ACTIVIDADES DEL TEMA 3: REGRESIÓN LINEAL

Alumno: Francisco Márquez

Planteamiento Actividad 1

Uno de los problemas con los que se encuentran las industrias químicas es el tratamiento de las aguas residuales. Dichas aguas son químicamente complejas puesto que se caracterizan por altos valores de gases tóxicos, sólidos volátiles y otras sustancias nocivas. Los datos siguientes se obtuvieron a partir de 33 muestras de aguas residuales tratadas en el Instituto Politécnico de Virginia. En dichos datos se muestran los porcentajes de reducción de gases tóxicos (Y) durante el proceso de depuración para distintos valores de reducción en los sólidos volátiles (X). El objetivo es determinar un modelo que permita predecir la disminución porcentual en la concentración de gases tóxicos conocida la disminución de solidos volátiles durante la depuración de las aguas residuales.

X	Y	X	Y	X	Y	X	Y
3	5	31	30	37	36	42	44
7	11	31	40	33	38	43	47
11	21	32	32	39	37	44	44
15	16	33	34	39	36	45	46
18	16	33	32	39	45	46	46
27	28	34	34	40	39	47	49
29	27	36	37	41	41	50	51
30	25	36	38	42	40		
30	36	36	34				

Tarea

Realizar un análisis de regresión simple sobre los datos, incluyendo el contraste de linealidad. Estudiar posibles datos atípicos y si afectan al ajuste realizado. ¿Afectan a alguna de las hipótesis del modelo? ¿Es admisible que la variabilidad de los residuos aumente o disminuya con la propia variable explicativa? Justica estas cuestiones.

Solución

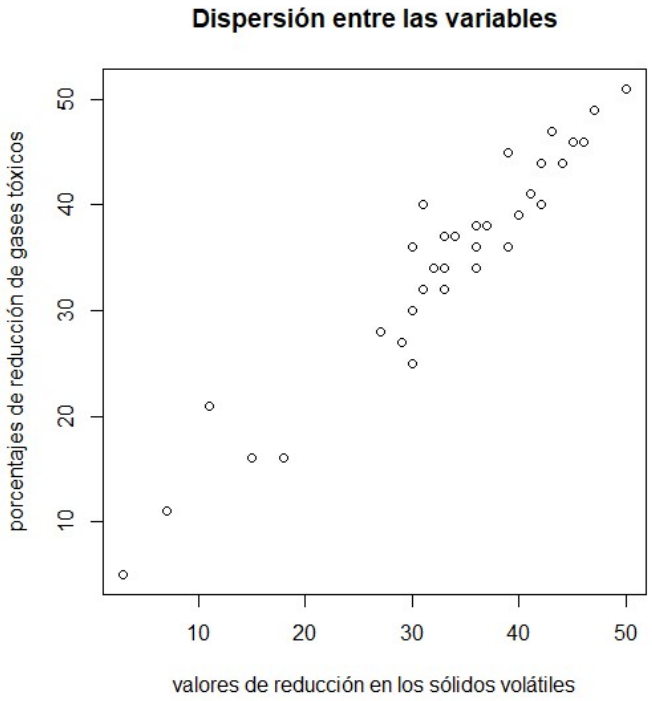
Para realizar el análisis voy a utilizar el software R. En primer lugar, examinamos visualmente la relación entre los *porcentajes de reducción de gases tóxicos* (Y) y los *valores de reducción en los sólidos volátiles* (X) a través de un gráfico de dispersión, esto lo hacemos con el fin de determinar si parece razonable una relación lineal entre las variables.

Sintaxis:

```
# grafico de las variables

plot(y~x,my_data, main="Dispersión entre las variables",xlab="valores de reducción en los sólidos volátiles",ylab="porcentajes de reducción de gases tóxicos")
```

Resultado:



El gráfico muestra una posible asociación lineal positiva entre las variables. El siguiente paso consiste en determinar el modelo que se ajusta de forma lineal entre las dos variables. Esto lo determinaremos al evaluar la hipótesis que la pendiente de la recta de regresión es igual a cero contra la alternativa de que la pendiente es distinta de cero (hipótesis de no linealidad entre X e Y).

Sintaxis:

```
# AOV de no linealidad - hipótesis de no linealidad entre X e Y

anova <- aov(modelo,my_data)

summary(anova)
```

Resultado:

```
            Df Sum Sq Mean Sq F value Pr(>F)
x            1   3564    3564   358.2 <2e-16 ***
Residuals   31    308      10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Con base en el resultado del contraste obtenemos un p-valor redondeado a cero (2e-16), lo cual es menor a 0.05. Esto indica que con un 95% de confianza podemos rechazar la hipótesis de que la pendiente es cero, en otras palabras, se sugiere una relación lineal significativa entre X y Y.

A continuación, construimos el modelo lineal usando la técnica del análisis de regresión.

Sintaxis:

```
#modelo de regresion

modelo <- lm(y~x,my_data)

summary(modelo)
```

Resultado:

```
Call:
lm(formula = y ~ x, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5712 -1.5989 -0.4751  1.2509  7.4973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.62650    1.71576   2.114   0.0427 *
x            0.93149    0.04921  18.927   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.154 on 31 degrees of freedom
Multiple R-squared:  0.9204,    Adjusted R-squared:  0.9178
F-statistic: 358.2 on 1 and 31 DF,  p-value: < 2.2e-16
```

Con base en los coeficientes obtenidos, podemos construir el siguiente modelo lineal

$$\mathbf{Y = 3.62650+0.93149*X}$$

Con un coeficiente de determinación R^2=0.92 esto indica que el 92% de la variabilidad de los *porcentajes de reducción de gases tóxicos* (Y) puede ser explicado por la variabilidad en los *valores de reducción en los sólidos volátiles* (X). A continuación haremos un análisis para validar que se cumplen los supuestos del modelo: normalidad, igualdad de varianzas (homocedasticidad) e independencia de los residuos.

Validación del modelo:

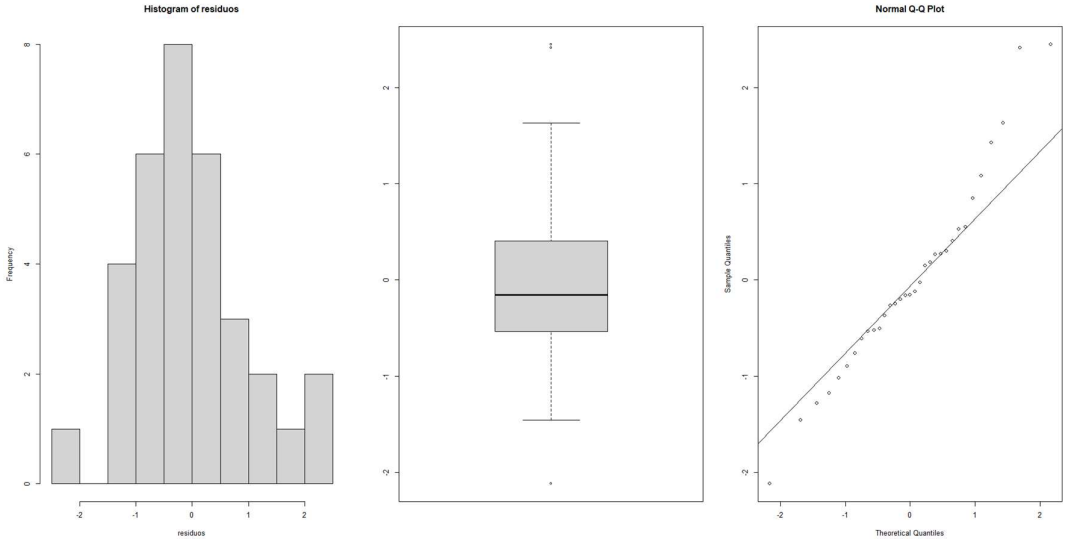
Normalidad:

Esta validación normalmente inicia analizando gráficos que nos puedan sugerir si la distribución de los residuos estandarizados presenta las características de una distribución normal. Para ello utilizamos el histograma, el gráfico de cajas y el gráfico de cuantiles.

Sintaxis:

```
res<-rstandard(modelo) # residuos estándar modelo ajustado
hist(residuos) # histograma residuos estandar
boxplot(residuos) # diagrama de cajas residuos estandar
qqnorm(residuos) # gráfico de cuantiles residuos estandar
qqline(residuos) # Linea de la distribución teórica Normal
```

Resultado:



Tanto el histrograma como el grafico de cajas sugieren un comportamiento de distribución que pudiera ajustar con la distribución normal, unimodal, forma de campana. En gráfico QQ desviaciones de los puntos del gráfico respecto de la diagonal indican alteraciones/desviaciones de la normalidad. En el gráfico, los puntos se aproximan en su mayoría a la diagonal lo que apoya la hipótesis de normalidad.

Igualdad de varianzas (Homocedasticidad):

Para validar este supuesto debemos comprobar que no exista una relación entre los Valores ajustados por el modelo y los residuos estandarizados. Una forma sencilla de comprobarlo es usando un grafico de dispersión con ambas variables.

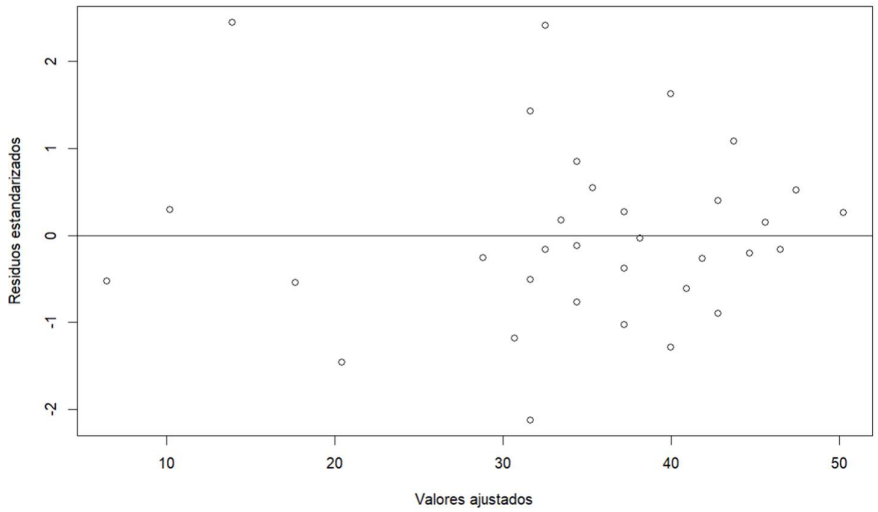
Sintaxis:

```
#  Varianza constante – La varianza de los errores  es constante

plot(fitted.values(modelo),rstandard(modelo),  xlab="Valores  ajustados",  ylab="Residuos
estandarizados")  #valores ajustados vs. residuos estandarizados

abline(h=0)
```

Resultado:



En vista de que no se observa ningún patrón que indique una posible relación entre estas variables podemos concluir que se cumple el supuesto de homocedasticidad.

Independencia de los residuos:

Para comprobar esta supuesto usaremos la prueba Durbin-Watson, cuya aplicación se basa en el valor del estadístico resultante de ejecutar el siguiente contraste de hipótesis:

H_0: no hay correlación entre los residuos

H_1: los residuos están autocorrelacionados

En nuestro caso ejecutamos la prueba a continuación.

Sintaxis:

```
#prueba Durbin-Watson

library(car)

durbinWatsonTest(modelo)
```

Resultados:

```
 lag Autocorrelation D-W Statistic p-value
   1       -0.186225      2.363832   0.388
 Alternative hypothesis: rho != 0
```

Dado que el p-valor es mayor a 0.05, con un 95% de confianza no podemos rechazar la hipótesis nula, en otras palabras, los residuos son independientes por no estar correlacionados.

¿Afectan a alguna de las hipótesis del modelo?

Sí, una significativa presencia de valores atípicos puede afectar a los tres supuestos que validan el modelo de regresión: normalidad, homocedasticidad e independencia de los residuos.

¿Es admisible que la variabilidad de los residuos aumente o disminuya con la propia variable explicativa?

No es admisible, puede sugerir presencia de autocorrelación. Se tendrían que realizar tareas adicionales de procesamiento dependiendo del tipo de correlación.

Planteamiento Actividad 2

Una industria automovilística desea conocer el promedio de vehículos por persona (Y) en una serie de países, en función de su densidad de población (X1), Renta per Cápita (X2), Precio del Litro de Gasolina (X3), Toneladas de Gasolina Consumida (X4) y Promedio de Kilómetros de Transporte Público Usados por Persona (X5). Se han obtenidos los siguientes datos:

Y	X_1	X_2	X_3	X_4	X_5
0.27	89	7.7	49	1.1	2.6
0.33	323	9.8	59	1	1.6
0.42	2	8.7	17	2.8	0.1
0.28	119	11	56	1.2	1.9
0.24	16	7.1	49	1.2	2.2
0.33	97	8.8	61	1	1.5
0.35	247	10.4	49	1.1	1.7
0.08	71	3.4	56	1.7	0.7
0.34	2	9.8	57	1.2	2
0.20	46	3.8	40	1.5	0.3
0.30	188	4.6	61	0.6	1.8
0.18	1309	8.5	49	1.2	3.5
0.43	138	9.8	44	1.6	0.8
0.30	412	9.4	56	1	1.5
0.40	12	5.9	34	1.3	0.2
0.28	13	9.8	61	1	1.7
0.10	107	1.8	68	0.7	0.9
0.18	73	4	44	0.8	1.3
0.34	18	10.6	42	1.3	1.7
0.32	153	13.3	56	1.3	2
0.014	55	1.2	36	3.3	0.1
0.27	229	5.5	35	1.2	1.6
0.53	23	9.7	17	2.7	0.3
0.09	86	2.1	40	1.1	2.1

Tarea

Con los datos obtenidos estudia si existe un patrón permita predecir el promedio de vehículos por habitante en un país.

Solución

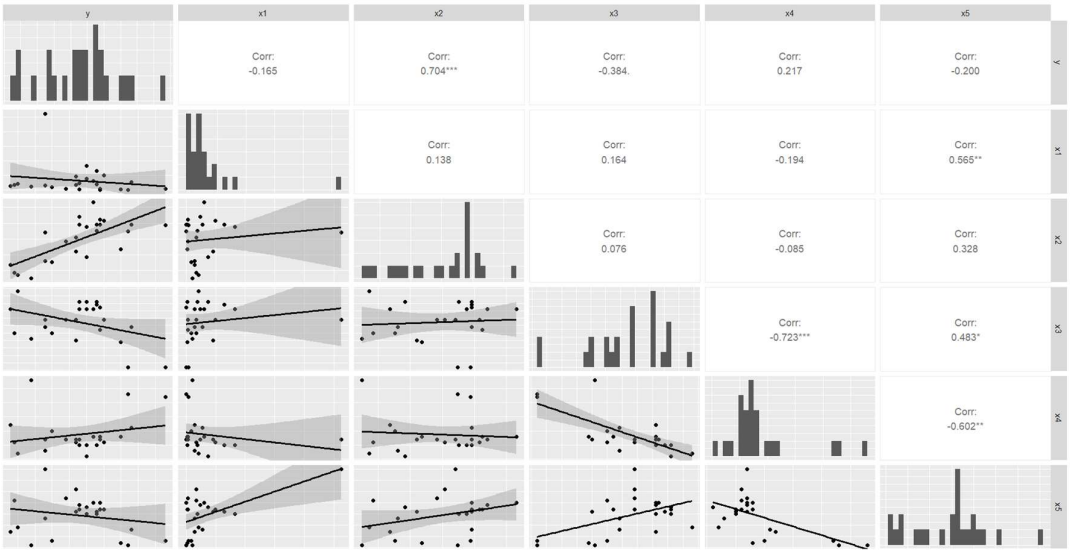
Lo primero que haremos es estudiar la relación que existe entre las variables. Esta información nos servirá para identificar cuáles pueden ser los mejores predictoresen el modelo, qué variables se relacionan linealmente (no serán incluidas) y para identificar colinealidad entre predictores.

En la solución de esta tarea usaremos el software R. Examinemos las correlaciones y los histogrogramas de las variables en estudio.

Sintaxis:

```
library(GGally)
ggpairs(my_data2, lower = list(continuous = "smooth"),diag = list(continuous = "barDiag"),
axisLabels = "none")
```

Resultado:



Del examen preliminar podemos observar lo siguiente:

Las variables que tienen una mayor relación lineal con promedio de vehículos por persona (Y) son: Renta per Cápita (X2) y Precio del Litro de Gasolina (X3), ellas a su vez no están correlacionadas por lo que puede ser útil introducir ambos predictores en el modelo.

Modelo

Para la construcción del modelo, hay diferentes formas de llegar a la mejor versión final. En este caso emplearemos el método mixto, iniciando el modelo con todas las variables como predictores y realizando la selección de los mejores la medición Akaike(AIC).

Sintaxis:

```
#modelo de regresion
modelo2 <- lm(y~x1+x2+x3+x4+x5,my_data2)
summary(modelo2)
```

Resultado:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = my_data2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.074595 -0.041992 -0.003251  0.037209  0.136575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.052e-01  1.001e-01   4.047 0.000757 ***
x1          -1.022e-05  5.549e-05  -0.184 0.855912
x2           2.905e-02  3.815e-03   7.614 4.92e-07 ***
x3          -3.937e-03  1.308e-03  -3.011 0.007506 **
x4          -5.203e-02  2.922e-02  -1.781 0.091824 .
x5          -5.672e-02  2.256e-02  -2.514 0.021652 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05681 on 18 degrees of freedom
Multiple R-squared:  0.8029,    Adjusted R-squared:  0.7482
F-statistic: 14.67 on 5 and 18 DF,  p-value: 8.249e-06
```

El modelo con todas las variables como predictores, tiene un R^2 =0.8029, lo cual es un valor alto y se interpreta como que el 80.29% de la variabilidad observada en promedio de vehículos por persona puede ser explicado por las variables iniciales. El p-value=8.249e-06 sugiere a un 95% de confianza, que al menos uno de los coeficientes parciales de regresión es distinto de 0 o que el modelo no es por azar. Muchos de ellos no son significativos, lo que es un indicativo de que podrían no contribuir al modelo.

Para la elección de los predictores, usamos la técnica paso a paso (stepwise) mixto, basándonos en el valor del Akaike (AIC).

Sintaxis:

```
# seleccion de predictores
step(object = modelo2, direction = "both", trace = 1)
```

Resultado:

```
Start:  AIC=-132.57
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq      RSS     AIC
- x1    1  0.000110 0.058199 -134.53
<none>              0.058090 -132.57
- x4    1  0.010235 0.068324 -130.68
- x5    1  0.020402 0.078492 -127.35
- x3    1  0.029256 0.087345 -124.78
- x2    1  0.187102 0.245191 -100.01

Step:  AIC=-134.53
y ~ x2 + x3 + x4 + x5

       Df Sum of Sq      RSS     AIC
<none>              0.058199 -134.53
+ x1    1  0.000110 0.058090 -132.57
- x4    1  0.010945 0.069144 -132.39
- x3    1  0.029197 0.087396 -126.77
- x5    1  0.032691 0.090891 -125.83
- x2    1  0.189791 0.247990 -101.74

Call:
lm(formula = y ~ x2 + x3 + x4 + x5, data = my_data2)

Coefficients:
(Intercept)           x2           x3           x4           x5
   0.407465     0.029119    -0.003932    -0.052977    -0.059076
```

De este resultado vemos que el mejor modelo resultante ha sido: Y= x2 + x3 + x4 + x5. Por ello construimos el modelo con estas variables y evaluamos sus métricas:

Sintaxis:

```
#modelo de regresion v2
modelo3 <- lm(y~x2+x3+x4+x5,my_data2)
summary(modelo3)
```

Resultados:

```
Call:
lm(formula = y ~ x2 + x3 + x4 + x5, data = my_data2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.074851 -0.039661 -0.005576  0.036890  0.136577

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.407465   0.096792   4.210 0.000475 ***
x2           0.029119   0.003699   7.871 2.13e-07 ***
x3          -0.003932   0.001274  -3.087 0.006063 **
x4          -0.052977   0.028026  -1.890 0.074070 .
x5          -0.059076   0.018083  -3.267 0.004057 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05535 on 19 degrees of freedom
Multiple R-squared:  0.8026,    Adjusted R-squared:  0.761
F-statistic: 19.31 on 4 and 19 DF,  p-value: 1.746e-06
```

Aún cuando el R^2 prácticamente no tuvo variación vemos como el R^2 ajustado si tuvo una mejora. Teniendo esto como resultado, el modelo obtenido es el siguiente:

promedio de vehículos por persona (Y) =
0.40746478
+0.02911871*Renta per Cápita (X2)
-0.00393221*Precio del Litro de Gasolina (X3)
-0.05297692*Toneladas de Gasolina Consumida (X4)
-0.05907625*Promedio de Kilómetros de Transporte Público Usados por Persona (X5)

Validación del modelo

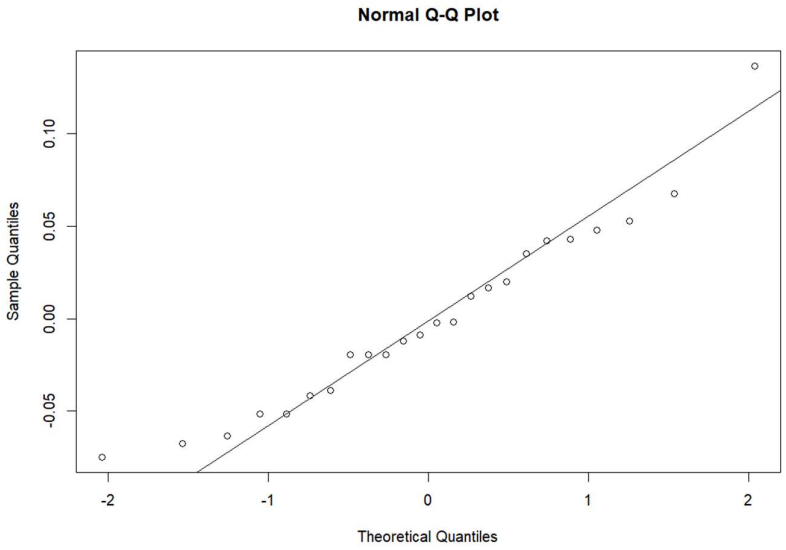
Normalidad:

Nos basamos en la forma del gráfico QQ.

Sintaxis:

```
# Normalidad
qqnorm(modelo3$residuals)
qqline(modelo3$residuals)
```

Resultado:



Se puede ver claramente como la mayoría de los puntos se ubican próximos a la línea de la distribución teórica. Por lo que se valida el supuesto.

Igualdad de varianzas (Homocedasticidad):

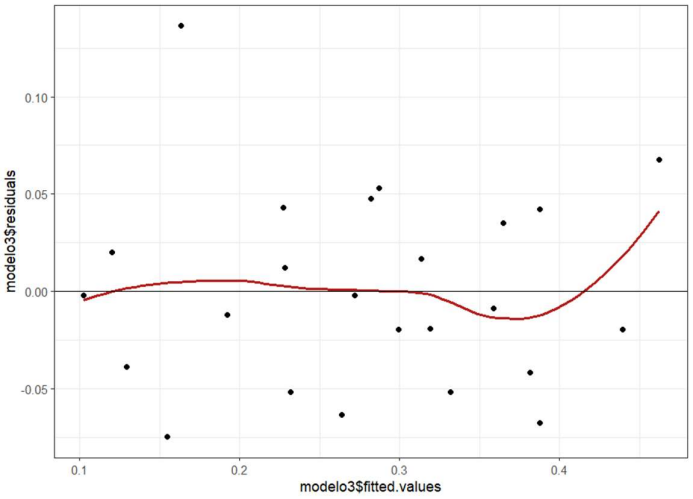
Graficamos los residuos vs los valores ajustados por el modelo, para comprobar que se distribuyen aleatoriamente en torno a cero, con la misma variabilidad a lo largo del eje X.

Sintaxis:

```
#Homocedasticidad

ggplot(my_data2, aes(modelo3$fitted.values, modelo3$residuals)) +
  geom_point() +
  geom_smooth(color = "firebrick", se = FALSE) +
  geom_hline(yintercept = 0) +
  theme_bw()
```

Resultado:



La forma como se distribuyen los valores a lo largo del eje, sin un patrón aparente, indica que el supuesto se cumple para el modelo.

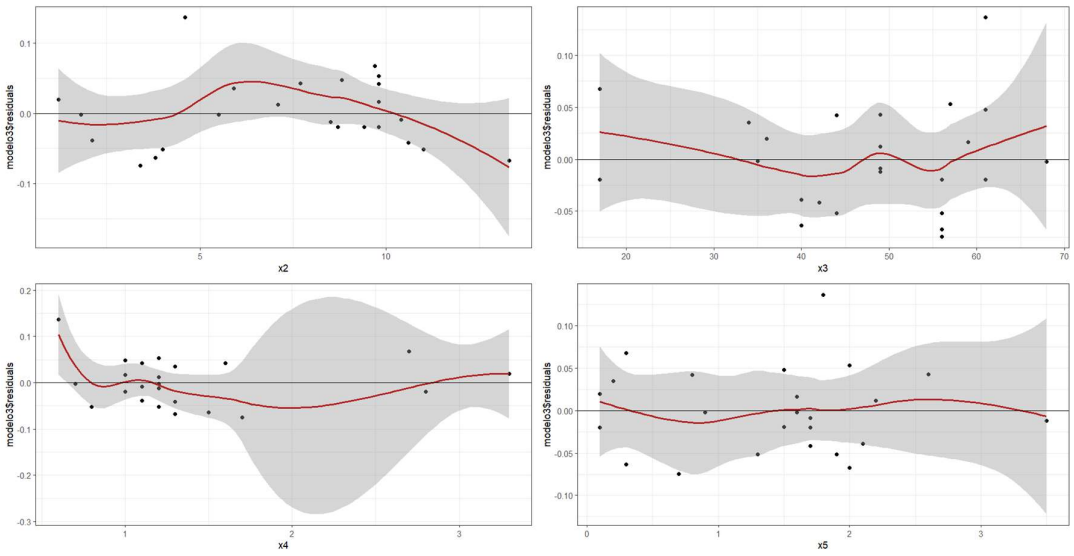
Linealidad entre variable respuesta y los predictores

Validaremos este supuesto con diagramas de dispersión entre cada uno de los predictores y los residuos del modelo. Si la relación es lineal, los residuos deben de distribuirse aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje X.

Sintaxis:

```
plot1 <- ggplot(data = my_data2, aes(x2, modelo3$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot2 <- ggplot(data = my_data2, aes(x3, modelo3$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot3 <- ggplot(data = my_data2, aes(x4, modelo3$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot4 <- ggplot(data = my_data2, aes(x5, modelo3$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
grid.arrange(plot1, plot2, plot3, plot4)
```

Resultado:



Podemos ver que para las cuatro variables, se cumple el supuesto de linealidad.