

Análisis de adaptabilidad de un modelo de regresión lineal para el estudio de la presencia del cloro en tanques de agua consecuencia de procesos químicos

Eduardo J. Sánchez P.* María G. Nuñez E.**

*Departamento de Ciencias de la Computación,
Facultad Experimental de Ciencias y Tecnología (FaCyT),
Universidad de Carabobo, Carabobo, Venezuela.*

7 de agosto de 2008

Resumen

En diversas aplicaciones estadísticas, cuando se desean realizar inferencias sobre una población en particular, también se desean realizar predicciones sobre su comportamiento a futuro. Estas predicciones deben realizarse sobre consideraciones de la población las cuales utilizan métodos de estimación sobre modelos matemáticos de los datos. En particular, uno de estos métodos, el método de regresión lineal simple, permite realizar predicciones y estimaciones muy versátiles desde el punto de vista del control de la precisión. En este trabajo presentamos su aplicación a un caso de prueba en particular, haciendo énfasis en el marco teórico propio del método, así como en los resultados y predicciones propias del caso de estudio.

*3er año. C.I.: V-18.410.742. eduardo.nogales@gmail.com. <http://www.bassy.bipolar.com.ve>

**3er año. C.I.: V-18.062.013. marygabnunes@gmail.com

Índice

1. Introducción	3
2. Planteamiento del problema	3
3. Objetivos	3
4. Marco teórico y resultados del estudio	4
4.1. Datos	4
4.2. Análisis exploratorio de los datos	4
4.3. Ajuste del modelo y análisis de variancia	7
4.4. Coeficientes de determinación y coeficientes de correlación	9
4.5. Estudio de los valores residuales y verificación de supuestos	9
4.6. Predicciones en el modelo	11
5. Ajuste de un modelo cuadrático	14
5.1. Estudio de los valores residuales y verificación de supuestos en el modelo parabolico	14
6. Conclusiones	17

Índice de figuras

1. Histogramas de las variables estudiadas	5
2. Grafico de dispersión	6
3. Grafico de residuales	10
4. Grafico de la prueba de normalidad	12
5. Grafico de predicciones del modelo	13
6. Grafico de los residuales para el modelo parabolico	15
7. Grafico de la prueba de normalidad para el modelo parabolico	16

1. Introducción

En múltiples ocasiones, en el análisis de datos, nos encontramos con situaciones en las que se requiere analizar la relación entre dos o más variables cuantitativas. En general, con este análisis se busca, por un lado, determinar si dichas variables están asociadas y en qué sentido se da dicha asociación (es decir, si los valores de una de las variables tienden a aumentar, o disminuir al aumentar los valores de la otra); y por otro, estudiar si los valores de una variable pueden ser utilizados para predecir el valor de la otra. Este proceso se conoce como análisis de regresión y en este trabajo se analizará una de las técnicas más simples para su implementación, a través de un caso del estudio de un caso de estudio particular en conjunto con una discusión apropiada del marco teórico asociado.

2. Planteamiento del problema

El caso de estudio seleccionado para el análisis de regresión busca estudiar la relación entre la cantidad de cloro, medida en litros por semana presente en 44 muestras de agua sometidas a un proceso químico y el tiempo transcurrido en semanas. Como se verá más adelante buscamos verificar si de hecho, el volumen de cloro en el agua depende del tiempo transcurrido desde la aplicación de este proceso para poder así, en función de esta información realizar predicciones sobre la cantidad de cloro, en función del tiempo.

3. Objetivos

Fundamentalmente se desea:

1. Estudiar la aplicabilidad del modelo lineal simple para el proceso de análisis de regresión a través de un caso de estudio concreto.
2. Analizar la relación de significación entre el tiempo transcurrido desde la realización del proceso químico y la cantidad de cloro en el agua.
3. Predecir la cantidad de cloro en función al tiempo transcurrido en semanas.

Más específicamente se busca

1. Realizar un estudio exploratorio previo de los datos que incluya un diagrama de dispersión, para determinar el tipo de relación que existe entre las variables involucradas.
2. Ajustar el modelo lineal y presentar todas las pruebas consideradas necesarias para determinar la idoneidad de este para luego, realizar un análisis de varianza

para determinar si la ecuación de regresión explica un porcentaje significativo de la variabilidad en la variable dependiente.

3. Calcular el coeficiente de determinación del modelo ajustado, y el coeficiente de correlación lineal muestral, probando este último con un nivel de significación de 0.05 y analizar las conclusiones apropiadas.
4. Verificar que se cumplen los supuestos del modelo.
5. Predecir cuánto sería la cantidad de cloro que tendría el agua al transcurrir 15 semanas y deducir el tiempo que debe transcurrir para poder medir en una muestra de agua 0.37 de cloro.
6. Estudiar las conclusiones.

4. Marco teórico y resultados del estudio

Como podrá observarse, el marco teórico será discutido en conjunto con los resultados obtenidos del estudio, de esta manera no solo se logra un orden de presentación adecuado sino que se justifica cada resultado al mostrar la fuente de su obtención.

El análisis de regresión se realizó con la ayuda de R v.- 2.6.1. Un programa para la asistencia computacional de estudios estadísticos. A manera de justificación adicional se utilizará cuando sea de ayuda, la notación utilizada en este para el análisis de los datos de entrada.

4.1. Datos

El análisis de regresión se realizó, como es de esperarse, sobre los datos de las cantidades de cloro en función al número de semanas que han transcurrido desde la aplicación de cierto proceso químico en el agua. En este caso en particular se consideraron las variables x como el tiempo transcurrido (en semanas) desde la aplicación de dicho proceso e y como la cantidad de cloro en esta, consecuencia del proceso. Después de la carga de los datos, a través de la siguiente orden:

```
cloro <- data.frame(y,x)
```

4.2. Análisis exploratorio de los datos

Después de cargar los datos, a través de las ordenes `hist(y)` e `hist(x)` obtenemos los histogramas para ambas variables mostrados en la figura 1. De igual forma podemos ver el gráfico de la relación entre ambas, el cual se muestra en la figura 2

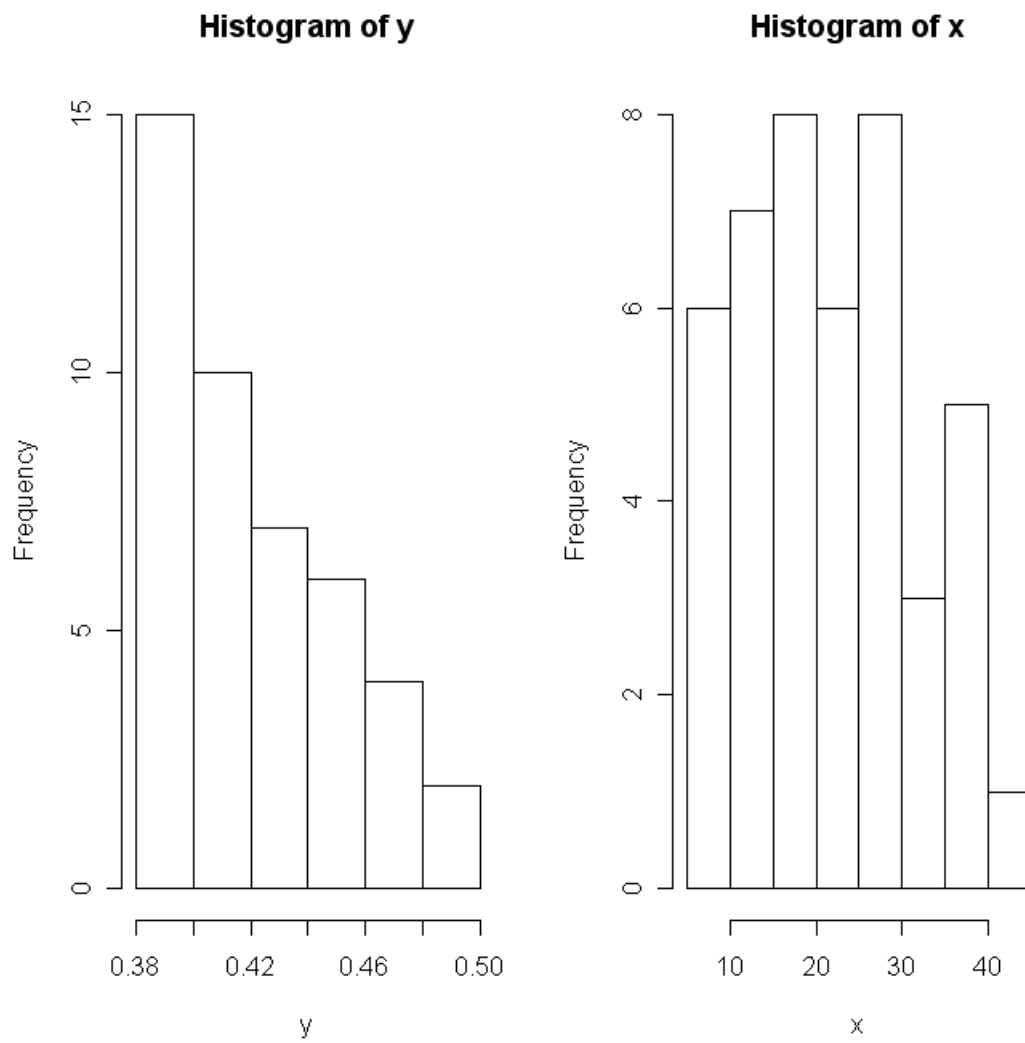


Figura 1: Histogramas de las variables estudiadas

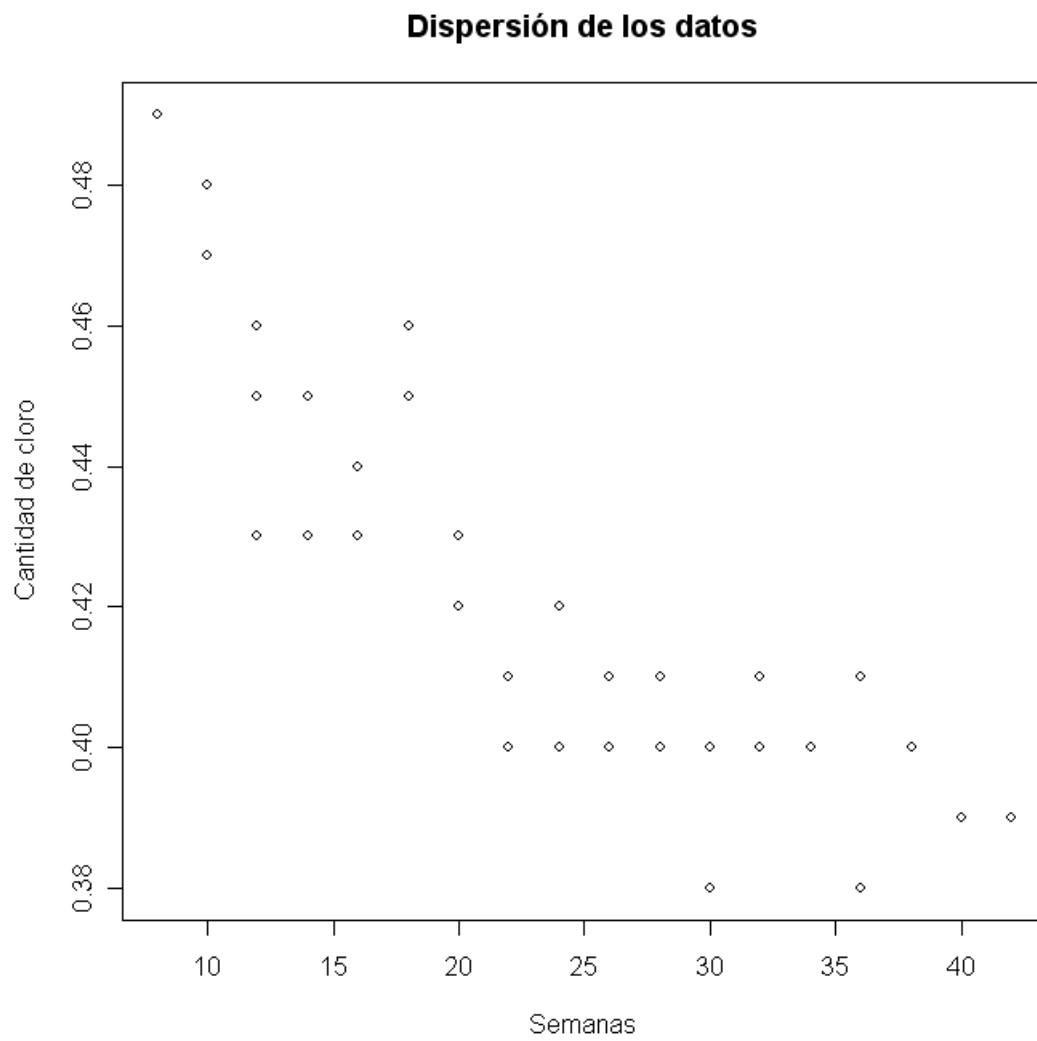


Figura 2: Grafico de dispersión

4.3. Ajuste del modelo y análisis de variancia

Como es sabido, ajustaremos un modelo lineal, para las variables x e y , de la forma

$$y_i = \beta_0 + \beta_1 x_i \quad (1)$$

pero bajo los parámetros estimados $\hat{\beta}_0$ y $\hat{\beta}_1$. Esto se logra a través de la siguiente instrucciones en R.

```
fp.fit <- lm(y~x, cloro)
```

Consecuencia de etos, ahora podemos obtener nuestro resumen del ajuste con la instrucción

```
summary(fp.fit)
```

Que produce, como resultado la siguiente tabla de resumen:

<u>Residuals:</u>					
Min	1Q	Median	3Q	Max	
-0.025741	-0.012042	-0.001608	0.012034	0.026224	
<u>Coefficients:</u>					
	Estimate	Std. Error	t value	<i>Pr</i> (> <i>t</i>)	
(Intercept)	0.4855103	0.0058907	82.42	< 2e-16	***
<i>x</i>	-0.0027168	0.0002431	-11.18	3.67e-14	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 0.01539 on 42 degrees of freedom					
Multiple R-Squared: 0.7483			Adjusted R-squared: 0.7423		
F-statistic: 124.9 on 1 and 42 DF			p-value: 3.675e-14		

Consideremos ahora la ejecución del comando

```
anova(fp.fit)
```

que produce la tabla ANOVA pertinente al modelo ajustado:

<u>Coefficients:</u>					
	DF	Sum Sq	Mean Sq	F value	Pr(> F)
x	1	0.0295587	0.0295587	124.88	3.675e-14 ***
Residuals	42	0.0099413	0.0002367		

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

De esta tabla podemos darnos cuentas de aspectos importantes sobre el modelo ajustado, como por ejemplo que la variable independiente x , es de hecho significativa en el modelo que obtuvimos el cual de acuerdo a la tabla, presenta la siguiente ecuación de regresión

$$y_i = 0,4855103 - (0,0027168)x_i \quad (2)$$

que nos permite darnos cuenta de que el número de semanas transcurridas afecta la cantidad de cloro en el agua; mas aún, siendo su p-valor 3.675e-14 y observando la estadística F calculada con un valor de 124.88, concluimos que debe rechazarse la hipótesis nula H_0 y por lo tanto podemos inferir que los valores se encuentran linealmente relacionados.

La siguiente prueba en la que se puede pensar seria establecer **intervalos de confianza** para los coeficientes de regresión, esto se logra, trabajando con un nivel de significación del 95 % con la siguiente orden:

```
confint(fp.fit, level = 0.95)
```

y obteniendo como resultado la siguiente salida

	2.5 %	97.5 %
(Intercept)	0.473622475	0.497398137
x	-0.003207415	-0.002226164

la cual no solo es consistente con los valores estimados de los coeficientes de regresión sino que tambien es consistente con la desición de rechazar la hipotesis nula, es decir, con el hecho de que hay una relación lineal entre las variables.

4.4. Coeficientes de determinación y coeficientes de correlación

Al ordenar lo siguiente:

`cor(cloro)`

obtenemos la matriz de coeficientes de correlación entre las variables implicadas en nuestro estudio, las cuales recordemos establece la magnitud de la pendiente para el modelo lineal de dos variables. Recordemos que éste coeficiente, se define y denota como

$$\rho(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 \sum_{i=1}^n (y_i - \bar{Y})^2}} ; -1 \leq \rho(x, y) \leq 1 \quad (3)$$

Para los datos de interes, la matriz tiene la forma

$$\begin{array}{cc} & x & y \\ x & 1 & -0.8650553 \\ y & -0.8650553 & 1 \end{array}$$

Dado que para y en función de x tenemos un valor cercano a -1 podemos concluir que la relación lineal es decreciente, es decir que mientras mas semanas pasen, menos cloro habrá en el agua.

En este punto parece apropiado mencionar el cálculo del **coeficiente de determinación** el cual nos da una idea de la precisión general del modelo ajustado. Este se define y denota de la siguiente manera

$$r^2 = 1 - \frac{\text{SCE}}{\text{STC}} ; 0 \leq r^2 \leq 1 \quad (4)$$

En nuestro caso, de acuerdo a la tabla podemos ver que $r^2 = 0,7423$, valor que, considerando que el caso de estudio en particular presenta muchos factores de variabilidad como por ejemplo el tipo de químico usado, los factores climáticos, entre otros, podemos concluir que el modelo se ajusta lo suficiente.

4.5. Estudio de los valores residuales y verificación de supuestos

Los valores residuales, a saber aquellos definidos, para un cierto valor y_i como

$$e_i = y_i - \hat{y}_i \quad (5)$$

Deben ser calculados y graficados pues esto representa un medio alternativo para corroborar la exactitud del modelo ajustado. Estos se calculan ordenando lo siguiente:

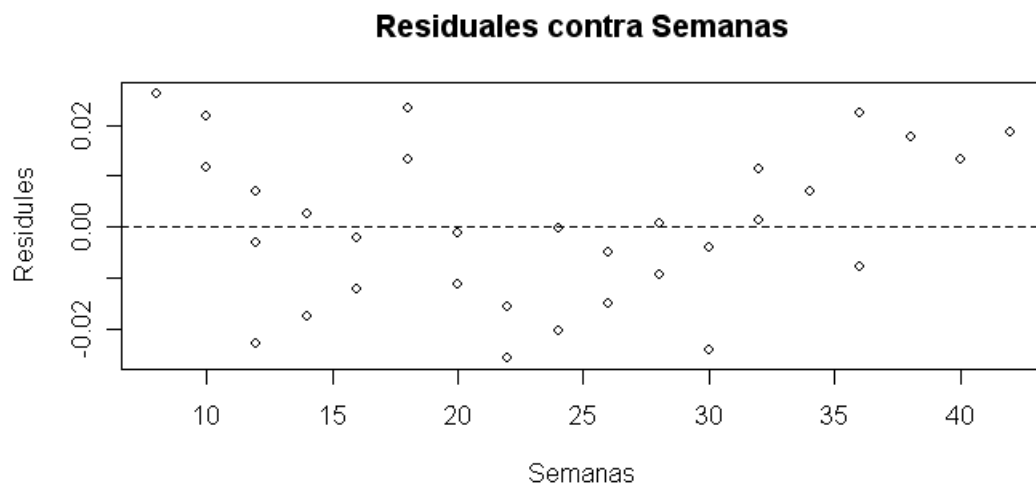


Figura 3: Grafico de residuales

```
clor.res<-residuals(fp.fit)
```

y se muestran en la figura 4. Recordemos que entre las suposiciones más importantes para este análisis, esta la condición de normalidad de la muestra. Esto puede verificarse con la siguiente orden, donde a su vez ordenamos la creación de la grafica.

```
qqnorm(traf.res, ylab="Residuales")  
qqline(traf.res)
```

Estos resultados se muestran en la figura ??, y como puede verse, los datos cumplen el supuesto de normalidad.

4.6. Predicciones en el modelo

Ya en este punto podemos estudiar las predicciones que permite hacer el modelo, que se reflejan en la figura 5. Este permite dar respuesta a las predicciones solicitadas:

1. Trancurridas las 15 semanas, la cantidad de cloro será ya de 0.44 ℓ .
2. Para obtener 0.37 ℓ de cloro deben pasar al menos semanas 40 semanas.

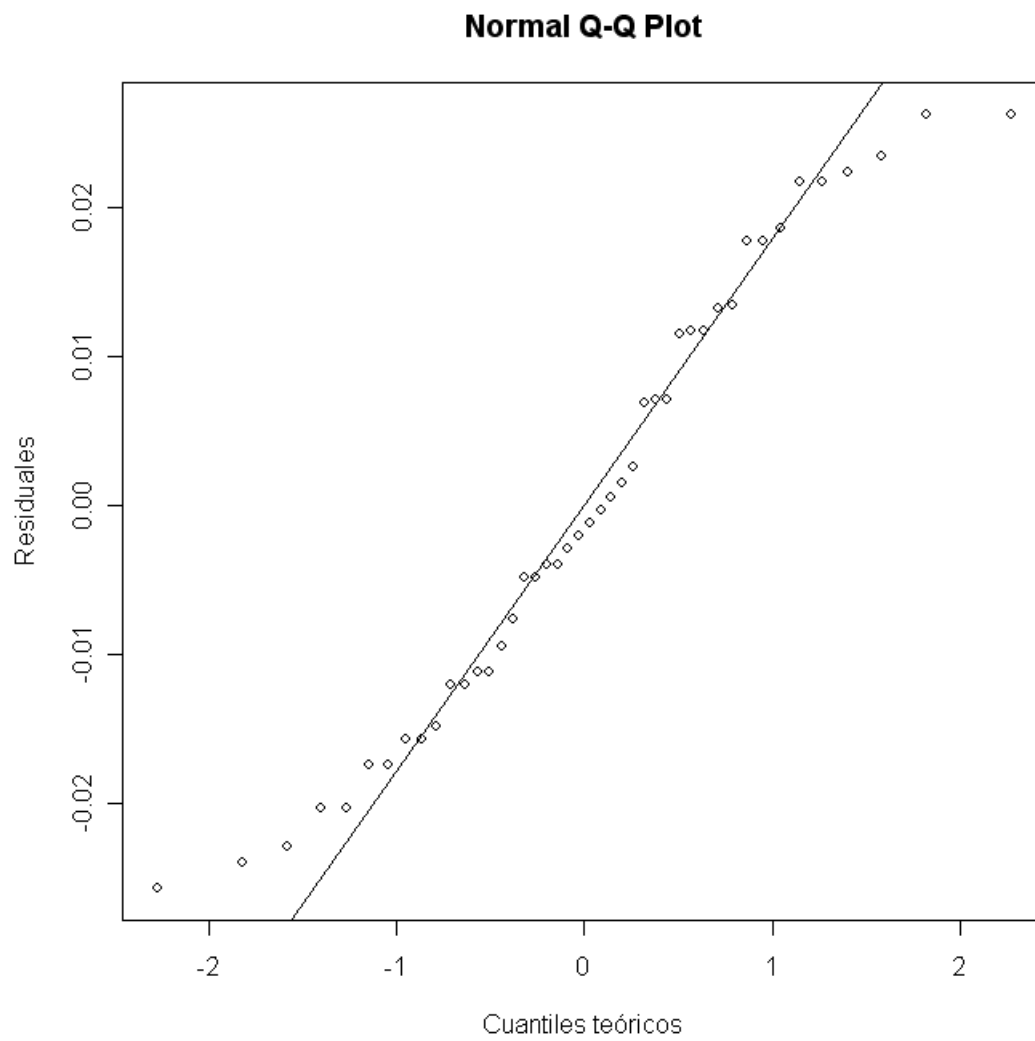


Figura 4: Grafico de la prueba de normalidad

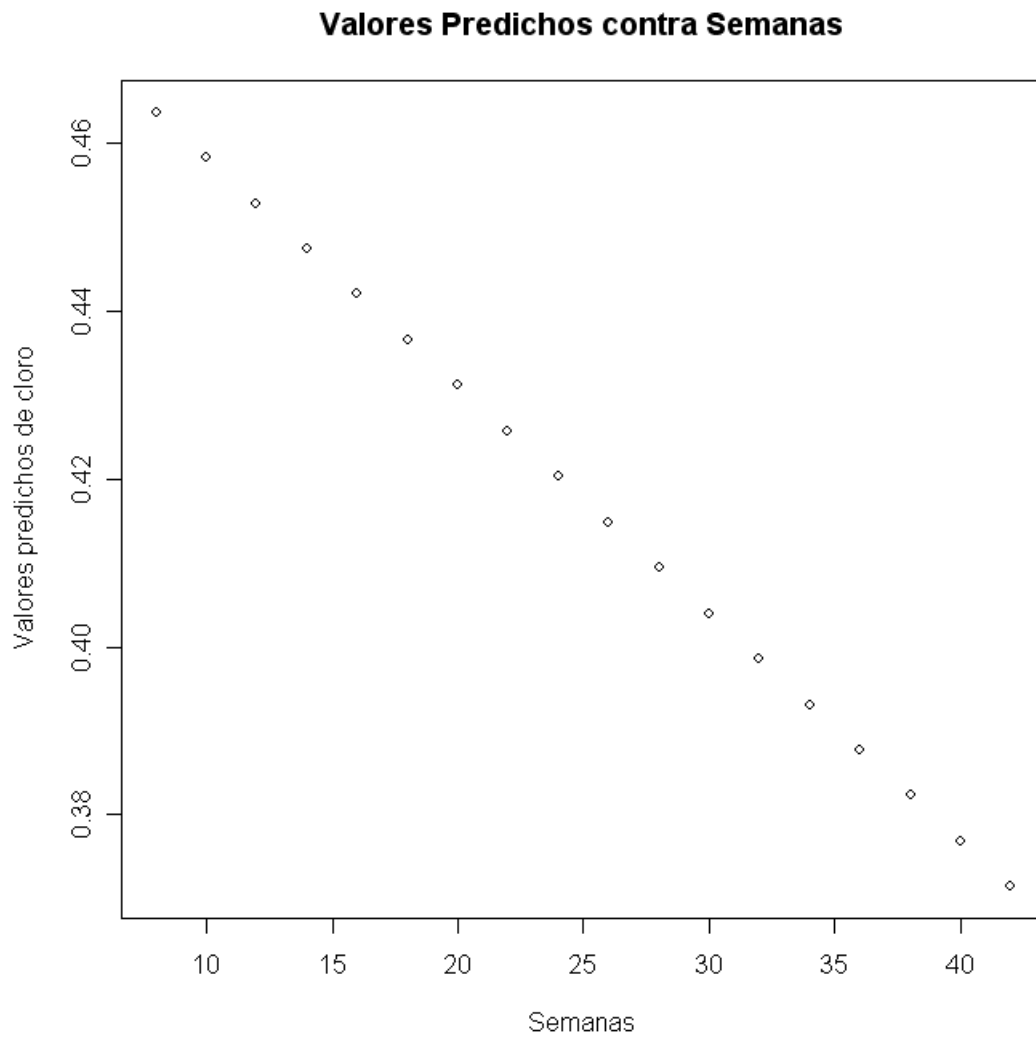


Figura 5: Grafico de predicciones del modelo

5. Ajuste de un modelo cuadratico

Para hacer un análisis comparativo, ordenaremos ajustar un modelo parabolico, a saber, uno de la forma

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \quad (6)$$

mediante las ordenes

```
fp.fit1 <- lm(y~x+I(x*x), cloro)
anova(fp.fit1)
```

lo que produce una tabla ANOVA generada por el comando que tiene la siguiente forma

Response: <i>y</i>						
	DF	Sum Sq	Mean Sq	F value	<i>Pr</i> (> <i>F</i>)	
<i>x</i>	1	0.0295587	0.0295587	226.331	2.2e-16	***
I(<i>x</i> * <i>x</i>)	1	0.0045868	0.0045868	35.121	5.508e-07	***
Residuals	41	0.0053546	0.0001306			

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

5.1. Estudio de los valores residuales y verificación de supuestos en el modelo parabolico

Estos se muestran en las figuras 6 y 7.

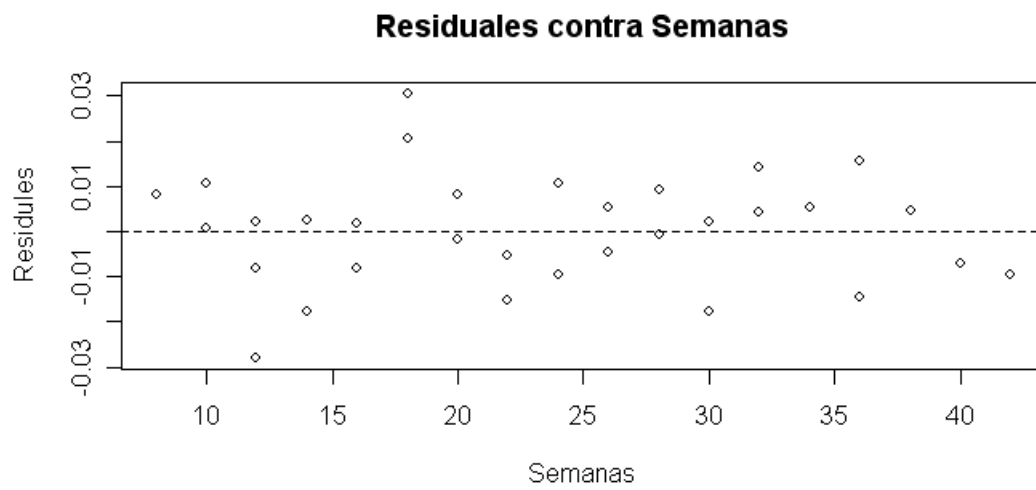


Figura 6: Grafico de los residuales para el modelo parabolico

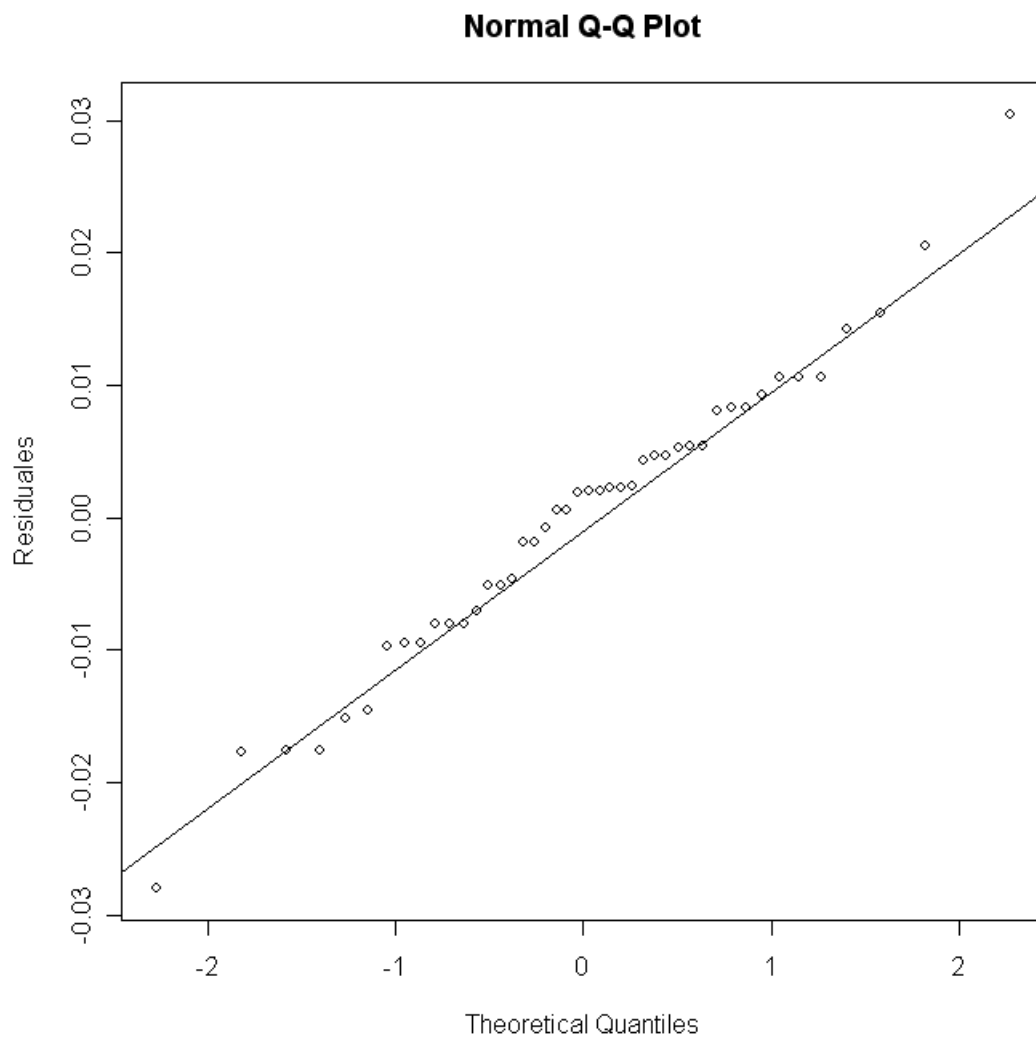


Figura 7: Grafico de la prueba de normalidad para el modelo parabolico

6. Conclusiones

Después de realizar el estudio del método y después de haberlo aplicado en el caso de prueba escogido podemos observar que su precisión siempre estuvo controlada directamente por el nivel de significación lo cual lo hace muy versátil y adaptable.

El caso de prueba aunque parece simple, en la realidad conlleva muchos factores a considerar y nuestras conclusiones fueron que al pasar el tiempo, las cantidades de cloro remanentes en el agua, consecuencias de la aplicación del proceso químico, van disminuyendo.

Referencias

- [1] Eduardo. J. Sánchez P. *Notas en Probabilidades y Estadísticas*. 2008.
- [2] The R Reference Manual.