Análisis de regresión

Capítulo 3: Supuestos y comprobación de la adecuación del modelo

Mg. Sc. J. Eduardo Gamboa U.

Introducción

Hemos visto que podemos explicar la relación de dependencia en un conjunto de variables a través de una línea de regresión.
Esa línea puede tener dos fines:

- (a) Explicativo
- (b) Predictivo

La validación de procedimientos inferenciales se realiza mediante el análisis de los errores, representados en la muestra por los **residuales**.

Un residual es la realización del error del modelo de regresión lineal:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

donde:

- y_i es la observación de la variable respuesta
- $\hat{y}_i$ es el valor estimado por el modelo

El diagnóstico del modelo se realiza a través de los residuales con dos fines:

1. Verificación de supuestos
2. Análisis de outliers y valores influyentes

Ajuste del modelo en R

```
library(readxl)
library(dplyr)
library(broom)
datos <- read_excel("U3_datos_1.xlsx")
modelo <- lm(Sueldo ~ ., data = datos)
summary(modelo)
```

```
Call:
lm(formula = Sueldo ~ ., data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81981 -0.56106 -0.04567  0.62732  2.15948

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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.135 on 19 degrees of freedom
Multiple R-squared:  0.7626,    Adjusted R-squared:  0.7126
F-statistic: 15.25 on 4 and 19 DF,  p-value: 9.639e-06
```

```
modelo |> augment()
```

```
# A tibble: 24 x 11
   Sueldo Educacion Sexo   Edad    X4 .fitted .resid   .hat .sigma .cooksd
    <dbl>     <dbl> <chr> <dbl> <dbl>   <dbl>  <dbl> <dbl>  <dbl>   <dbl>
 1   4.22      12   M        31    10    4.76 -0.545 0.205  1.16  0.0150
 2   5.89      18   M        23    -3    5.46  0.431 0.417  1.16  0.0354
 3   9.88      20   M        50    -6    8.73  1.15  0.217  1.13  0.0720
 4   2.35       7.5 F        19    -4    2.62 -0.266 0.235  1.16  0.00443
 5   7         20   F        39    -2    6.58  0.420 0.123  1.16  0.00437
 6   1.25       9   M        18     7    3.07 -1.82  0.255  1.05  0.236
 7   6.78       3   M        39    -7    4.96  1.82  0.536  0.981 1.28
 8   5.19      15   F        32     8    4.61  0.581 0.144  1.16  0.0103
 9   8.16      21   F        35     5    6.00  2.16  0.162  1.02  0.167
10   6.1       18   F        34     7    5.33  0.766 0.131  1.15  0.0158
# i 14 more rows
# i 1 more variable: .std.resid <dbl>
```

Obtención de residuales

```
residuales <- modelo |> resid() |> as.vector()
residuales
```

```
 [1] -0.5446927  0.4307501  1.1457749 -0.2660204  0.4202928 -1.8198056
 [7]  1.8194646  0.5811903  2.1594785  0.7657081  1.0019604  0.1048631
[13] -0.9479326 -0.2384626 -1.7944816 -0.1961957 -0.4135295 -0.4165775
[19]  0.1369553 -0.6101662  0.2032022 -1.1690284 -1.3551731  1.0024253
```

Obtención de residuales

```
residuales <- modelo |> residuals() |> as.vector()
residuales
```

```
 [1] -0.5446927  0.4307501  1.1457749 -0.2660204  0.4202928 -1.8198056
 [7]  1.8194646  0.5811903  2.1594785  0.7657081  1.0019604  0.1048631
[13] -0.9479326 -0.2384626 -1.7944816 -0.1961957 -0.4135295 -0.4165775
[19]  0.1369553 -0.6101662  0.2032022 -1.1690284 -1.3551731  1.0024253
```

Obtención de residuales

```
residuales <- modelo$residuals |> as.vector()
residuales
```

```
 [1] -0.5446927  0.4307501  1.1457749 -0.2660204  0.4202928 -1.8198056
 [7]  1.8194646  0.5811903  2.1594785  0.7657081  1.0019604  0.1048631
[13] -0.9479326 -0.2384626 -1.7944816 -0.1961957 -0.4135295 -0.4165775
[19]  0.1369553 -0.6101662  0.2032022 -1.1690284 -1.3551731  1.0024253
```

Obtención de residuales

```
residuales <- modelo |> augment() |> pull(.resid)
residuales
```

```
 [1] -0.5446927  0.4307501  1.1457749 -0.2660204  0.4202928 -1.8198056
 [7]  1.8194646  0.5811903  2.1594785  0.7657081  1.0019604  0.1048631
[13] -0.9479326 -0.2384626 -1.7944816 -0.1961957 -0.4135295 -0.4165775
[19]  0.1369553 -0.6101662  0.2032022 -1.1690284 -1.3551731  1.0024253
```

Normalidad de errores

Supuesto

Los errores se distribuyen normalmente

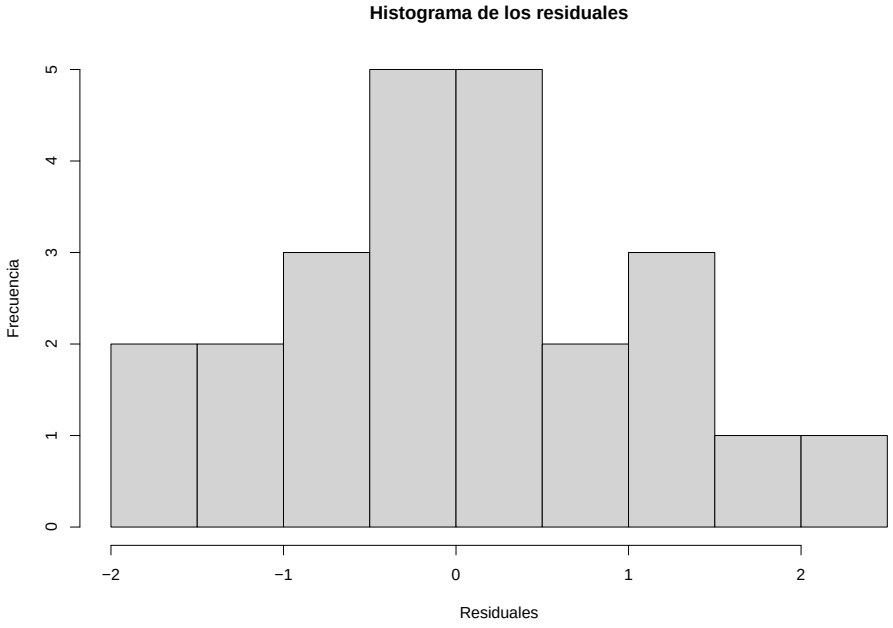
$$\varepsilon_i \sim N(0, \sigma^2)$$

El modelo de regresión lineal es robusto a la falta de cumplimiento de este supuesto si el tamaño de muestra es grande.

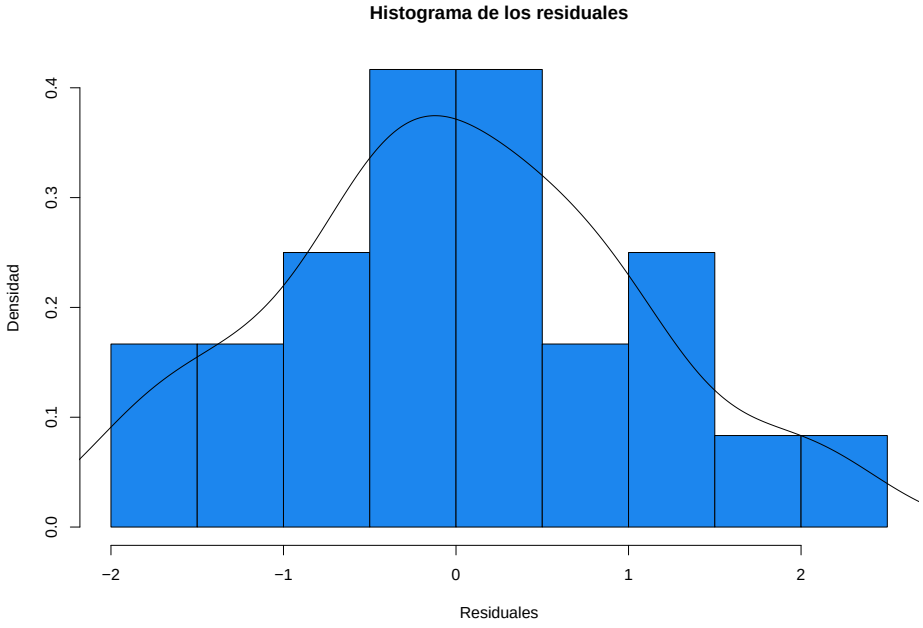
Medios de verificación

Histograma

```
residuales |> hist(main="Histograma de los residuales",
                   xlab="Residuales",
                   ylab="Frecuencia")
```


Histograma de los residuales
Frecuencia
0
1
2
3
4
5
−2
−1
0
1
2
Residuales

```r
residuales |> hist(freq=FALSE,
                   main="Histograma de los residuales",
                   xlab="Residuales",
                   ylab="Densidad",
                   col ="dodgerblue2")
residuales |> density() |> lines()
```


Histograma de los residuales
Densidad
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
Residuales

```r
library(ggplot2)
data.frame(residuales) |>
  ggplot(aes(x=residuales,))+
  geom_histogram(aes(y =..density..),
                 bins = round(1+3.3*log10(nrow(datos))),
                 fill = "dodgerblue2",
                 alpha = 0.6)+
  geom_density(size = 1.5)+
  labs(x="Residuales",y="Densidad")+
  theme_minimal()
```

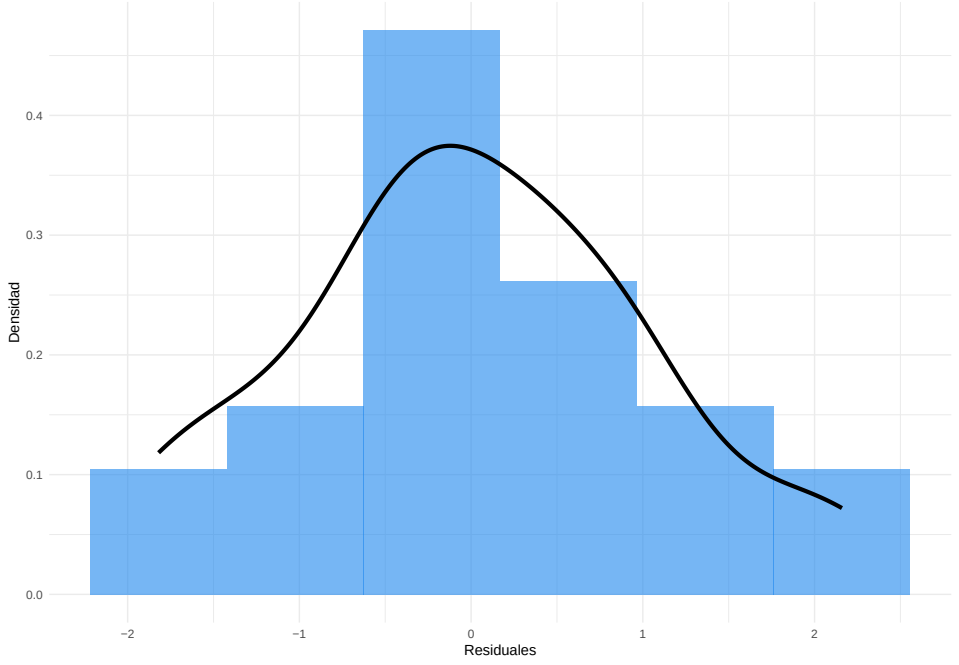
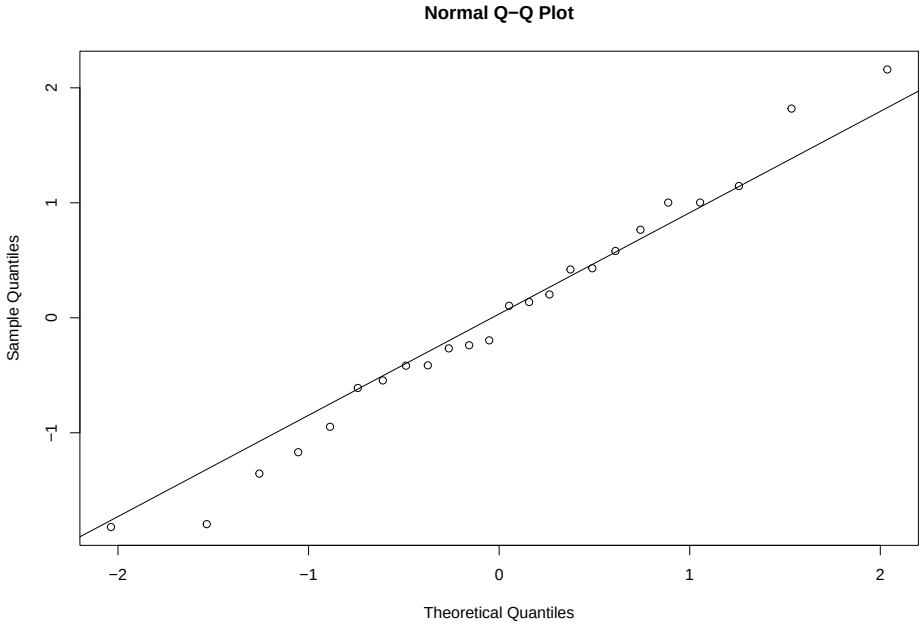

Densidad
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
Residuales

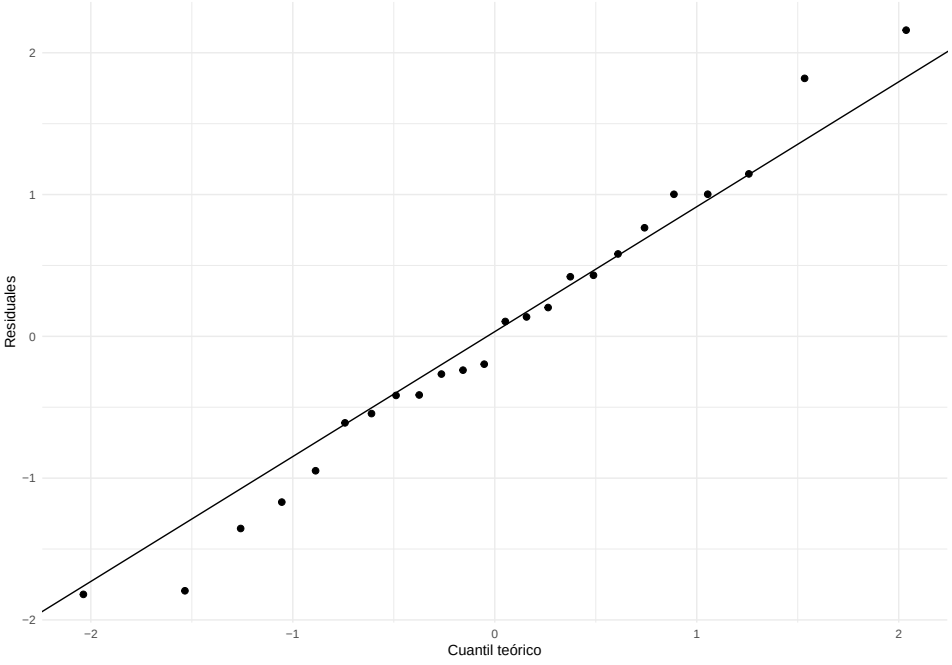
Gráfico de probabilidad normal

Residuales (estudentizados) ordenados t(i) versus el cuantil teórico, para i $= 1, \ldots, n$. Se sugiere que n > 20 (Peck, Vining y Montgomery, 2012). Permite detectar outlier, asimetría y curtosis

```
qqnorm(residuales); qqline(residuales)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−1
0
1
2

```r
data.frame(residuales) |>
  ggplot(aes(sample=residuales))+
  stat_qq(size = 2) +
  stat_qq_line(distribution = stats::qnorm)+
  labs(x = "Cuantil teórico",
       y = "Residuales")+
  theme_minimal()
```


Residuales
Cuantil teórico
2
1
0
−1
−2
−2
−1
0
1
2

Coeficiente de asimetría y curtosis:

Deben ser igual a 0 y 3, respectivamente

```
library(moments)
residuales |> skewness()
```

```
[1] 0.1257975
```

```
residuales |> agostino.test()
```

```
    D'Agostino skewness test

data:  residuales
skew = 0.1258, z = 0.3019, p-value = 0.7627
alternative hypothesis: data have a skewness
```

```r
residuales |> kurtosis()
```

```
[1] 2.592246
```

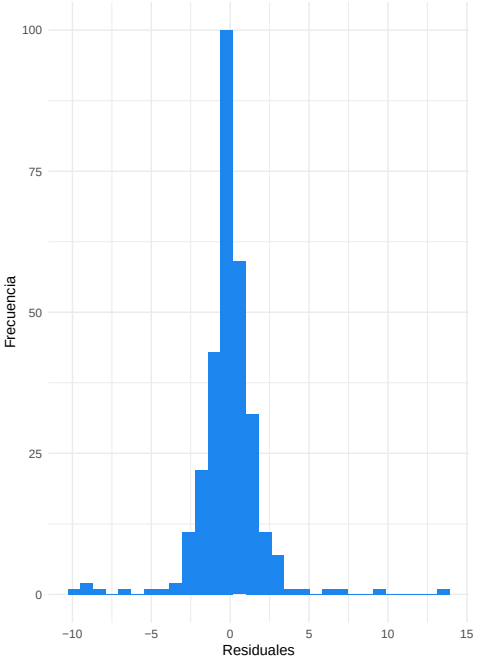
```r
residuales |> anscombe.test()
```

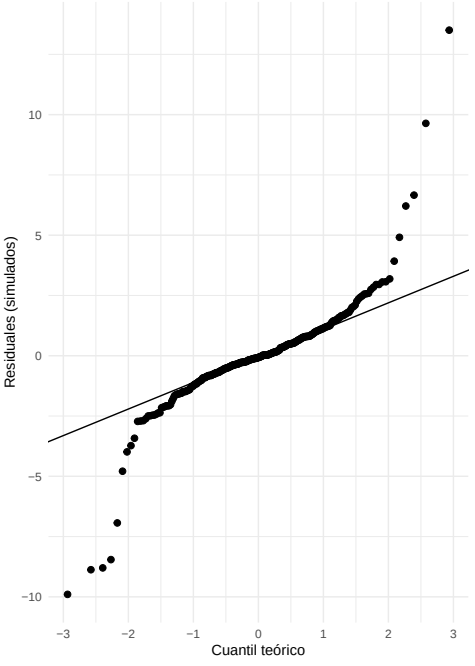
```
    Anscombe-Glynn kurtosis test

data:  residuales
kurt = 2.592246, z = -0.028832, p-value = 0.977
alternative hypothesis: kurtosis is not equal to 3
```

¿Qué sucede con los gráficos de probabilidad normal con datos asimétricos o que tienen colas pesadas?

Simulamos a partir de una distribución t (colas pesadas, especialmente cuando los grados de libertad son pequeños

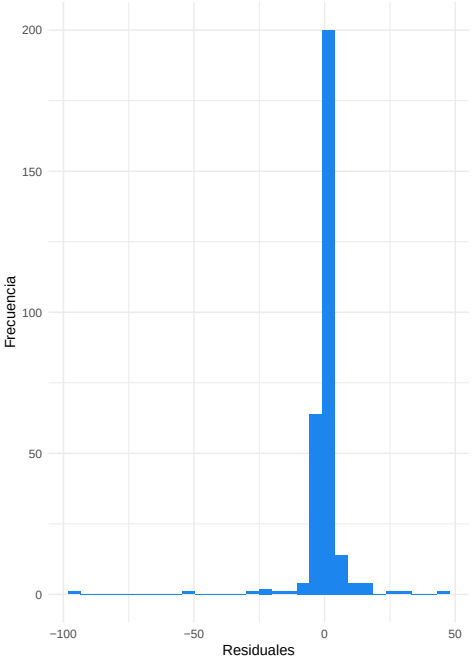
```
set.seed(4)
y = rt(300,2)
g1 = data.frame(y) |>
  ggplot(aes(x=y,))+
  geom_histogram(fill = "dodgerblue2")+
  labs(x="Residuales",y="Frecuencia")+
  theme_minimal()
g2 = data.frame(y=y) |>
  ggplot(aes(sample=y))+
  stat_qq(size = 2) +
  stat_qq_line(distribution = stats::qnorm)+
  labs(x = "Cuantil teórico",
       y = "Residuales (simulados)")+
  theme_minimal()
library(gridExtra)
grid.arrange(g1,g2,ncol=2)
```

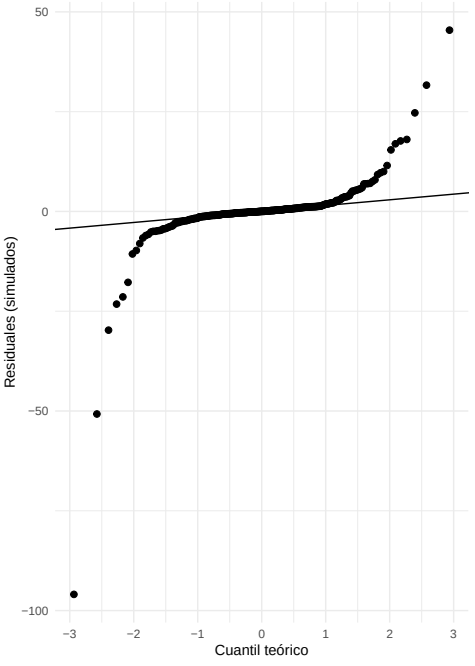

100
75
50
25
0
Frecuencia
−10
−5
0
5
10
15
Residuales


10
5
0
−5
−10
Residuales (simulados)
−3
−2
−1
0
1
2
3
Cuantil teórico

La distribución Cauchy también tiene colas pesadas (y alta curtosis)

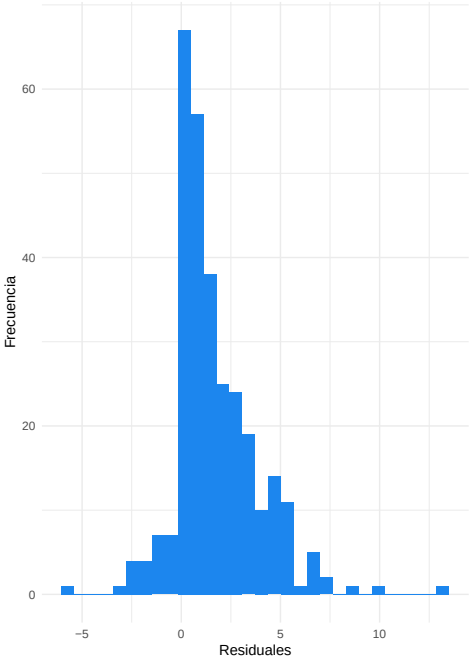
```
set.seed(4)
y = rcauchy(300)
g3 = data.frame(y) |>
  ggplot(aes(x=y,))+
  geom_histogram(fill = "dodgerblue2")+
  labs(x="Residuales",y="Frecuencia")+
  theme_minimal()
g4 = data.frame(y=y) |>
  ggplot(aes(sample=y))+
  stat_qq(size = 2) +
  stat_qq_line(distribution = stats::qnorm)+
  labs(x = "Cuantil teórico",
       y = "Residuales (simulados)")+
  theme_minimal()
grid.arrange(g3,g4,ncol=2)
```

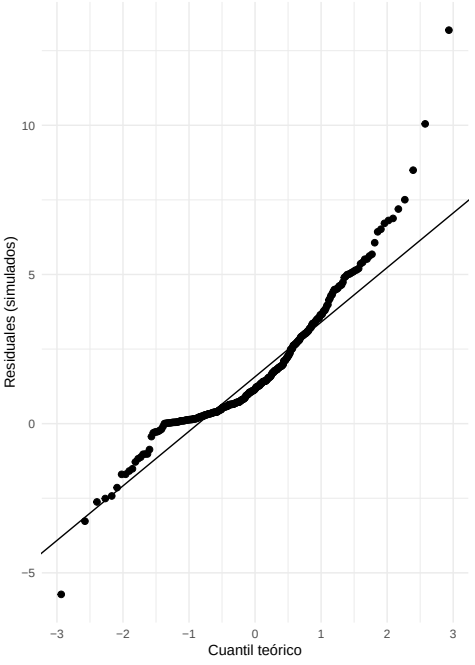

200
150
100
50
0
Frecuencia
−100
−50
0
50
Residuales


50
0
−50
−100
Residuales (simulados)
−3
−2
−1
0
1
2
3
Cuantil teórico

Simulamos distribuciones asimétricas usando la distribución Chi cuadrado

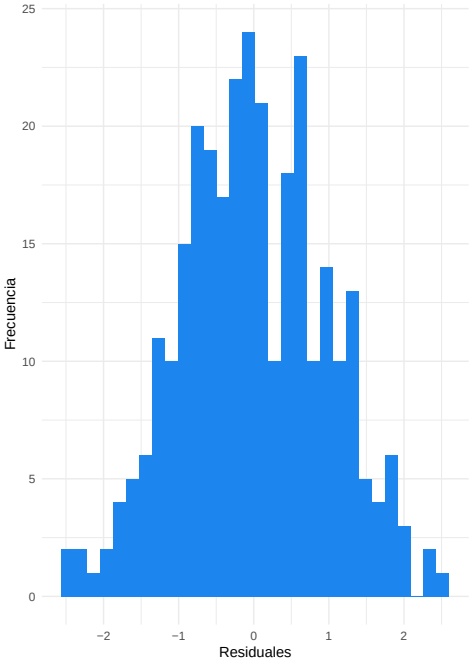
```
set.seed(4)
y = c(rchisq(275,2),-rchisq(25,2))
g5 = data.frame(y) |>
  ggplot(aes(x=y,))+
  geom_histogram(fill = "dodgerblue2")+
  labs(x="Residuales",y="Frecuencia")+
  theme_minimal()
g6 = data.frame(y=y) |>
ggplot(aes(sample=y))+
  stat_qq(size = 2) +
  stat_qq_line(distribution = stats::qnorm)+
  labs(x = "Cuantil teórico",
       y = "Residuales (simulados)")+
theme_minimal()
grid.arrange(g5,g6,ncol=2)
```

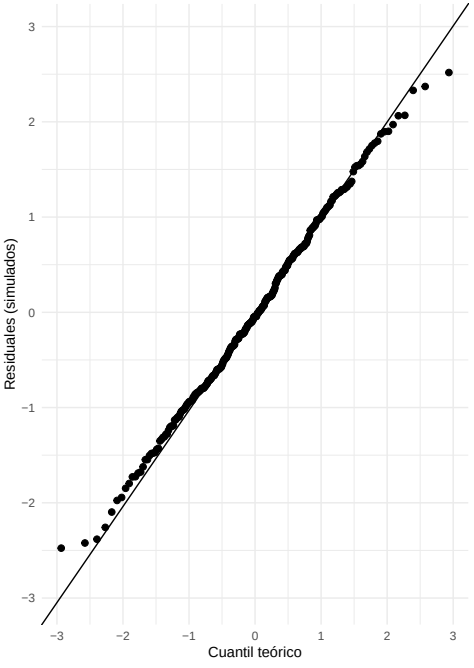

Frecuencia
60
40
20
0
−5
0
5
10
Residuales


Residuales (simulados)
10
5
0
−5
−3
−2
−1
0
1
2
3
Cuantil teórico

Entonces, ¿cómo debería ser el comportamiento? Como el que ya se vio en los datos. Simulando

```
set.seed(4)
y = rnorm(300)
g9 = data.frame(y) |>
  ggplot(aes(x=y,))+
  geom_histogram(fill = "dodgerblue2")+
  labs(x="Residuales",y="Frecuencia")+
  theme_minimal()
g10 = data.frame(y=y) |>
  ggplot(aes(sample=y))+
  stat_qq(size = 2) +
  stat_qq_line(distribution = stats::qnorm)+
  labs(x = "Cuantil teórico",
       y = "Residuales (simulados)")+
  theme_minimal()
grid.arrange(g9,g10,ncol=2)
```


Frecuencia
25
20
15
10
5
0
−2
−1
0
1
2
Residuales


Residuales (simulados)
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Cuantil teórico

Pruebas de normalidad:

Shapiro Wilk, Anderson Darling, Kolmogorov Smirnov (Lilliefors). La primera de ellas es la más potente, sin embargo en todos los casos disminuye la potencia de prueba conforme disminuye el tamaño de muestra (Kyun & Hong, 2019)

```
library(nortest)
residuales |> shapiro.test()
```

```
    Shapiro-Wilk normality test

data:  residuales
W = 0.98304, p-value = 0.9446
```

```
residuales |> ad.test()
```

```
    Anderson-Darling normality test

data:  residuales
A = 0.13547, p-value = 0.9741
```

```
residuales |> lillie.test()
```

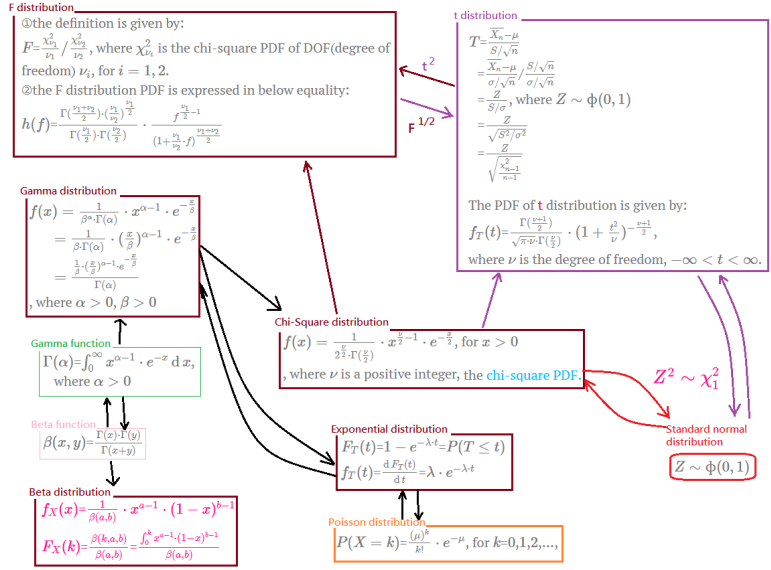
```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuales
D = 0.075434, p-value = 0.977
```

Motivos

- Errores siguen otra distribución
- Asimetría
- Colas pesadas
- Heterocedasticidad

Consecuencias

- Pruebas de hipótesis e intervalos de confianza que utilizan las distribuciones t, Chi cuadrado o F.
- Intervalos de confianza no son simétricos o son más (o menos) amplios, y podrían tener límites que no son establecidos en una distribución Normal.
- Distribución simétrica con colas pesadas: outliers tendrán mayor efecto en las estimaciones


F distribution
①the definition is given by:
$F=\frac{\chi_{\nu_1}^2}{\nu_1}/\frac{\chi_{\nu_2}^2}{\nu_2}$, where $\chi_{\nu_i}^2$ is the chi-square PDF of DOF(degree of freedom) ν_i, for $i=1,2$.
②the F distribution PDF is expressed in below equality:
$h(f)=\frac{\Gamma(\frac{\nu_1+\nu_2}{2})\cdot(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}}{\Gamma(\frac{\nu_1}{2})\cdot\Gamma(\frac{\nu_2}{2})}\cdot\frac{f^{\frac{\nu_1}{2}-1}}{(1+\frac{\nu_1}{\nu_2}\cdot f)^{\frac{\nu_1+\nu_2}{2}}}$
t distribution
$T=\frac{\overline{X_n}-\mu}{S/\sqrt{n}}$
$=\frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}/\frac{S/\sqrt{n}}{\sigma/\sqrt{n}}$
$=\frac{Z}{S/\sigma}$, where $Z\sim\phi(0,1)$
$=\frac{Z}{\sqrt{S^2/\sigma^2}}$
$=\frac{Z}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}}$
The PDF of t distribution is given by:
$f_T(t)=\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\cdot\nu}\cdot\Gamma(\frac{\nu}{2})}\cdot(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}$,
where ν is the degree of freedom, $-\infty<t<\infty$.
t^2
$F^{1/2}$
Gamma distribution
$f(x)=\frac{1}{\beta^\alpha\cdot\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-\frac{x}{\beta}}$
$=\frac{1}{\beta\cdot\Gamma(\alpha)}\cdot(\frac{x}{\beta})^{\alpha-1}\cdot e^{-\frac{x}{\beta}}$
$=\frac{\frac{1}{\beta}\cdot(\frac{x}{\beta})^{\alpha-1}\cdot e^{-\frac{x}{\beta}}}{\Gamma(\alpha)}$
, where $\alpha>0,\beta>0$
Gamma function
$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}\cdot e^{-x}\,\mathrm{d}\,x$,
where $\alpha>0$
Beta function
$\beta(x,y)=\frac{\Gamma(x)\cdot\Gamma(y)}{\Gamma(x+y)}$
Beta distribution
$f_X(x)=\frac{1}{\beta(a,b)}\cdot x^{a-1}\cdot(1-x)^{b-1}$
$F_X(k)=\frac{\beta(k,a,b)}{\beta(a,b)}=\frac{\int_0^k x^{a-1}\cdot(1-x)^{b-1}}{\beta(a,b)}$
Chi-Square distribution
$f(x)=\frac{1}{2^{\frac{\nu}{2}}\cdot\Gamma(\frac{\nu}{2})}\cdot x^{\frac{\nu}{2}-1}\cdot e^{-\frac{x}{2}}$, for $x>0$
, where ν is a positive integer, the chi-square PDF.
$Z^2\sim\chi_1^2$
Standard normal distribution
$Z\sim\phi(0,1)$
Exponential distribution
$F_T(t)=1-e^{-\lambda\cdot t}=P(T\le t)$
$f_T(t)=\frac{\mathrm{d}\,F_T(t)}{\mathrm{d}\,t}=\lambda\cdot e^{-\lambda\cdot t}$
Poisson distribution
$P(X=k)=\frac{(\mu)^k}{k!}\cdot e^{-\mu}$, for k=0,1,2,...,

Acciones a tomar

- Transformar la variable respuesta (Box Cox podría ser útil) para reducir asimetría
- Evaluar retirar outlier(s)
- Bootstrapping
- Evaluar un modelo teórico distinto (Distribuciones para datos de conteo y proporciones)

Homocedasticidad de errores

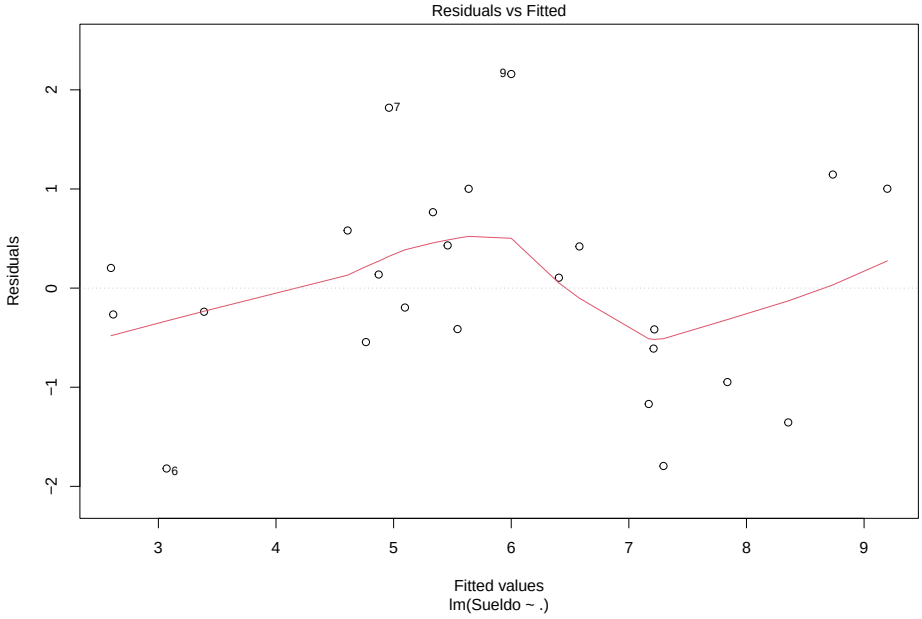
Supuesto

Los errores tienen varianza homogénea, es decir $V(\varepsilon_i) = \sigma^2 \forall i = 1, ..., n$

Medios de verificación

Gráfico de valores ajustados versus residuales y versus la raíz cuadrada del valor absoluto de los residuales estudentizados

```
modelo |> plot(which=1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Sueldo ~ .)
9
7
6
2
1
0
−1
−2
3
4
5
6
7
8
9

```r
modelo |> augment() |>
  with(lowess(x = .fitted, y = .resid)) |> as.data.frame() -> smoothed
modelo |> augment() |>
  ggplot(aes(x=.fitted,y=.resid))+
  geom_point(size = 3) +
  geom_hline(yintercept=0)+
  geom_path(data = smoothed, aes(x = x, y = y), col = "red")+
  labs(x = "Valor ajustado", y = "Residual",
       title = "Evaluación de homocedasticidad", subtitle =
         "Modelo")+
  theme_minimal()
```

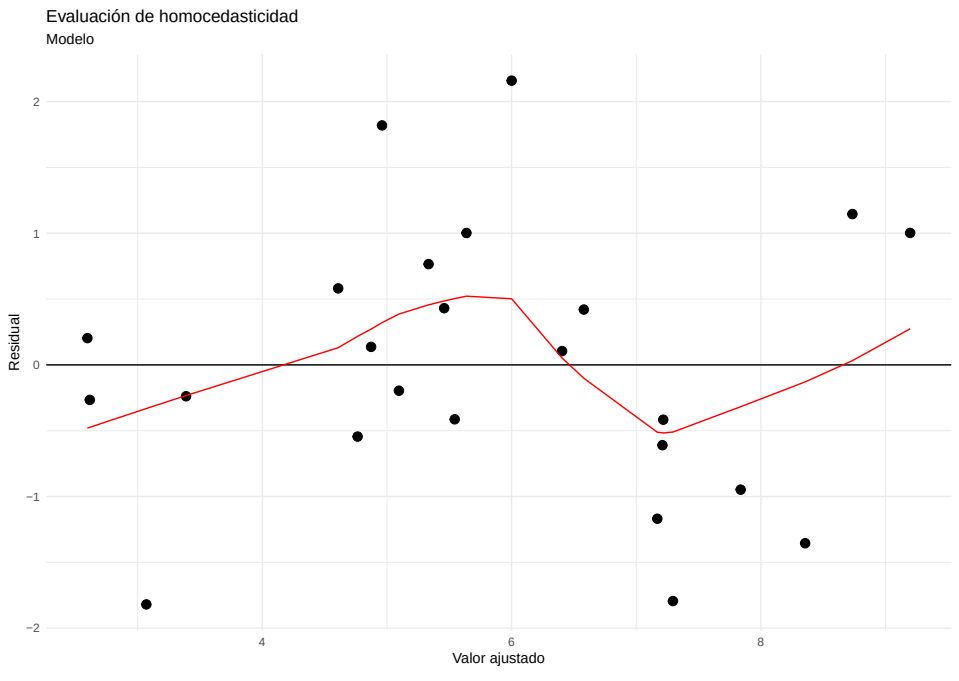
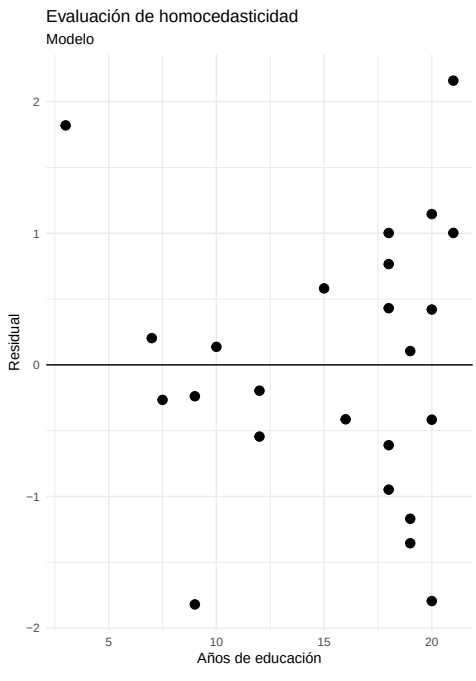

Evaluación de homocedasticidad
Modelo
Residual
Valor ajustado
2
1
0
−1
−2
4
6
8

Gráfico de valores de cada variable explicativa versus residuales (estudentizados).

Conviene también intentar con variables independientes no consideradas en el modelo

```
modelo |> augment() |>
ggplot(aes(x=Educacion,y=.resid))+
  geom_point(size = 3) +
  geom_hline(yintercept=0)+
  labs(x = "Años de educación",
       y = "Residual",
       title = "Evaluación de homocedasticidad",
       subtitle = "Modelo")+
  theme_minimal() -> gra1
modelo |> augment() |>
  ggplot(aes(x=Sexo,y=.resid))+
  geom_point(size = 3) +
  geom_hline(yintercept=0)+
  labs(x = "Sexo",
       y = "Residual",
       title = "Evaluación de homocedasticidad",
       subtitle = "Modelo")+
  theme_minimal() -> gra2
grid.arrange(gra1,gra2,ncol=2)
```


Evaluación de homocedasticidad
Modelo
Residual
2
1
0
−1
−2
5
10
15
20
Años de educación

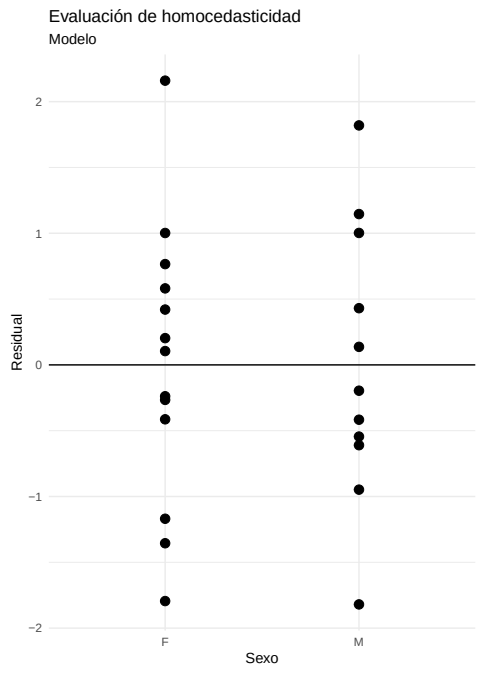
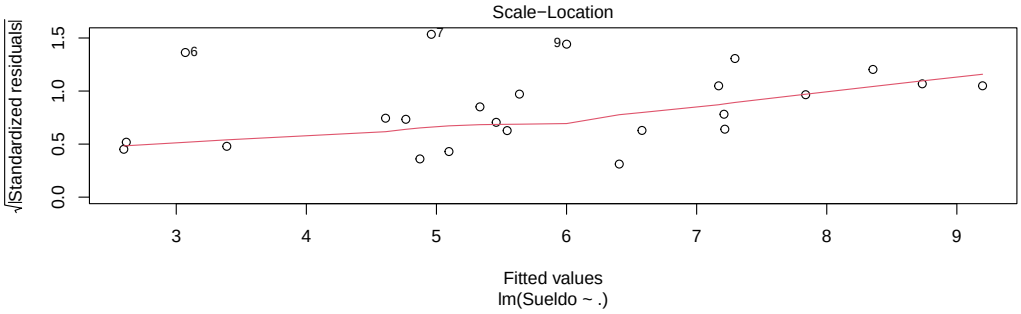

Evaluación de homocedasticidad
Modelo
Residual
2
1
0
−1
−2
F
M
Sexo

Gráfico de valores ajustados versus la raíz cuadrada de los residuales estudentizados en valor absoluto

```
modelo |> plot(which=3)
```



Prueba de Breusch Pagan

```
library(car)
library(olsrr)
library(lmtest)
modelo |> ncvTest()
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.3867398, Df = 1, p = 0.53402
```

```
modelo |> ols_test_breusch_pagan()
```

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

              Data
---------------------------------
Response : Sueldo
Variables: fitted values of Sueldo

       Test Summary
----------------------------
DF            =    1
Chi2          =    0.3867398
Prob > Chi2   =    0.5340181
```

```r
modelo |> bptest(studentize = F)
```

```
    Breusch-Pagan test

data:  modelo
BP = 0.86645, df = 4, p-value = 0.9293
```

```r
modelo |> bptest(studentize = T)
```

```
    studentized Breusch-Pagan test

data:  modelo
BP = 1.0883, df = 4, p-value = 0.8961
```

Motivos

- La varianza es función de la media
- Los errores son multiplicativos
- Falta de normalidad
- Desgaste o mejora en el proceso de toma o recolección de datos

Consecuencias

- Falta de precisión en las estimaciones intervalares.
- Pruebas de hipótesis basadas en las distribuciones, t, Chi cuadrado, F no son válidas.
- Predicciones ineficientes.

Acciones a tomar

- Transformar la variable respuesta. Se sugiere Box Cox. ¿Qué sucede con las estimaciones y predicciones?
- Utilizar mínimos cuadrados ponderados, de tal modo que $V(Y|X) = \sigma^2 X$
- Considerar un modelo que contemple heterocedasticidad

Independencia de errores

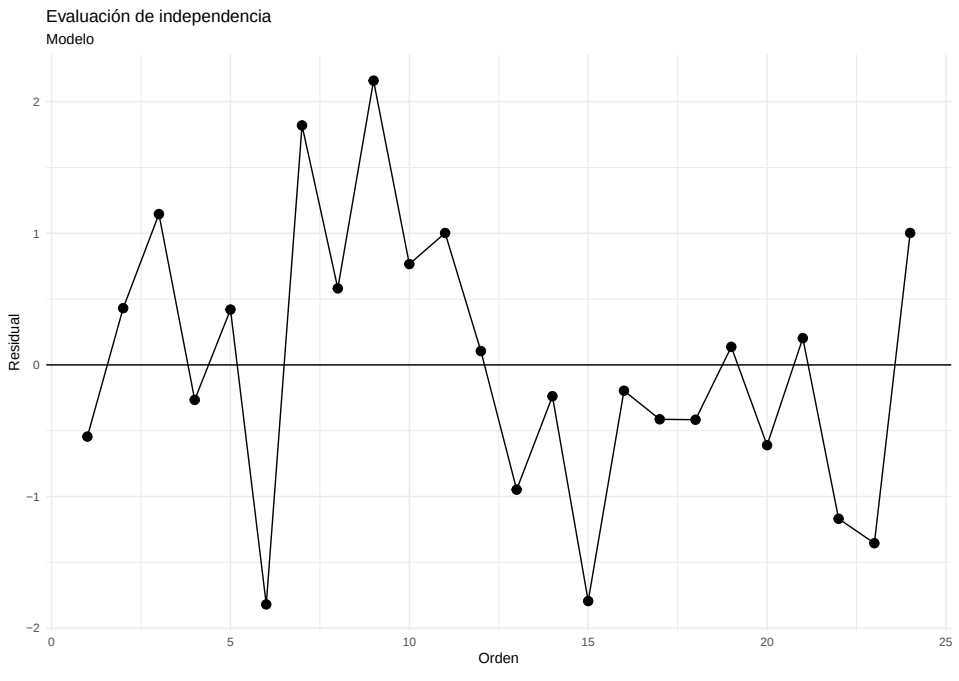
Supuesto

Los errores son independientes = los errores no están autocorrelacionados

Medios de verificación

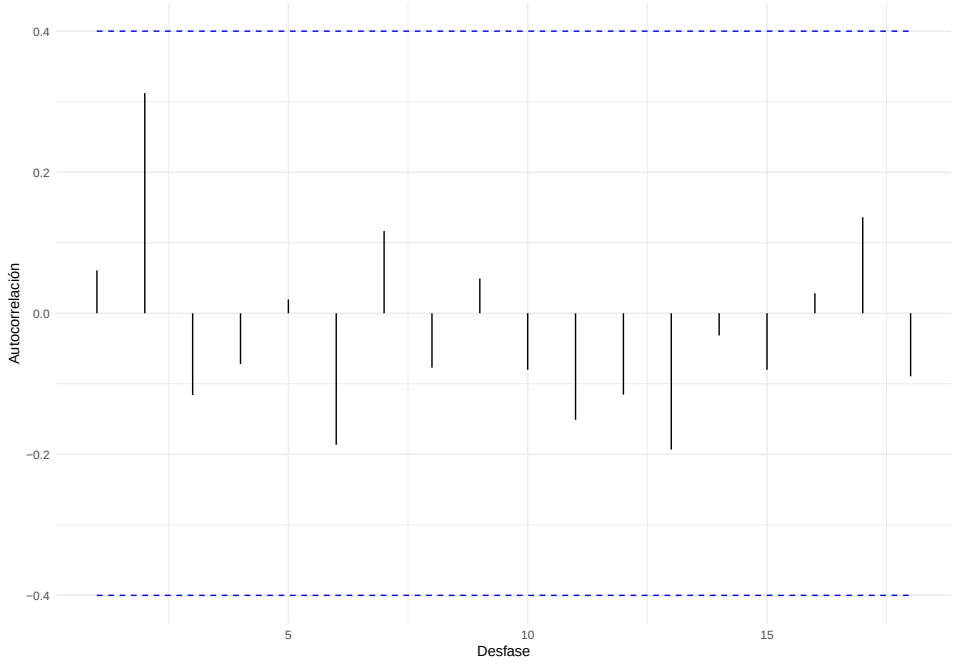
Gráfico de los residuales en orden

```
data.frame(residuales) |>
ggplot(aes(x=1:nrow(datos),y=residuales))+
  geom_point(size = 3) +
  geom_line()+
  geom_hline(yintercept=0)+
  labs(x = "Orden", y = "Residual", title = "Evaluación de independencia",
  theme_minimal()
```


Evaluación de independencia
Modelo
Residual
Orden
2
1
0
−1
−2
0
5
10
15
20
25

Correlograma

```
library(ggfortify)
residuales |>
TSA::acf(lag = 18, plot=F) |>
autoplot() +
labs(x = "Desfase",y = "Autocorrelación") +
theme_minimal()
```


0.4
0.2
0.0
−0.2
−0.4
Autocorrelación
5
10
15
Desfase

Prueba de Durbin Watson

```
modelo |> dwtest(alternative = "two.sided")
```

```
    Durbin-Watson test

data:  modelo
DW = 1.8251, p-value = 0.6178
alternative hypothesis: true autocorrelation is not 0
```

```r
modelo |> durbinWatsonTest(alternative = "two.sided",
                           max.lag = 10,
                           reps = 1e5)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.06083440      1.825133 0.61290
   2       0.31228900      1.239576 0.07010
   3      -0.11592874      1.986495 0.64150
   4      -0.07200947      1.894077 0.67282
   5       0.01955072      1.688519 0.86622
   6      -0.18651831      1.964531 0.32574
   7       0.11658913      1.215914 0.36538
   8      -0.07713629      1.582569 0.65156
   9       0.04915486      1.137808 0.53220
  10      -0.08010438      1.240744 0.97334
 Alternative hypothesis: rho[lag] != 0
```

Bibliografía

- Kyun, T., Hong, J. (2019). More about the basic assumption of t-test: normality and sample size
- Mendenhall, W. (2012). A Second Course in Statistics Regression Analysis. Pearson.
- Montgomery, D., Peck, E., Vining, G. (2012). Introduction to Linear Regression Analysis. Wiley.
- Rawlings, J. (1998). Applied Regression Analysis: A Research Tool. Springer.
- Weisberg, S. (2014) Applied Linear Regression. Wiley.