Practica 2.1

Agustin Riquelme y Heriberto Espino

2023-10-01

1. Lectura de la base de datos

Podemos observar que la base contiene tres columnas, una de nombre edad, otra número de postores y el último precio de subasta

```
## [1] "Edad"              "Num Postores"      "Precio de subasta"
```

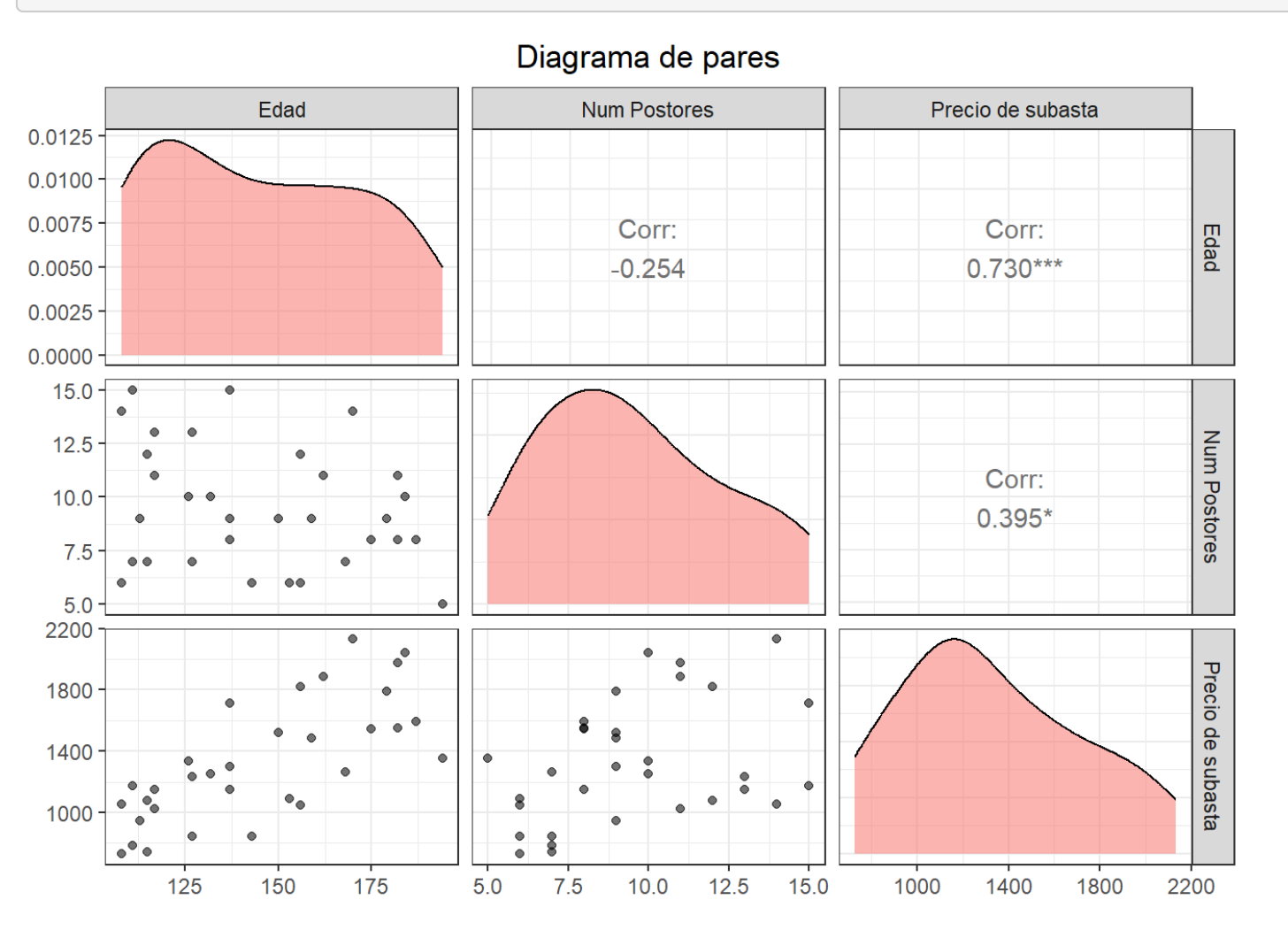
```
## # A tibble: 6 × 3
##    Edad `Num Postores` `Precio de subasta`
##   <dbl>          <dbl>               <dbl>
## 1   127             13                1235
## 2   115             12                1080
## 3   127              7                 845
## 4   150              9                1522
## 5   156              6                1047
## 6   182             11                1979
```

Para la creación de los modelos vamos a definir nuestra variable de respuesta que será el precio de subasta, mientras que las otras dos serán la edad y el número de postores, a cada una de estas variables predictoras se le realizará un modelo de regresión por separado, a continuación lo veremos:

```
x1 <- data$Edad # Predictora 1
x2 <- data$`Num Postores` # Predicora 2
y <- data$`Precio de subasta` # Variable de respuesta
n <- length(x1) # Número de datos en cada variable
```

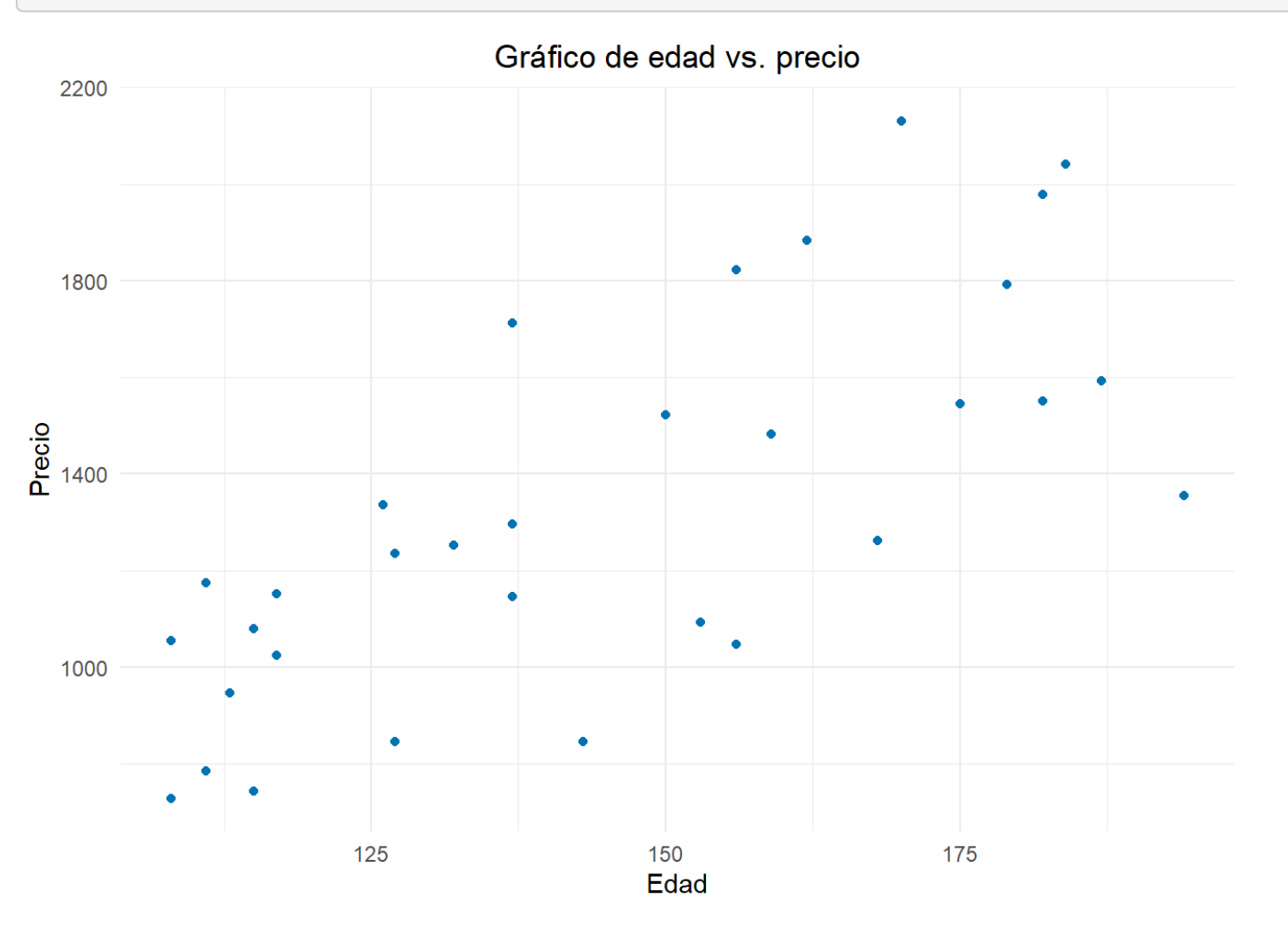
Vamos a ver el diagrama de pares para irnos dando una idea de como son los datos de la muestra de la subasta

```
ddp(data)
```



2. Modelo 1 Precio~edad

```
graf.d.x1(x1, x2, y, 0.05)
```



Para ver el comportamiento de los datos, una gráfica de dispersión siempre es la mejor opción, observando dicha herramienta, podemos observar que el comportamiento entre las variables tienen una relación positiva, además de poder observar los límites de los datos que van de 100 a 200 para el caso de la edad, y para el caso del precio desde 600 hasta 2200.

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -485.04 -192.36   31.03  157.50  541.47
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -192.047    264.372  -0.726    0.473
## x1            10.480      1.793   5.844 2.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.5 on 30 degrees of freedom
## Multiple R-squared:  0.5324, Adjusted R-squared:  0.5168
## F-statistic: 34.15 on 1 and 30 DF,  p-value: 2.158e-06
```

En el modelo podemos observar la relación positiva que habíamos visto anteriormente con el valor de β_1. En el apartado de residuales no observamos un comportamiento tan simétrico y el promedio es distinto de cero, por lo que podemos decir que los residuos no se comportan normalmente. Para el caso de β_1 su variación es baja, sin embargo para $ _0$ su variación es bastante alta. Sim embargo para el valor estadístico t, para β_0 su valor es cercano a cero, indicando que la hipótesis nula puede ser aceptada para este caso, sin embargo para β_1 sucede lo contrario, esto igualmente puede ser observado en el valor p. Por otro lado nuesto error residual es bastante alto, es decir, nuestro error en cada predicción, se va a desviar mucho de nuestra línea de regresión. Podemos igualmente observar que la precisión de nuestro modelo no es tan alta, pero al menos está poco encima del 50%, es decir, o bien, poco más del 50% de los datos observados pueden ser explicados con nuestro modelo. Finalmente podemos decir con el estadístico F que hay evidencia suficiente para rechazar H_0: No existe relación entre ambas variales.

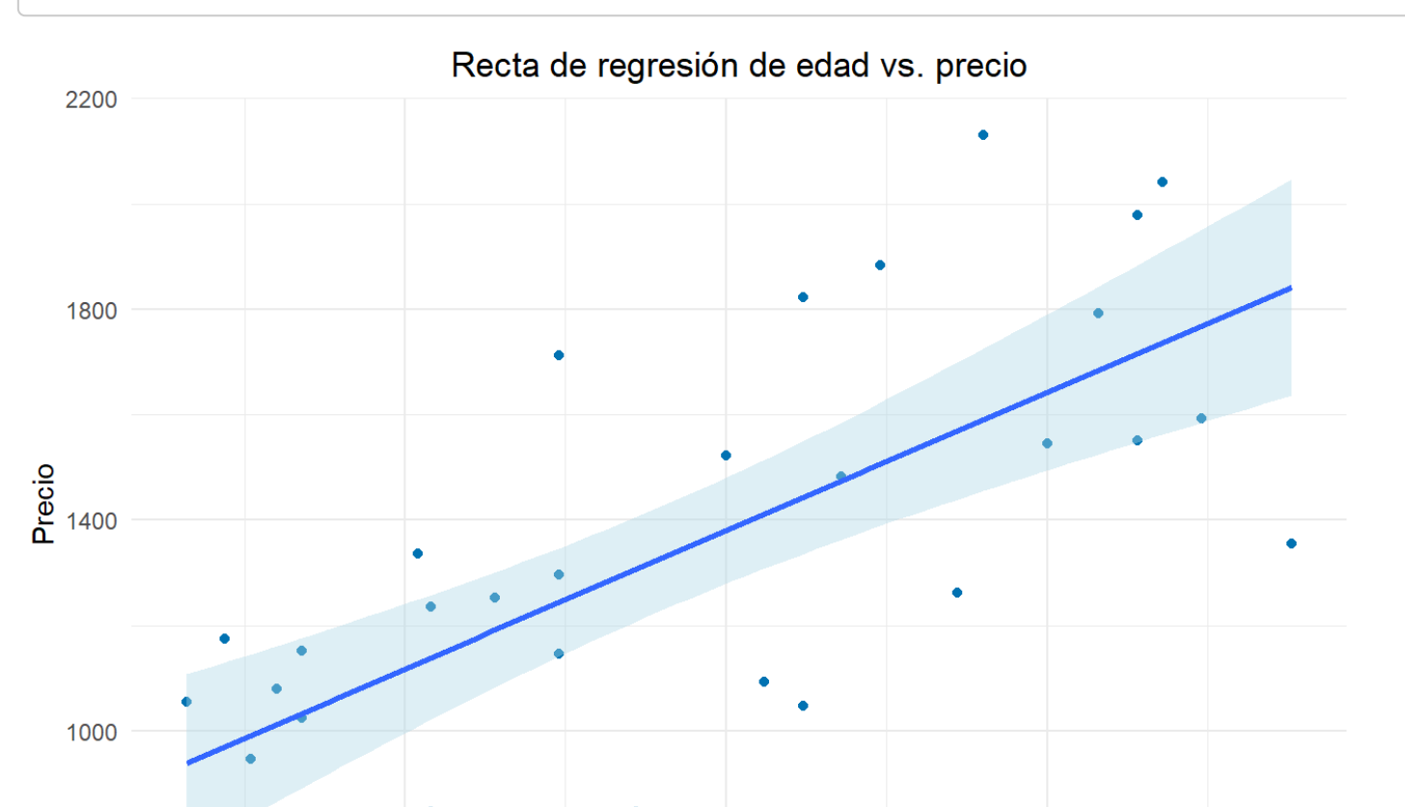
```
AIC(modelo1)
```

```
## [1] 453.8772
```

Por otro lado el criterio de Información de Akaike nos va a servir para comparar ambos modelos, es decir, el que tenga menor valor AIC, será el mejor modelo. A continuación se muestra la línea de regresión del modelo.

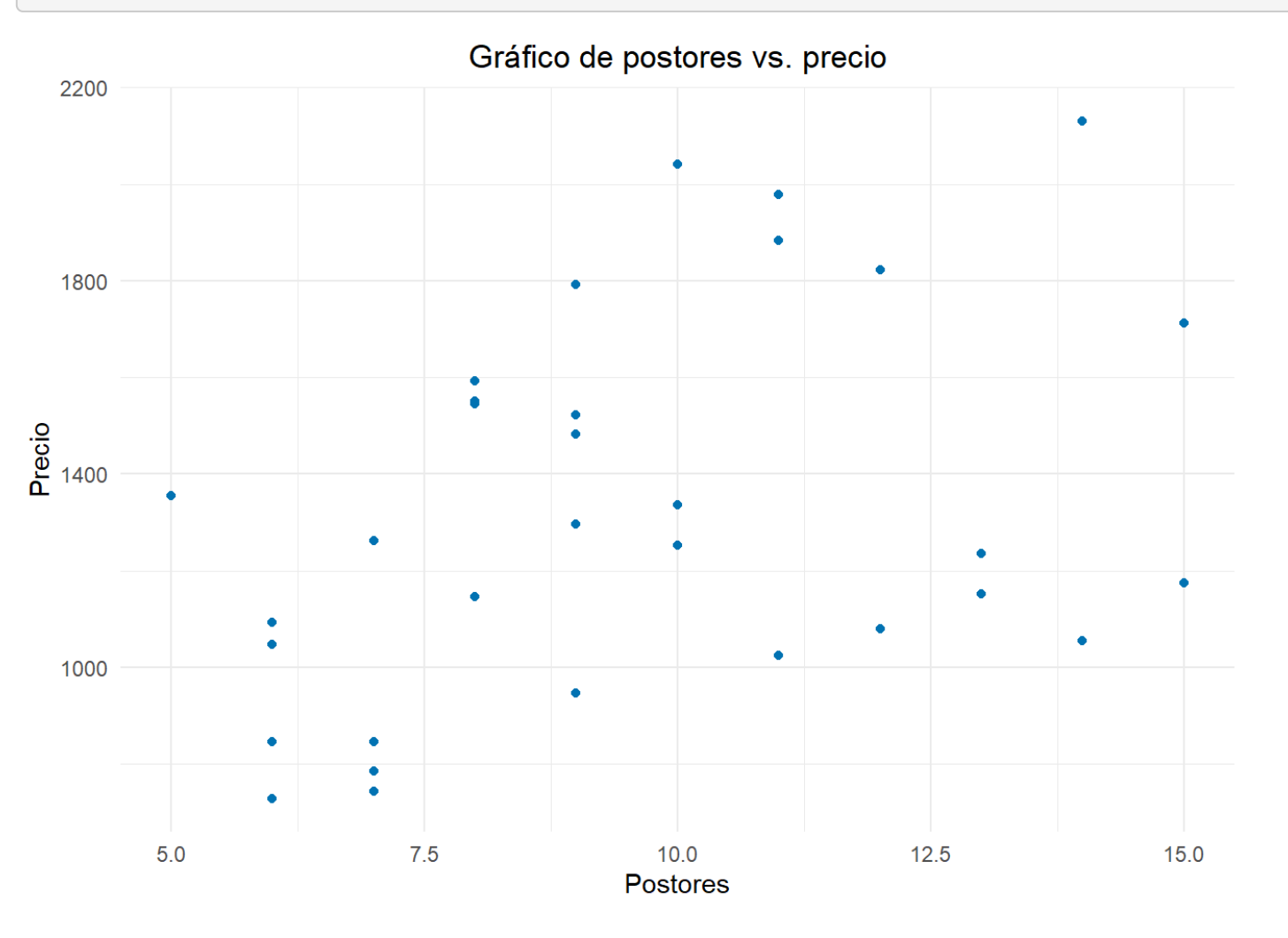
```
graf.mrls.x1(x1, x2, y, 0.05)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



3. Modelo 2: Precio~Num Postores

```
graf.d.x2(x1, x2, y, 0.05)
```



Como mencionamos anteriormente, la gráfica de dispersión es de ayuda para poder observar el comportamiento de nuestras variables, en este caso podemos observar un comportamiento más disperso que el antrior el cual ronda entre los 5 a 15 postores, observando una relación positiva

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -516.60 -355.05  -29.02  303.23  688.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   804.91     230.83   3.487  0.00153 **
## x2             54.76      23.24   2.356  0.02518 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 367.4 on 30 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1281
## F-statistic: 5.553 on 1 and 30 DF,  p-value: 0.02518
```

En el modelo podemos observar la relación positiva que habíamos visto anteriormente con el valor de β_1, además de observar que los residuos no se comportan tanto de manera simétrica. En este caso, tanto β_0 como β_1 su varianza es baja, mucho menor que las betas, y a su vez en el estadístico, ambos valores están alejados de cero, pero no son tan grandes comparadas con la varianza. Por otro lado, el error residual es más elevado que el pasado y en conjunto nos da que nuestra precisión del modelo es muy baja, siendo 0.15, indicando que sólo el 15% de nuestros datos pueden ser descritos con el modelo. Finalmente podemos decir con el estadístico F que parece no ser lo suficientemente grande comparado con nuestra muestra pudiendo indicar que no existe una relación entre las variables.

```
AIC(modelo2)
```

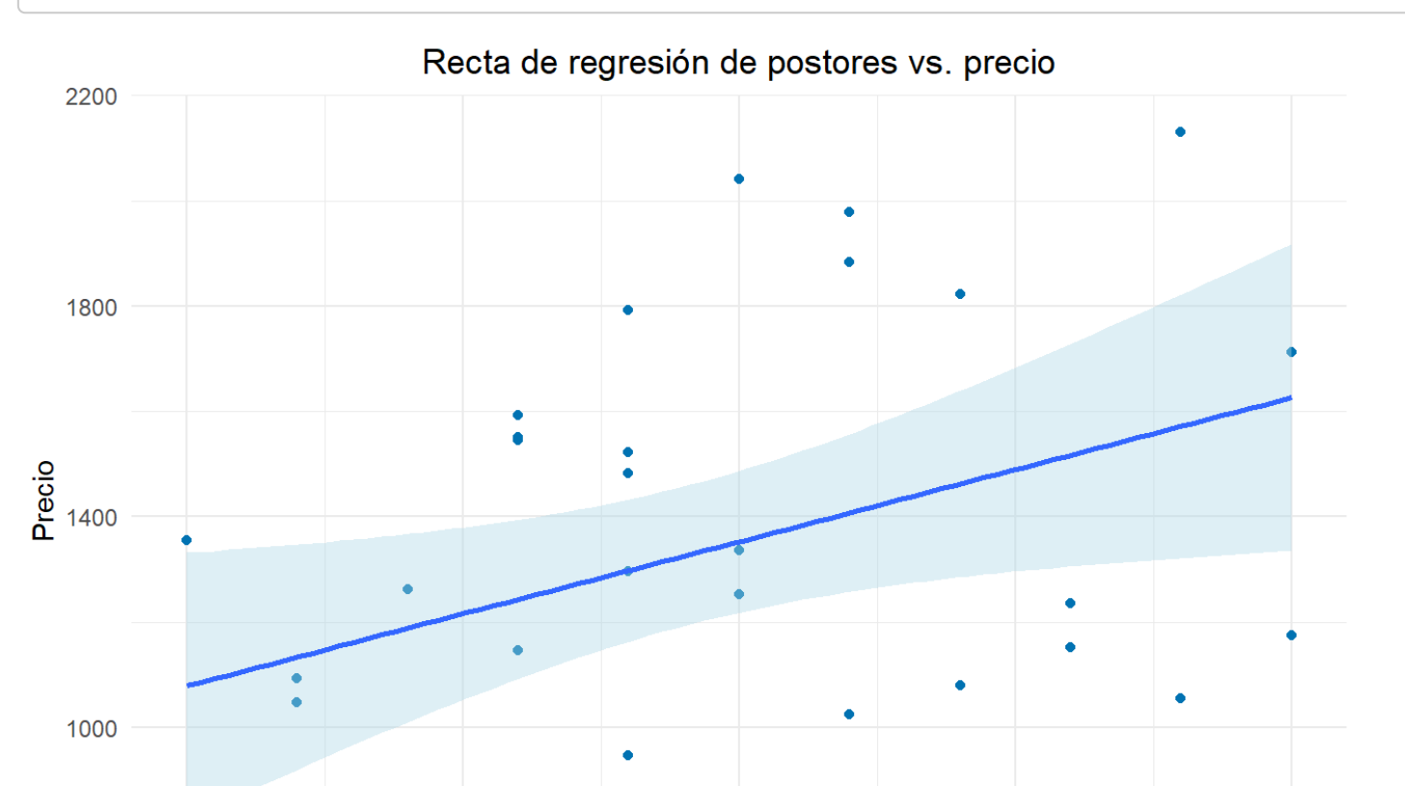
```
## [1] 472.7648
```

Comparando el AIC de ambos modelos, se puede decir que el primero es mejor que este segundo modelo. A continuación se observa la gráfica de dispersión para este modelo.

```
##
## Call:
## lm(formula = modelo1)
##
## Coefficients:
## (Intercept)           x1
##     -192.05        10.48
```

```
graf.mrls.x2(x1, x2, y, 0.05)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



4. Construcción del modelo de regresión múltiple

Para este caso, se incluirán ambas variables como predictoras, es decir, comparado con los anteriores modelos, aquí se incluirá otro criterio para saber si existe una relación predecible entre las variables, su construcción es similar a la del modelo pasado, únicamente, como se djo al inicio, se incluirán dos variables predictoras que fueron las de los modelos pasados. Para ello se usará el mismo modelo lm que en casos pasados pero haremos la suma de ambas variables.

```
modmult <- lm(y ~ x1+x2)
anova <- aov(modmult)
summary(anova)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## x1           1 2555224 2555224  143.41 9.53e-13 ***
## x2           1 1727838 1727838   96.97 9.34e-11 ***
## Residuals   29  516727   17818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Además, al observar el anova podemos ver mucha significancia en las variables predictoras al tener tanto un estadístico alto así como un p valor significativo y que ambas variables predictoras contribuyen al modelo. Igual podemos observar la comparación de media y varianzas y a su vez podemos decir que al menos una β_i es distinta de cero.

```
modmult <- lm(y ~ x1+x2)
anova <- aov(modmult)
summary(modmult)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.49 -117.34   16.66  102.55  213.50
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1338.9513   173.8095  -7.704 1.71e-08 ***
## x1             12.7406     0.9047  14.082 1.69e-14 ***
## x2             85.9530     8.7285   9.847 9.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.5 on 29 degrees of freedom
## Multiple R-squared:  0.8923, Adjusted R-squared:  0.8849
## F-statistic: 120.2 on 2 and 29 DF,  p-value: 9.216e-15
```

Al observar el p valor del estadístico F del modelo, podemos ver que es muy bajo, es decir, es significativo, ya anteriormente habíamos visto el de cada variable por separado e igualmente eran significativos. El error residual de este modelo es mucho más bajo por lo que podemos decir que nuestras predicciones pueden variar en 133.5 a comparación de los datos originales y si se observa con detalle, la confianza de este modelo aumenta significativamente con el número de variables consideradas.

```
aic(x1, x2, y, 0.05)
```

Test Akaike

Modelo	AIC
1	453.8772
2	472.7648
Múltiple	408.8771

Por el criterio de Akaike, parece ser el mejor modelo este de regresión lineal multiple, pues su valor es el más bajo. Esto debido a que toma en cuenta más argumentos para hacer las predicciones.

5. Extracción de Información del modelos

Para la extracción de la información del modelo, lo importante a considerar serán nuestras betas, nuestra S^2 y nuestro estadístico así como su p valor, para ello, en el primer caso, las betas serán obtenidas con los coeficientes del modelo, especificamente aquellos que se encuentran en la primer fila y en cada una de las columnas.

```
extr.datos(x1, x2, y, 0.05)
```

Extracción de datos

Variable	Valor
β₀	-1338.95134
β₁	12.74057
β₂	85.95298
s	133.48467
s²	17818.15655

6. Prueba general del Modelo

La prueba general del modelo se realizará mediante la comparación del valor p, si este es significativo, es decir, es menor que 0.05, el modelo de igual manera será significativo, en caso contrario no existirá evidencia para esta afirmación. Siendo su validación de la siguiente manera y concluyendo que el modelo sí es significativo al rechazar H_0

```
prueb.glob(x1, x2, y, 0.05)
```

Pruebas global para βi

Variable	Valor
FC	120.188161679417
P value f	0.00000000000000921635859536383
Conclusión	Se rechaza H0, es decir, el modelo es significativo

7. Correlación

```
corr(x1, x2, y, 0.05)
```

Coeficiente de correlación

Rho	Conclsuión
R cuad	Hay una correlación fuerte
Valor	0.892343916353148
R adj	Hay una correlación fuerte
Valor	0.884919358860262

El coeficiente de correlación (Rho) es una medida numérica que varía entre -1 y 1. Un valor de 1 indica una correlación positiva perfecta, mientras que un valor de -1 indica una correlación negativa perfecta. En este caso, el valor de Rho es 0.892343916353148, lo que sugiere una correlación positiva fuerte entre las variables que estás analizando.

R cuadrado (R cuad) es una medida que indica qué porcentaje de la variación en una variable puede explicarse por la otra variable en el modelo de regresión. Un valor de 0.892343916353148 sugiere que aproximadamente el 89.23% de la variación en una variable puede explicarse por la otra variable en este análisis.

R cuadrado ajustado (R adj) es similar a R cuadrado, pero tiene en cuenta el número de predictores en el modelo. Si el valor de R adj es alto, indica que el modelo es adecuado para explicar la variación en la variable dependiente.

En resumen, los valores proporcionados indican que hay una correlación positiva fuerte entre las variables analizadas, y aproximadamente el 89.23% de la variación en una variable puede explicarse por la otra variable en el modelo. Además, tanto R cuadrado como R cuadrado ajustado sugieren que el modelo es adecuado para explicar la relación entre las variables.

8. Pruebas Individuales

Para el caso de las pruebas individuales, se va a extraer del modelo cada uno de los p valores de las variables predictoras. Al estar estas en la sección de los coeficientes, bastará con indicar su posición en la matriz para obtenerlas:

Y para sus respectivas pruebas, como en el caso anterior y ya muchos otros modelos, los p valores se compararán con nuestro alpha y podemos observar que en ambos casos se rechaza H_0, queriendo decir que ambas betas son significativas para el modelo

```
prueb.indiv1(x1, x2, y, 0.05)
```

Prueba individual para β₁

Variable	Valor
P value β₁	0.0000000000000169275643184263
Concusión	Se rechaza H0, es decir, β₁ es significativo para el modelo

```
prueb.indiv2(x1, x2, y, 0.05)
```

Prueba individual para β₂

Variable	Valor
P value β₂	0.0000000000934495315616398
Conclusión	Se rechaza H0, es decir, β₂ es significativo para el modelo

9. Intervalos de confianza para β_i

Para el calculo de los intervalos de confianza de cada beta, se obtendrá el valor de probabilidad acumulada de una t student de dos colas con $n-1$ grados de libertad y se realizará el calculo para cada beta restandole el producto de la probabilidad acumulada con su error estándar, siendo su calculo de la siguiente manera:

Obteniendo los siguientes resultados, que afirman a un nivel alpha = 0.05 que cada beta se encuentra en ese intervalo.

```
int.conf.b1(x1, x2, y, 0.05)
```

Int. de confianza β₁

Límites	Valor
Inferior	10.89534
Superior	14.58580

```
int.conf.b2(x1, x2, y, 0.05)
```

Int. de confianza β₂

Límites	Valor
Inferior	68.15104
Superior	103.75493

10. Gráfica del modelo

Al graficar el modelo, podemos observar como puntos rojos cada uno de los datos que tenemos en el modelo y el plano representa nuestro modelo de regresión, la distancia entre cada uno de estos va a ser el error de predicción y al observarlo a detalle podemos ver que todos los errores parecen pequeños.

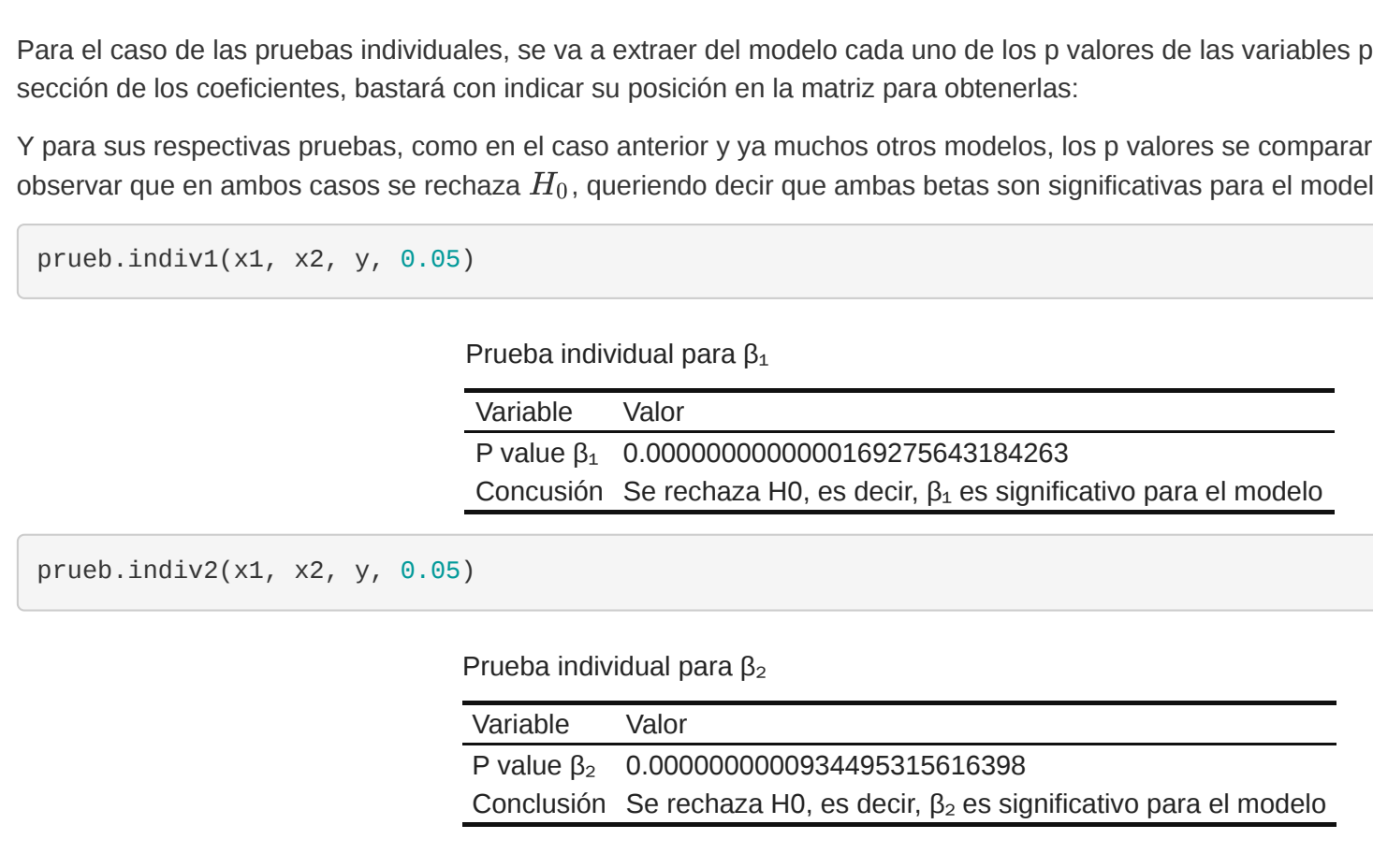
```
graf.mrlm(x1, x2, y, 0.05)
```

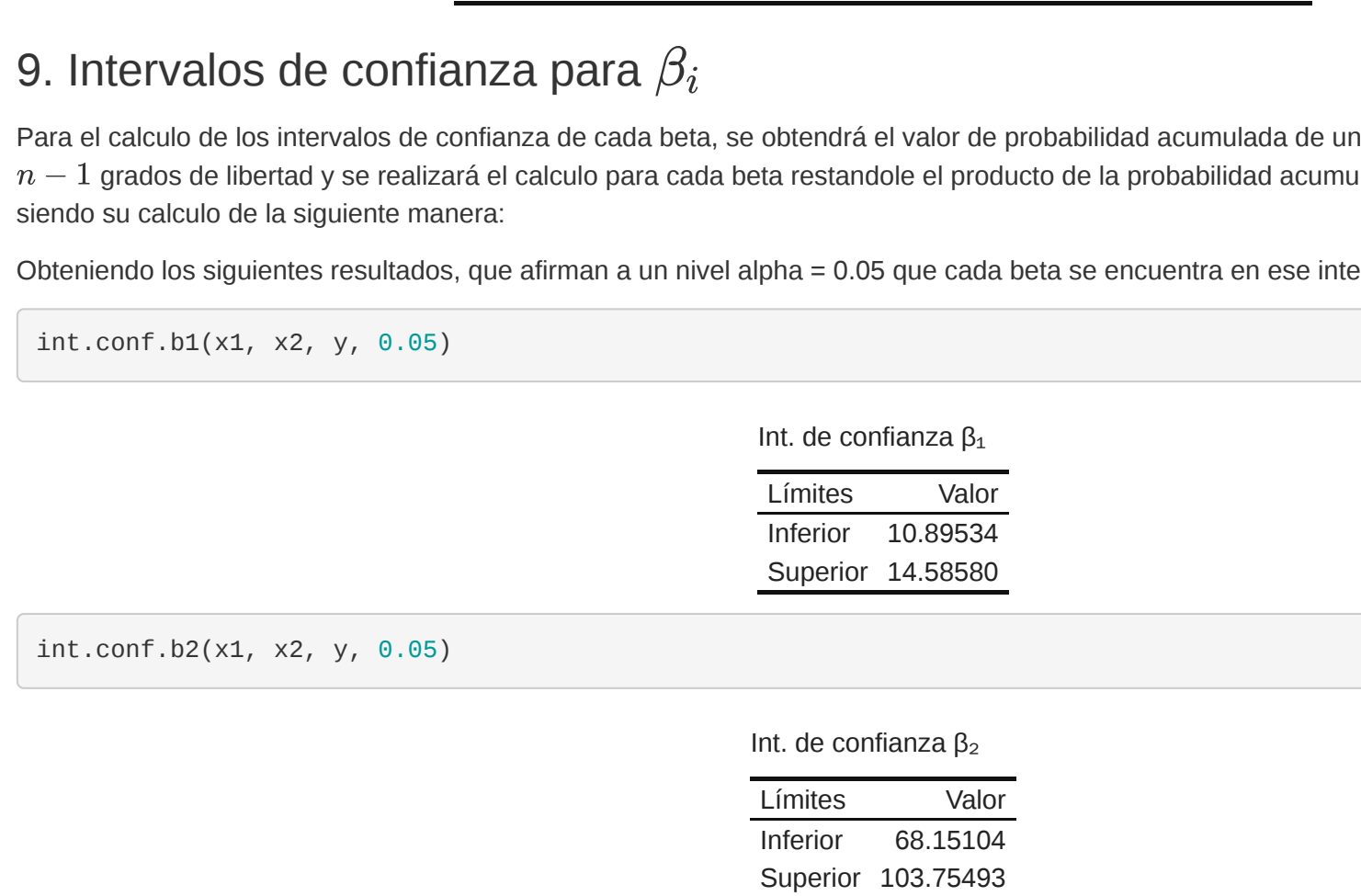
```
## [1] 0.05138413
```

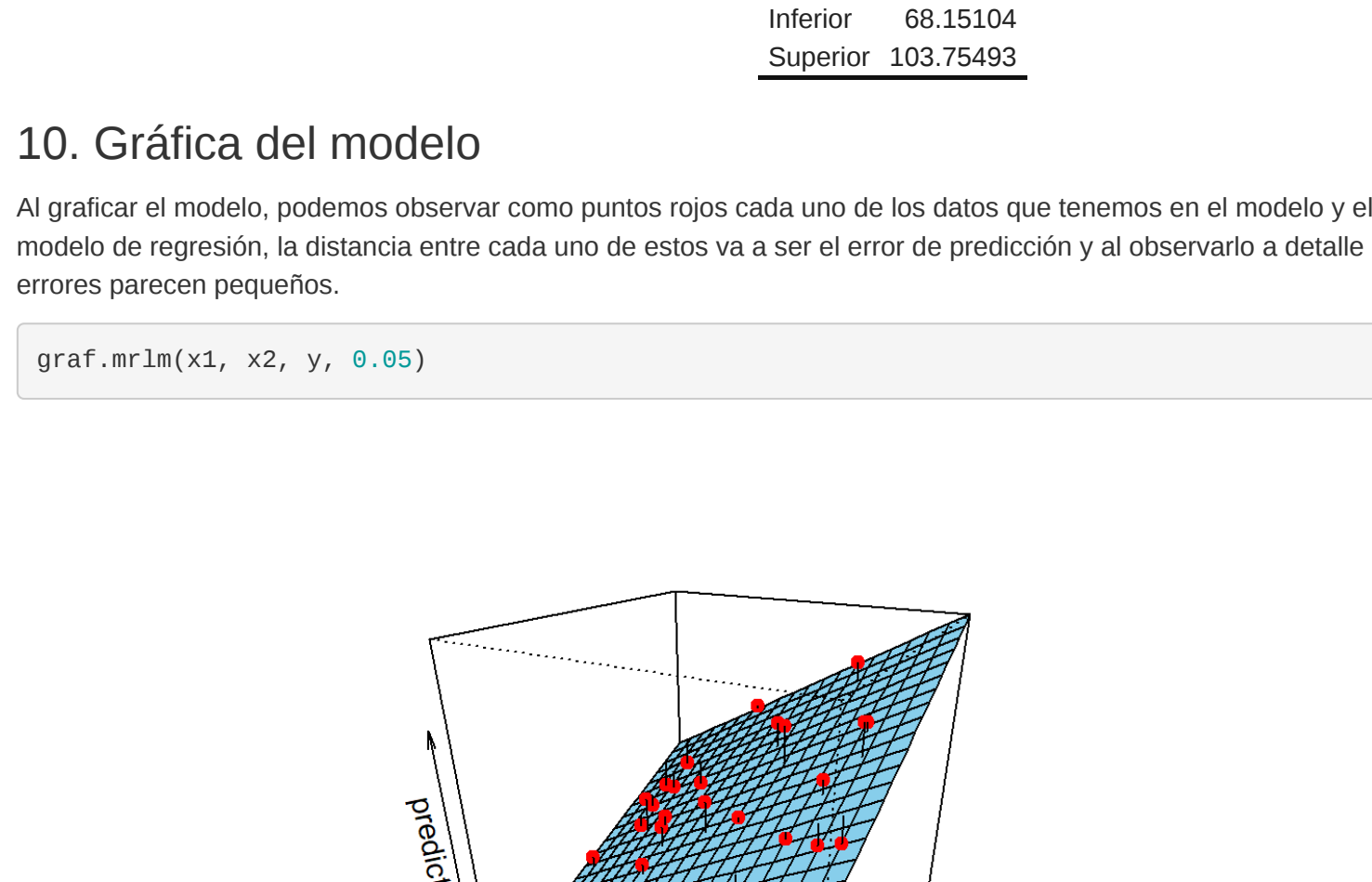
11. Análisis de residuales

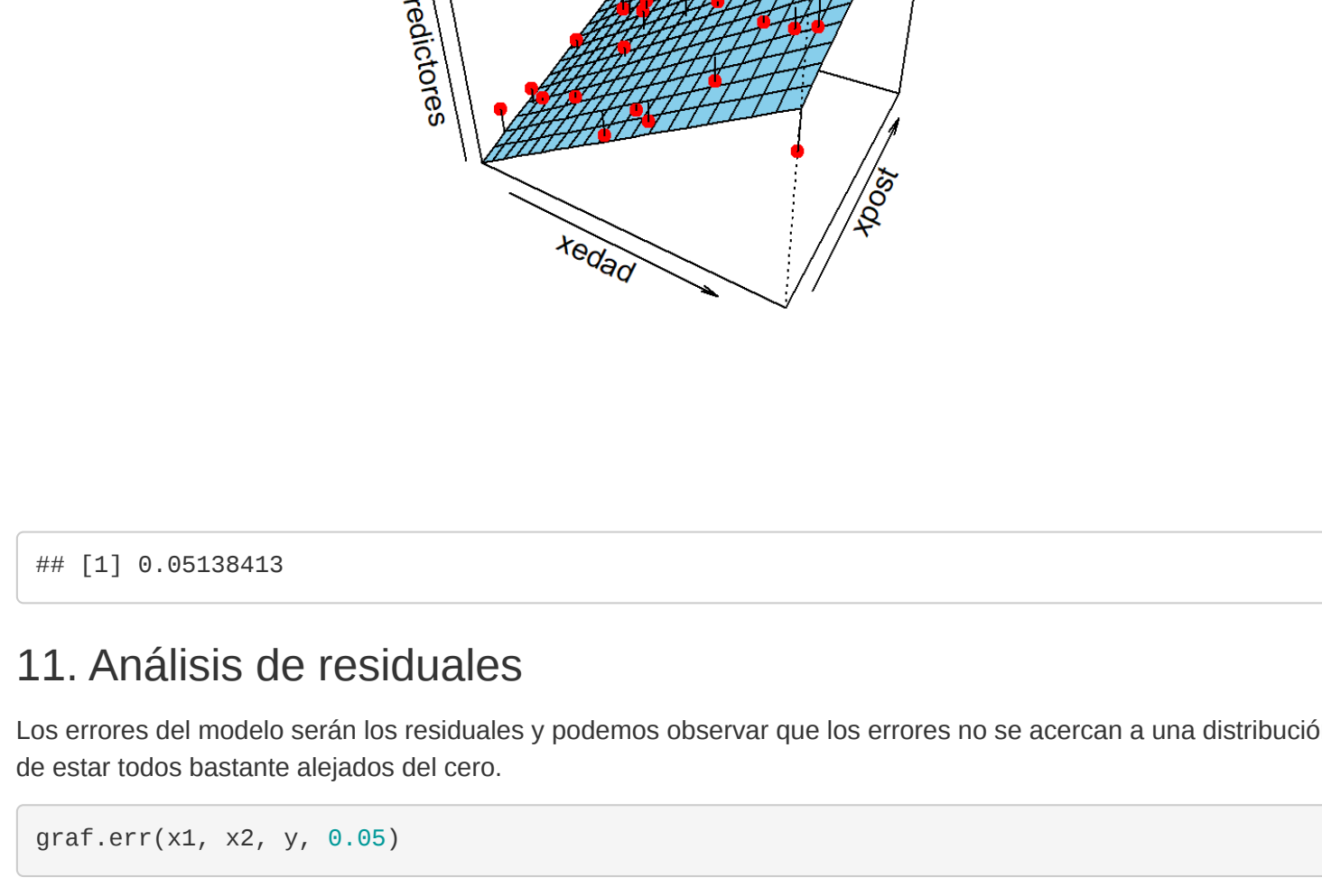
Los errores del modelo serán los residuales y podemos observar que los errores no se acercan a una distribución normal al graficarlos, además de estar todos bastante alejados del cero.

```
graf.err(x1, x2, y, 0.05)
```









Como el número de datos es mayor que 30, se usará la prueba Lillie y la Anderson para comprobar si los errores se comportan de manera normal con mayor evidencia.

```
norm.err(x1, x2, y, 0.05)
```

Normalidad H0: Err. norm vs H1: Err. no norm

Test.Lillie	Valor
P valor	0.181286635723582
Conclusión	p valor ≥ α, no rechazamos H0 .

```
norm.err2(x1, x2, y, 0.05)
```

Normalidad H0: Err. norm vs H1: Err. no norm

Test.Anderson	Valor
P valor	0.204437140763369
Conclusión	p valor ≥ α, no rechazamos H0

12. Correlación de los errores

```
corr.err(x1, x2, y, 0.05)
```

Test Durbin-Watson

Test	Conclusión
DW	1.87200832574999
P valor	0.355352817154974
Conclusión	No hay una correlación en los residuos

En el análisis estadístico realizado, se aplicó el Test Durbin-Watson para evaluar la presencia de autocorrelación en los residuos del modelo. El resultado del test arrojó un valor de 1.87200832574999 para el estadístico Durbin-Watson (DW). En este contexto, un valor de DW cercano a 2 sugiere que no hay autocorrelación en los residuos. Además, se calculó el valor P, que fue encontrado como 0.355352817154974. Un valor P mayor que el nivel de significancia (generalmente 0.05) indica que no hay evidencia suficiente para rechazar la hipótesis nula de que no hay autocorrelación en los residuos.

En base a estos resultados, se llega a la conclusión de que no hay una correlación significativa en los residuos del modelo. Esto significa que los errores residuales del modelo analizado no muestran un patrón sistemático en su distribución, lo que refuerza la fiabilidad de las conclusiones obtenidas a partir de este análisis estadístico