r_{ij} una vez calculada la Tabla ANOVA podemos detectar:

- Heterogeneidad de varianzas
- Falta de independencia entre las observaciones
- Presencia de outliers
- Omisión de alguna variable importante
- Falta de Normalidad

Podemos investigar la distribución de los residuos a través de diagramas de tallo-hoja, histogramas, box-plots. Podemos detectar asimetría, presencia de outliers, etc.

Si el tamaño de n_i es razonable, es aconsejable realizarlos para cada nivel del factor. Recordemos que como en regresión, los residuales no son independientes. En general, esta dependencia suele ser despreciable.

Si el tamaño de n_i es razonable, podemos chequear el supuesto de normalidad realizando qq-plots y aplicando el test de Shapiro-Wilk para la observaciones originales en cada nivel. Si no es así, los haremos para todos los residuos juntos.

Si detectásemos residuos grandes o alejados del grueso de los residuos deberíamos estudiar cuidadosamente la situación.

Otros gráficos

- Diagrama de puntos: se construyen graficando los residuos (o las observaciones originales) de cada nivel del factor en paralelo y nos darán una idea de si el supuesto de homogeneidad de varianzas entre los niveles es razonable o no.
- Valores predichos $\overline{Y}_i$ vs. Residuos: en este gráfico podemos apreciar la bondad del ajuste del modelo y las varianzas de los residuos.
- Gráfico de residuos vs. secuencia temporal: si se tiene registrado el orden en que fueron tomadas las observaciones es aconsejable hacer este gráfico con el fin de detectar alguna tendencia.

- Gráfico de residuales vs. alguna variable de interés: si se midió alguna otra variable (edad, peso, etc) puede ser útil graficar los residuos vs. esta variable. Esto puede contribuir a:

  1. la comprensión del problema
  2. sugerir variables a controlar en una nueva experiencia
  3. ayudar a detectar un factor confundido si no se aleatorizó correctamente.

Para detectar heterogeneidad de varianzas en este modelo existen varios tests específicos cuando la distribución de los datos es normal. Veremos una opción, que es la del Test de Levene, que es válida en un contexto más general.

Supongamos que tenemos un **Anova 1 Factor** en el que comparamos k tratamientos.

Las hipótesis a testear son:

Ho: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ vs. $H_1: \sigma_i^2 \neq \sigma_j^2$ para algún par de índices $i \neq j$

Test de Levene Modificado

El test de Levene modificado testea la igualdad de varianzas. Puede calcularse fácilmente transformando la variable de respuesta y calculando una nueva Tabla de ANOVA para las variables transformadas.

Los pasos a seguir son:

0) Computamos la mediana de la i-ésima casilla $\tilde{Y}_i = \underset{j}{med}(Y_{ij})$

1) Calculamos las variables transformadas: $W_{ij} = \left| Y_{ij} - \tilde{Y}_i \right|$

2) Calculamos la Tabla de ANOVA para W_{ij}

3) Rechazamos la hipótesis de igualdad de varianzas si el estadístico *F* del paso anterior es grande.

Entre las propuestas para testear homogeneidad de varianzas, este test figura entre los más potentes y resistentes a la violación del supuesto de normalidad.

Si se rechaza la hipótesis de igualdad de varianzas, tenemos algunas alternativas.

Si la varianza no es constante, pero se sustenta el supuesto de normalidad, es recomendable usar mínimos cuadrados ponderados o pesados.

Muchas veces la heterogeneidad de varianzas está acompañada por la no normalidad de las observaciones. En este caso, la transformación de la variable de respuesta suele ser un remedio.

Con frecuencia, la misma transformación que estabiliza las varianzas también corrige la falta de normalidad de los datos.

Si esto no se puede lograr, puede combinarse una transformación estabilizadora de varianzas con una alternativa al test de F que sea no paramétrica.

Una posibilidad para encontrar la función transformadora es realizar un gráfico de $\bar{Y}_i$ vs. S_i para visualizar qué tipo de relación tienen.

Veamos un ejemplo.

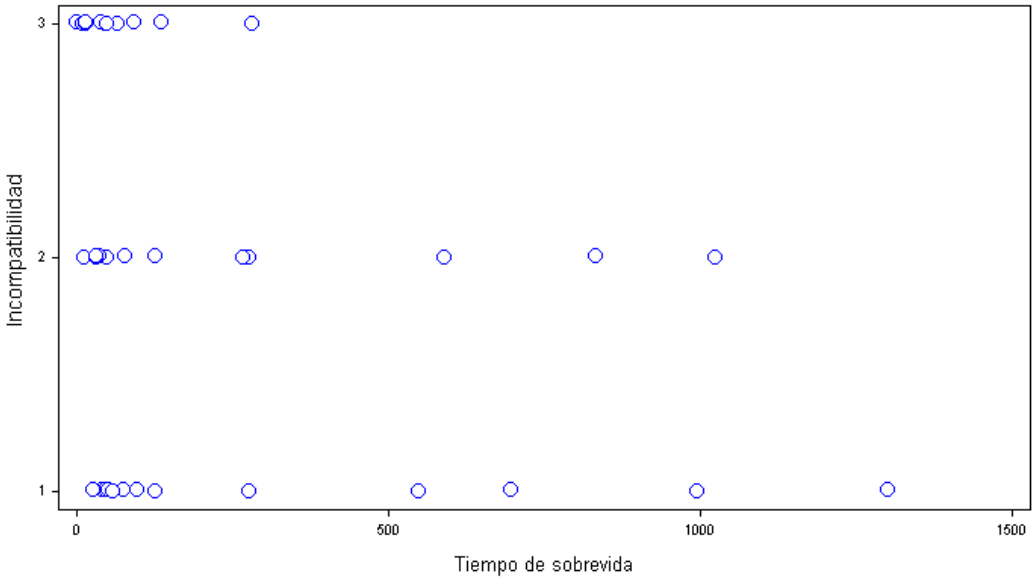
Transplante de Corazón

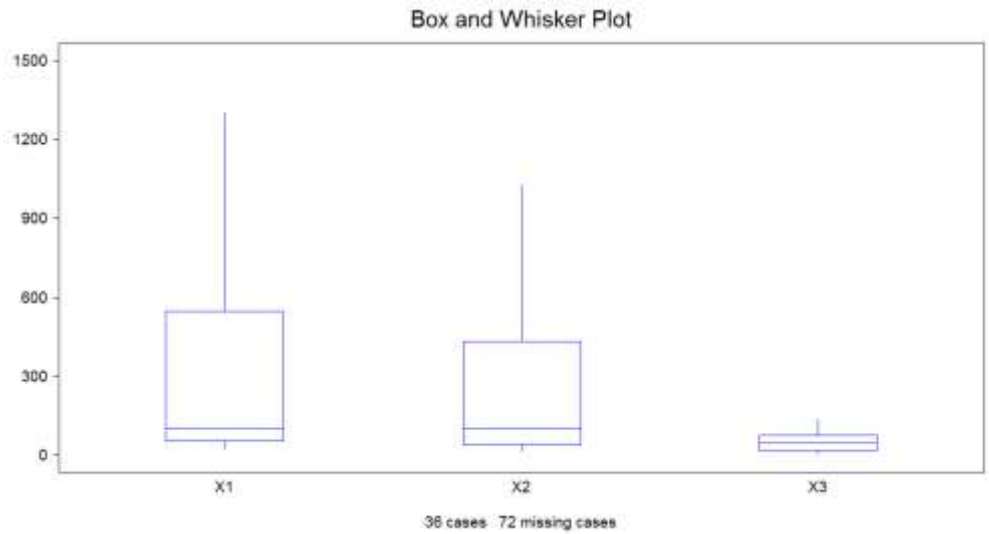
En los transplantes de corazón la similitud entre el tipo de tejido del donante y del receptor es importante, pues grandes diferencias aumentan la probabilidad de rechazo del corazón transplantado. Los datos que analizaremos a continuación corresponden al tiempo de supervivencia de 36 pacientes transplantados. Los datos fueron agrupados en tres categorías de acuerdo con el grado de incompatibilidad entre el tejido del donante y del receptor (baja=1, media=2 y alta=3). Los investigadores quieren determinar si el tiempo medio de sobrevida depende del grado de incompatibilidad.
Las hipótesis a testear son:

$H_o: \mu_1 = \mu_2 = \mu_3$ vs. H_1: no todas las μ_i son iguales

Diagrama de Puntos

El diagrama de puntos sugiere que el tiempo de sobrevida puede disminuir cuando crece la incompatibilidad.


Incompatibilidad
3
2
1
0
500
1000
1500
Tiempo de sobrevida



Inicialmente se realizó un ANOVA y se calcularon los residuos con fines de diagnóstico. A continuación ofrecemos la salida y algunos gráficos.

SOURCE	DF	SS	MS	F	P
BETWEEN	2	455385	227693	2.13	0.1351
WITHIN	33	3530419	106982		
TOTAL	35	398580			

Incompat.	MEAN	SAMPLE SIZE	GROUP STD DEV
1	334.92	13	421.99
2	281.08	12	347.32
3	69.818	11	81.607
TOTAL	235.97	36	327.08

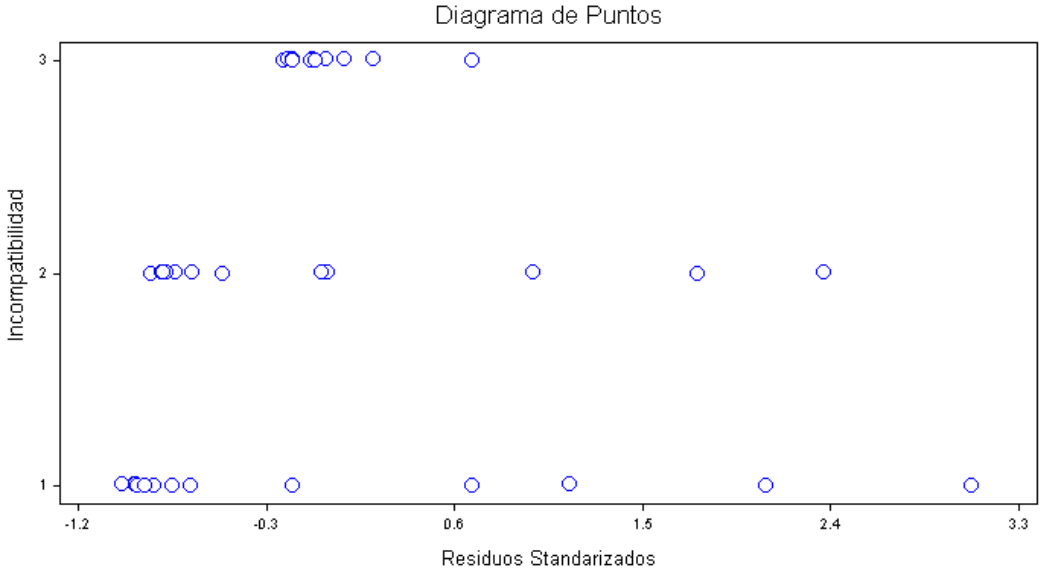
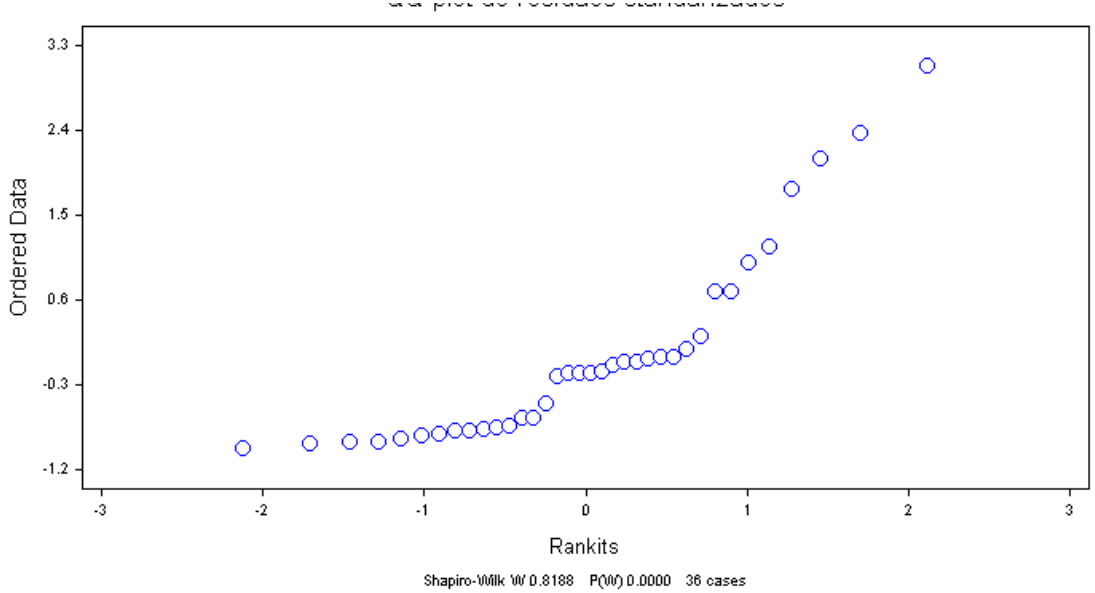


Diagrama de Puntos
Incompatibilidad
3
2
1
-1.2
-0.3
0.6
1.5
2.4
3.3
Residuos Standarizados

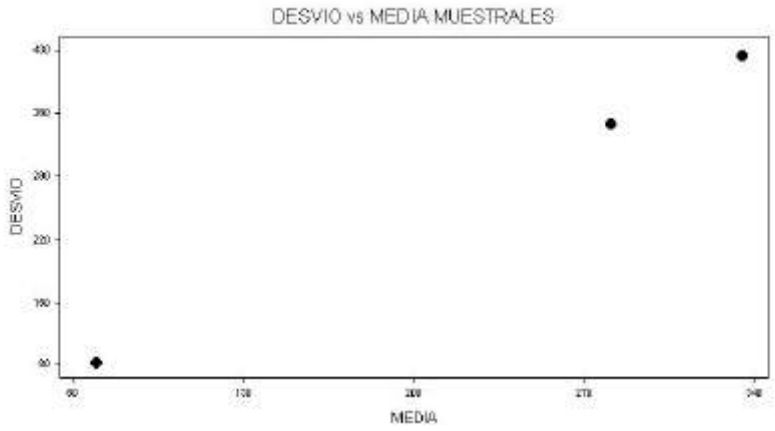

Ordered Data
3.3
2.4
1.5
0.6
-0.3
-1.2
-3
-2
-1
0
1
2
3
Rankits
Shapiro-Wilk W 0.8188 P(W) 0.0000 36 cases

El diagrama de puntos de los residuos standarizados sugiere que la distribución de los residuos es asimétrica a derecha y que la varianza de los residuos podría ser menor cuando hay una alta incompatibilidad.

El test de Levene modificado fue aplicado obteniéndose un p-valor igual a 0.1504.

Por otro lado, el qq-plot de todos los residuos standarizados revela cierta asimetría a derecha y el test de Shapiro-Wilk tiene un p-valor menor que 0.0001.

Si realizamos un scatter plot de $\overline{Y_i}$ vs S_i parece haber una relación lineal entre ambos.



Trabajamos con la nueva variable $Y' = \log(Y)$ y realizamos el análisis de la varianza para ella.

Tabla de ANOVA

SOURCE	DF	SS	MS	F	P
BETWEEN	2	12.9734	6.48670	3.57	0.0394
WITHIN	33	59.9250	1.81591		
TOTAL	35	72.8984			

IND	MEAN	SAMPLE SIZE	GROUP STD DEV
1	5.0174	13	1.3338
2	4.8098	12	1.4213
3	3.6281	11	1.2790
TOTAL	4.5237	36	1.3476

El p-valor del test de Levene modificado para la variable transformada es 0.7282. El diagrama de puntos y el qq-plot de los residuos standarizados (p-valor del test de Shapiro -Wilk = 0.1463) también sugieren que la transformación logarítmica es apropiada.

En la tabla de ANOVA vemos que el estadístico F = 3.57 con un p-valor = 0.0394. Para un nivel α=0.05 concluiríamos que la media del logaritmo del tiempo de sobrevida de los transplantados depende del grado de incompatibilidad del tejido entre donante y receptor.

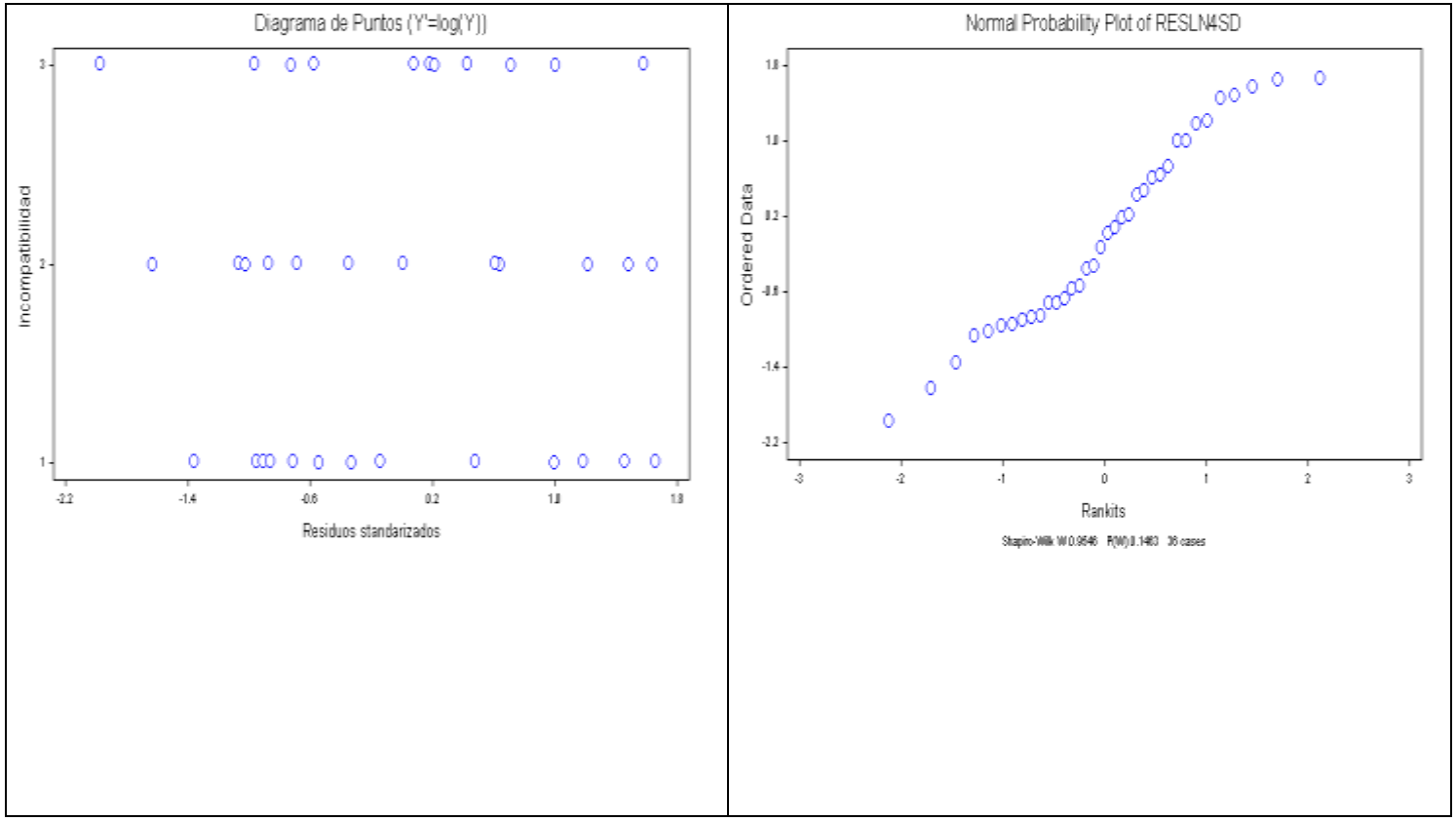


Diagrama de Puntos (Y'=log(Y))
Incompatibilidad
Residuos standarizados
Normal Probability Plot of RESLN4SD
Ordered Data
Rankits
Shapiro-Wilk W 0.9646 P(W) 0.1463 36 cases