Universidad de Granada
Escuela Internacional de Posgrado
Máster en Estadística Aplicada
Materia: Diseño estadístico experimental y control de calidad. Aplicaciones en Biociencias e ingeniería.
Alumno: Francisco Javier Marquez Rosales

ACTIVIDADES DE RECUPERACION:

Tema 2. Análisis de la varianza de una vía

Tema 3. Regresión lineal

Tema 4. Diseño por bloques aleatorizado

Septiembre, 2022

Tema 2. Análisis de la varianza de una vía

Planteamiento Actividad 2.1

Se desea comprobar si ciertos cambios en un proceso de fabricación aumentan la calidad del producto. Para ello se comparan los resultados con el método tradicional (A) frente a los obtenidos por los procedimientos que se desean probar (B y C). Los datos corresponden a una medida de calidad del proceso.

A	B	C
32	40	37
44	46	30
31	33	28
35	29	33
33	35	37
33	32	39

Tarea

Comprobar si existen diferencias entre los tres tratamientos. En caso de existir diferencias entre los tratamientos, determinar de cuál de ellos proviene. Estudiar la validez del modelo, es decir, que los residuos sean normales e independientes y la varianza constante.

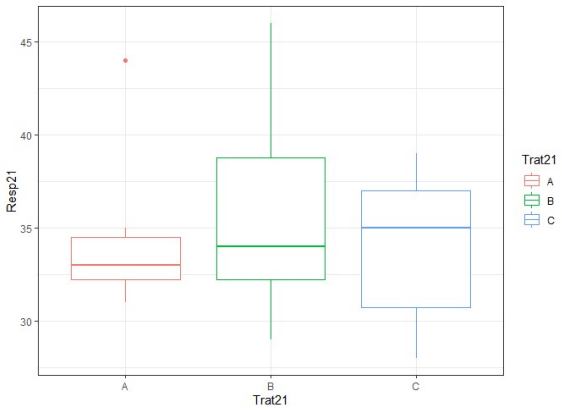
Solución

Para realizar el análisis se utilizará el software R. En primer lugar, examinamos visualmente la distribución de los datos en los tres métodos de fabricación basados en un gráfico de cajas.

```
### construcción de los datos
Resp21<- c(32,44,31,35,33,33,40,46,33,29,35,32,37,30,28,33,37,39)
Trat21<-c("A","A","A","A","A","A","B","B","B","B","B","B","C","C","C","C","C","C")

## el data frame
dat_21<-data.frame(Resp21,Trat21)
```

```
trat_21<-factor(Trat21)

#Determinacion de los factores
trat21f<-factor(trat_21)

#analisis los datos
require(ggplot2)
ggplot(data = dat_21, aes(x = Trat21, y = Resp21, color = Trat21)) +
  geom_boxplot() +
  theme_bw()
```

Resultado



El gráfico nos muestra que la dispersión de los datos para el método de fabricación A es menor que para B y C al igual que, de forma leve, el valor del promedio de la medida de calidad. Lo anterior indica inicialmente que el método de fabricación A ofrece resultados más estables (menos dispersos) y en donde la media de calidad es un poco menor.

Validación de los supuestos

Supuesto 1: Independencia

Por la forma como está descrita la recolección de los datos en el ejercicio, asumimos que fueron recolectados en forma aleatoria.

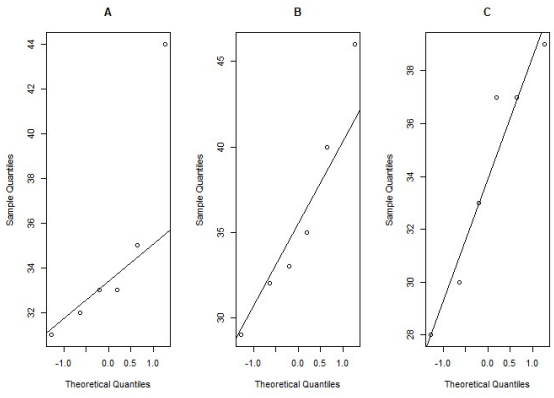
Supuesto 2: Distribucion normal

Debemos comprobar si la distribución que siguen los datos de fabricación de los tres métodos se pudiera considerar como 'Distribución Normal'. Para ello generamos los respectivos gráficos QQ y la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es *H0: los datos siguen una distribución normal*.

En este caso la prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

```
par(mfrow = c(1,3))
qqnorm(dat_21[dat_21$Trat21 == "A","Resp21"], main = "A")
qqline(dat_21[dat_21$Trat21 == "A","Resp21"])
qqnorm(dat_21[dat_21$Trat21 == "B","Resp21"], main = "B")
qqline(dat_21[dat_21$Trat21 == "B","Resp21"])
qqnorm(dat_21[dat_21$Trat21 == "C","Resp21"], main = "C")
qqline(dat_21[dat_21$Trat21 == "C","Resp21"])
par(mfrow = c(1,1))

#test de normalidad (menos de 50 observaciones usamos el test Shapiro - Wilk)
hist(dat_21$Resp21)
shapiro.test(dat_21$Resp21)
```

Resultado




```
Shapiro-Wilk normality test

data:  dat_21$Resp21
W = 0.92699, p-value = 0.1719
```

El gráfico nos muestra como la nube de puntos se ajusta a lo largo de la recta del modelo normal teórico, lo cual sugiere normalidad. El resultado del test Shapiro-Wilk ofrece un p-value del 0.17 que al hacer la prueba al 95% de confianza no tenemos evidencia para rechazar la hipótesis nula, por lo que aceptamos que los datos tienen una distribución normal.

Supuesto 3: homocedasticidad o varianza constante entre grupos

Utilizaremos la prueba Barlett para evaluar la homocedasticidad (homogeneidad de varianza). Esta prueba no mantiene sensibilidad frente al supuesto de normalidad que acabamos de comprobar. La hipótesis nula de esta prueba es *H0: los datos presentan homogeneidad de varianza entre los grupos*.

```
#supuesto 3: homocedasticidad o varianza constante entre grupos
bartlett.test(Resp21~Trat21,dat_21)
```

Resultado

```
Bartlett test of homogeneity of variances
data:  Resp21 by Trat21
Bartlett's K-squared = 0.61404, df = 2, p-value = 0.7356
```

El resultado del p-valor indica que no hay evidencias significativas de falta de homocedasticidad. De esta forma hemos comprobado los supuestos necesarios para ejecutar el ANOVA.

Análisis de Varianza

Para comprobar si existen diferencias entre los tres tratamientos. Ahora aplicaremos el ANOVA. Este contrate plantea como hipótesis nula *Ho: las medias de los tratamientos son iguales*.

```
modelo21 <- lm(Resp21~trat21f)
ANOVA21 <- aov(modelo21)
RESUMEN_ANOVA21 <- summary(ANOVA21)
RESUMEN_ANOVA21
```

Resultado

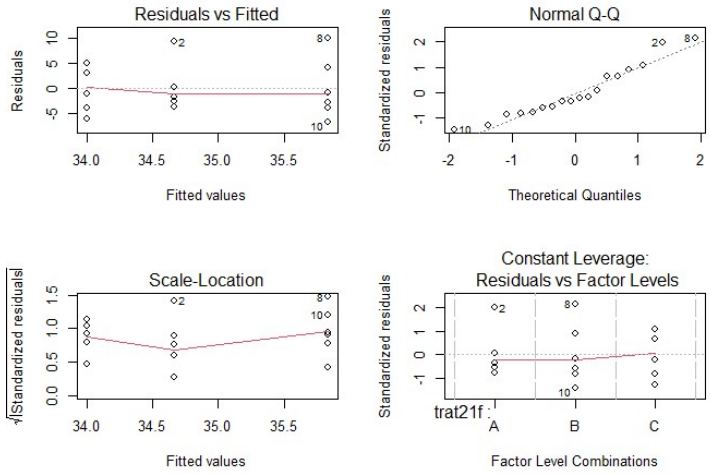
```
            Df Sum Sq Mean Sq F value Pr(>F)
trat21f      2   10.3   5.167   0.194  0.826
Residuals   15  400.2  26.678
```

Con base en este resultado, no tenemos suficiente evidencia para rechazar la hipótesis nula, por tanto concluimos que no existen diferencias significativas en los resultados de la calidad de los tres métodos de producción. Dado este resultado, no es válido estudiar el efecto particular de los tratamientos.

Validez del modelo

Estudiamos la validez del modelo analizando los residuos. Evaluamos que los mismos sean normales e independientes y la varianza constante. Para ello usamos la opción '*plot*' de R sobre el modelo obtenido con la siguiente sintaxis:

```
par(mfrow = c(2,2))
plot(modelo21)
par(mfrow = c(1,1))
```



La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residual vs fitted) y en el Gráfico NormalQ-Q (qqplot) los residuos se distribuyen muy cercanos a la linea de la normal, lo que indica un comportamiento normal. Estos valores nos permiten confirmar la validez del modelo.

Planteamiento Actividad 2.2

Se comparan las emisiones de distintas empresas de C02, para ello se miden las emisiones a la atmósfera de cuatro empresas, obteniendo los siguientes resultados:

Empresa 1	Empresa 2	Empresa 3	Empresa 4
32	54	30	61
45	23	23	34
68	33	41	38
29	35	42	39
41	11	31	28
37		37	29
78		22	
		45	

Tarea

Existen diferencias entre las emisiones de las diferentes empresas. En caso de existir, ¿de dónde proceden? Estudiar la validez del modelo.

Solución

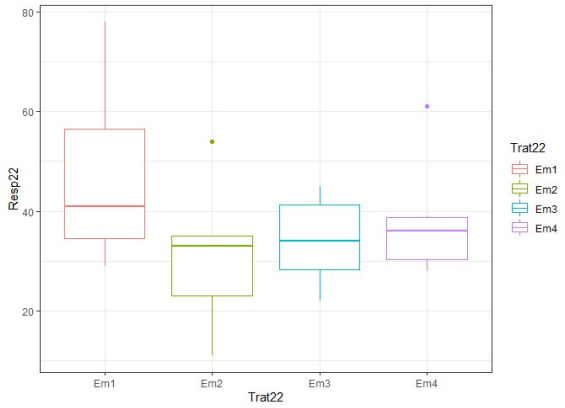
Para la solución de este problema usaremos el software R. Iniciamos haciendo la carga de los datos y la identificación de los factores que influyen en el análisis.

```
### construccion de los datos
Resp22<-
c(32,45,68,29,41,37,78,54,23,33,35,11,30,23,41,42,31,37,22,45,61,34,38,39,28,29)
Trat22<-c("Em1","Em1","Em1","Em1","Em1","Em1","Em1","Em2","Em2","Em2","Em2","Em2","Em3",
          "Em3","Em3","Em3","Em3","Em3","Em3","Em4","Em4","Em4","Em4","Em4","Em4")

## el data frame
dat_22<-data.frame(Resp22,Trat22)
trat_22<-factor(Trat22)

#Determinacion de los factores
trat22f<-factor(trat_22)
```

Estudiamos la distribución de los datos, sus valores medios y su dispersión los cual nos permitirá saber si el análisis planteado es razonable. Para ello utilizamos el gráfico de cajas

```
#analisis los datos
require(ggplot2)
ggplot(data = dat_22, aes(x = Trat22, y = Resp22, color = Trat22)) +
  geom_boxplot() +
  theme_bw()
```

Resultado



Examinando el gráfico podemos observar inicialmente como la emisión de CO2 de la empresa 1 tiene una dispersión mayor con relación a las otras empresas, por ello tendremos que comprobar el supuesto de homocedasticidad. La emisión de Co2 producida por la empresa 2 presenta una ligera asimetría. De igual forma la media de la empresa 1 parece tener una diferencia con relación a las medias de las otras empresas.

Procedemos entonces a validar los supuestos.

Validación de los supuestos

Supuesto 1: Independencia

Por la forma como está descrita la recolección de los datos en el ejercicio, asumimos que fueron recolectados en forma aleatoria.

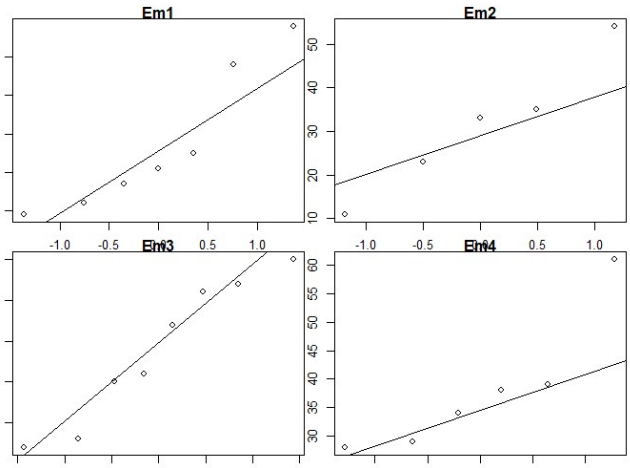
Supuesto 2: Distribución normal

Debemos comprobar si la distribución de las emisiones de CO2 de las 4 empresas se pueden considerar como 'Distribución Normal'. Para ello generamos los respectivos gráficos QQ y la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es *H0: los datos siguen una distribución normal*.

En este caso la prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

```
#supuesto 2: distribucion normal
par(mar=c(1,1,1,1)) #cambio de margenes para que ajusten los 4 graficos
par(mfrow = c(2,2))
qqnorm(dat_22[dat_22$Trat22 == "Em1","Resp22"], main = "Em1")
qqline(dat_22[dat_22$Trat22 == "Em1","Resp22"])
qqnorm(dat_22[dat_22$Trat22 == "Em2","Resp22"], main = "Em2")
qqline(dat_22[dat_22$Trat22 == "Em2","Resp22"])
qqnorm(dat_22[dat_22$Trat22 == "Em3","Resp22"], main = "Em3")
qqline(dat_22[dat_22$Trat22 == "Em3","Resp22"])
qqnorm(dat_22[dat_22$Trat22 == "Em4","Resp22"], main = "Em4")
qqline(dat_22[dat_22$Trat22 == "Em4","Resp22"])
par(mfrow = c(1,1))

#test de normalidad (menos de 50 observaciones usamos el test Shapiro - Wilk)
shapiro.test(dat_22$Resp22)
```

Resultado




```
Shapiro-Wilk normality test

data:  dat_22$Resp22
W = 0.9223, p-value = 0.05092
```

Los gráficos nos muestran como las nubes de puntos se ajustan a lo largo de la recta del modelo normal teórico, lo cual sugiere normalidad. El resultado del test Shapiro-Wilk ofrece un p-value del 0.05092 que indica que al hacer la prueba al 95% de confianza no tenemos evidencia para rechazar la hipótesis nula, por lo que aceptamos que los datos tienen una distribución normal. Este resultado está al borde del criterio por que hay que tomar en cuenta la sensibilidad de los datos al considerarlos normal.

Supuesto 3: homocedasticidad o varianza constante entre grupos

Dado que se encuentra en el límite para aceptar que se distribuye de forma normal, la prueba de Fisher y la de Bartlett no son recomendables. En su lugar es mejor emplear una prueba basada en la mediana, por lo que emplearemos la prueba de Levene y la prueba de Fligner-Killeen. En ambas pruebas contrastamos la hipótesis nula *Ho: la varianza entre grupos es constante*.

```
require(car)
leveneTest(Resp22~Trat22,dat_22,center = "median")
fligner.test(Resp22~Trat22,dat_22)
```

Resultado

```
Levene's Test for Homogeneity of Variance (center = "median")
```

```
      Df F value Pr(>F)
group  3  0.5875 0.6296

Fligner-Killeen test of homogeneity of variances
data:  Resp22 by Trat22
Fligner-Killeen:med chi-squared = 1.7094, df = 3, p-value = 0.6349
```

Como resultado en ambas pruebas obtenemos valores p-value mayores a 0.05, esto indica que no tenemos evidencia para rechazar la hipótesis nula, por lo que concluimos que se cumple el criterio de homocedasticidad o igualdad de la varianza entre empresas.

Análisis de Varianza

Para comprobar si existen diferencias entre los tres tratamientos. Ahora aplicaremos el ANOVA. Este contrate plantea como hipótesis nula *Ho: las medias de las emisiones de CO2 de las empresas son iguales*.

```
modelo22 <- lm(Resp22~trat22f)
ANOVA22 <- aov(modelo22)
RESUMEN_ANOVA22 <- summary(ANOVA22)
RESUMEN_ANOVA22
```

Resultado

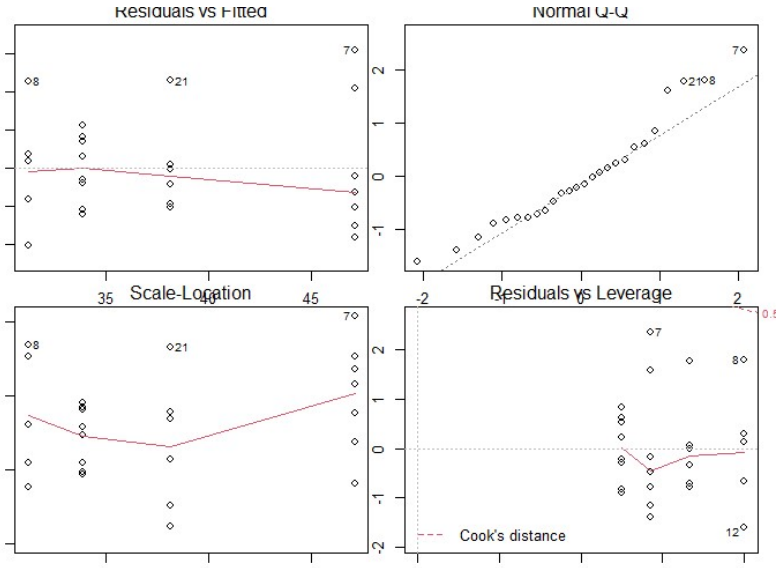
```
            Df Sum Sq Mean Sq F value Pr(>F)
trat22f      3    952   317.5   1.601  0.218
Residuals   22   4363   198.3
```

Con base en este resultado, no tenemos suficiente evidencia para rechazar la hipótesis nula, por tanto concluimos que no existen diferencias significativas en los resultados de emisión de Co2 de las cuatro empresas. Dado este resultado, no resulta útil estudiar el efecto individual de la emisión de CO2 de las empresas.

Validez del modelo

Estudiamos la validez del modelo analizando los residuos. Evaluamos que los mismos sean normales e independientes y la varianza constante. Para ello usamos la opción '*plot*' de R sobre el modelo obtenido con la siguiente sintaxis:

```
par(mfrow = c(2,2))
plot(modelo21)
par(mfrow = c(1,1))
```



La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residual vs fitted) y en el Gráfico NormalQ-Q (qqplot) los residuos se distribuyen muy cercanos a la linea de la normal, lo que indica un comportamiento normal. Estos valores nos permiten confirmar la validez del modelo.

Planteamiento Actividad 2.3

Una empresa produce telas en diferentes telares. Les gustaría que los telares fuesen lo más homogéneos posibles para obtener tejidos de la misma calidad, aunque el ingeniero sospecha que por la variación en la fuerza puede producir diferencias significativas entre las telas producidas en diferentes telares. Para investigar esto, se selecciona al azar, cuatro telares y hace cuatro determinaciones fuerza en el tejido fabricado en cada telar. Obteniendo la siguiente tabla:

TELAR	OBSERVACIONES			
1	98	97	99	96
2	91	90	93	92
3	96	95	97	95
4	95	96	99	98

Tarea

¿Está en lo cierto el investigador? Comprobar el modelo y determinar el origen de las posibles diferencias.

Solución

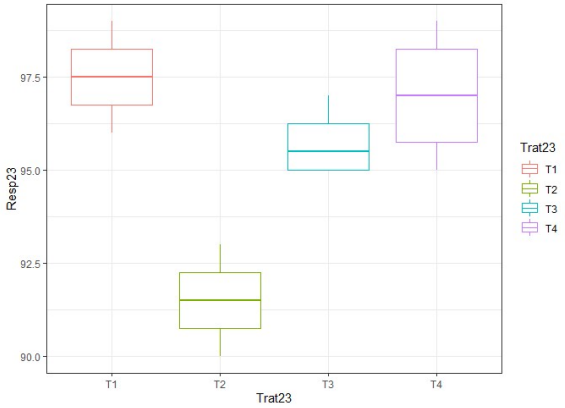
Para realizar el análisis se utilizará el software R. En primer lugar, examinamos visualmente la distribución de los datos en los tres métodos de fabricación basados en un gráfico de cajas.

Sintaxis:

```
### construcción de los datos
### construccion de los datos
Resp23<- c(98,97,99,96,91,90,93,92,96,95,97,95,95,96,99,98)
Trat23<-
c("T1","T1","T1","T1","T2","T2","T2","T2","T3","T3","T3","T3","T4","T4","T4","T4")

## el data frame.
dat_23<-data.frame(Resp23,Trat23)
trat_23<-factor(Trat23)
```

```
#Determinacion de los factores
trat23f<-factor(trat_23)

#analisis visual de los datos
require(ggplot2)
ggplot(data = dat_23, aes(x = Trat23, y = Resp23, color = Trat23)) +
  geom_boxplot() +
  theme_bw()
```

Resultado



El gráfico nos muestra que los valores medios de medida de tensión de los cuatro telares es diferente, en especial el telar 2. La dispersión de los datos para los cuales telares puede considerarse parecida. En ninguno de los casos se observa asimetría.

Validación de los supuestos

Supuesto 1: Independencia

Dado que los telares fuero seleccionados al azar para el regitro de las meiciones de fuerza, podemos considerar que los datos presentan la característica de independencia.

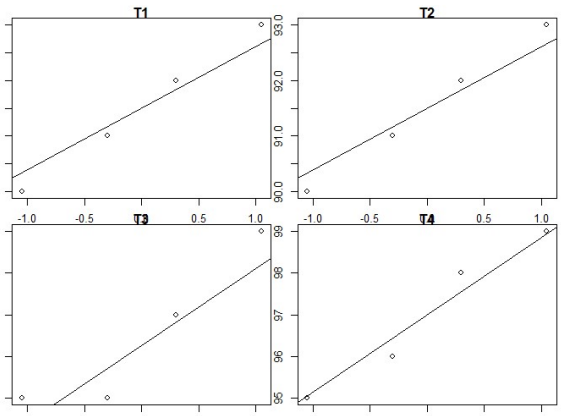
Supuesto 2: Distribución normal

Debemos comprobar si la medida de fuerza de los 4 telares empresas se pueden considerar que tienen una 'Distribución Normal'. Para ello generamos los respectivos gráficos QQ y la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es *H0: los datos siguen una distribución normal*.

En este caso la prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

```
#supuesto 2: distribucion normal

par(mar=c(1,1,1,1)) #cambio de márgenes para que ajusten los 4 gráficos
par(mfrow = c(2,2))
qqnorm(dat_23[dat_23$Trat23 == "T1","Resp23"], main = "T1")
qqline(dat_23[dat_23$Trat23 == "T1","Resp23"])
qqnorm(dat_23[dat_23$Trat23 == "T2","Resp23"], main = "T2")
qqline(dat_23[dat_23$Trat23 == "T2","Resp23"])
qqnorm(dat_23[dat_23$Trat23 == "T3","Resp23"], main = "T3")
qqline(dat_23[dat_23$Trat23 == "T3","Resp23"])
qqnorm(dat_23[dat_23$Trat23 == "T4","Resp23"], main = "T4")
qqline(dat_23[dat_23$Trat23 == "T4","Resp23"])
par(mfrow = c(1,1))

#test de normalidad (menos de 50 observaciones usamos el test Shapiro - Wilk)
shapiro.test(dat_23$Resp23)
```

Resultado



```
        Shapiro-Wilk normality test

data:  dat_23$Resp23
W = 0.93207, p-value = 0.2629
```

Los gráficos nos muestran como las nubes de puntos se ajustan a lo largo de la recta del modelo normal teórico, lo cual sugiere normalidad. El resultado del test Shapiro-Wilk ofrece un p-value del 0.26 lo que indica al 95% de confianza, que no tenemos evidencia para rechazar la hipótesis nula, por lo que aceptamos que los datos provienen de una distribución normal.

Supuesto 3: homocedasticidad o varianza constante entre grupos

Utilizaremos la prueba Barlett para evaluar la homocedasticidad (homogeneidad de varianza). Esta prueba no mantiene sensibilidad frente al supuesto de normalidad que acabamos de comprobar. La hipótesis nula de esta prueba es *H0: los datos presentan homogeneidad de varianza entre los grupos*.

```
bartlett.test(Resp23~Trat23,dat_23)
```

Resultado

```
Bartlett test of homogeneity of variances
data:  Resp23 by Trat23
Bartlett's K-squared = 1.1064, df = 3, p-value = 0.7755
```

El resultado del p-valor, p-value=0.73, indica que no hay evidencias significativas de falta de homocedasticidad. De esta forma hemos comprobado los supuestos necesarios para ejecutar el ANOVA.

Análisis de Varianza

Para comprobar si existen diferencias significativas entre la tensión de fuerza de los cuatro telares. Ahora aplicaremos el ANOVA. Este contraste plantea como hipótesis nula *Ho: las medias de la tensión de fuerza de los cuatro telares son iguales*.

```
            Df Sum Sq Mean Sq F value   Pr(>F)
```

```
trat23f      3  89.19  29.729   15.68 0.000188 ***
Residuals   12  22.75   1.896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado

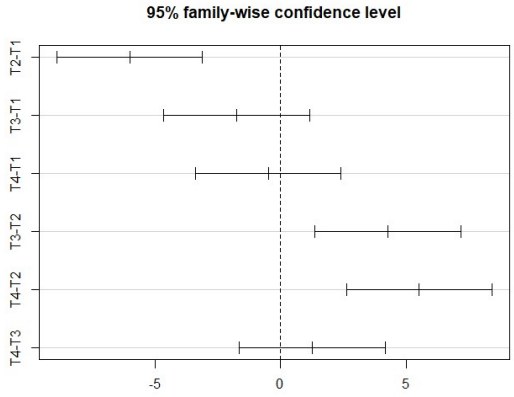
```
            Df Sum Sq Mean Sq F value   Pr(>F)
trat23f      3  89.19  29.729   15.68 0.000188 ***
Residuals   12  22.75   1.896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Con base en los datos analizados y con un 95% de confianza, no podemos aceptar la hipótesis nula, por tanto concluimos que existen diferencias significativas en las medidas de fuerza ejercidas por los cuatro telares. Dado este resultado, procedemos a investigar el origen de las posibles diferencias, para ello utilizaremos prueba de comparación múltiple de Tukey, llamada prueba TukeyHSD.

```
#comparacion multiple
TukeyHSD(anova23)
plot(TukeyHSD(anova23))
```

Resultado

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = modelo23)
$trat23f
       diff       lwr       upr     p adj
T2-T1 -6.00 -8.890552 -3.109448 0.0002455
T3-T1 -1.75 -4.640552  1.140552 0.3209518
T4-T1 -0.50 -3.390552  2.390552 0.9542581
T3-T2  4.25  1.359448  7.140552 0.0044029
T4-T2  5.50  2.609448  8.390552 0.0005377
T4-T3  1.25 -1.640552  4.140552 0.5894146
```



Con base en los resultados de la prueba y del gráfico obtenido observamos como se tienen diferencias significativas en las siguientes combinaciones de telares:

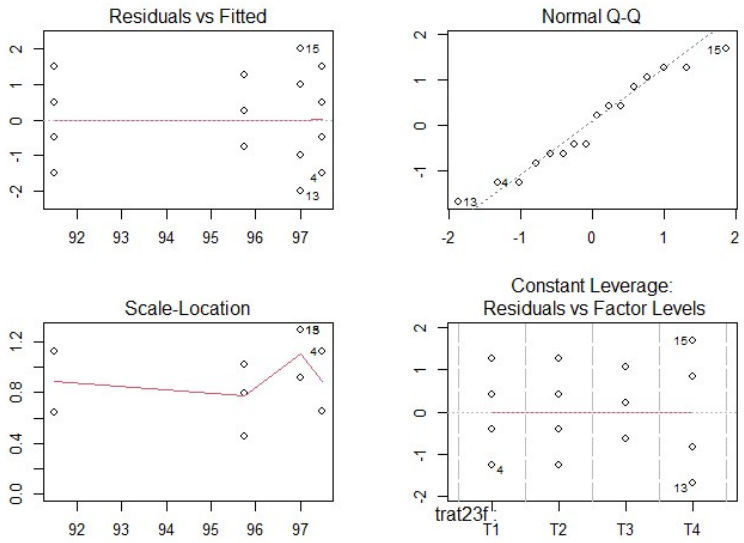
T1 con T2, T2 con T3 y T2 con T4

Debido a este resultado, podemos afirmar que la diferencia identificada por el ANOVA proviene principalmente de que las medidas del telar 2 (T2), en promedio, fueron significativamente menores al resto de los telares.

Validez del modelo

Estudiamos la validez del modelo analizando los residuos. Evaluamos que los mismos sean normales e independientes y la varianza constante. Para ello usamos la opción '*plot*' de R sobre el modelo obtenido con la siguiente sintaxis:

```
par(mfrow = c(2,2))
plot(modelo23)
par(mfrow = c(1,1))
```



La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residual vs fitted) y en el Gráfico NormalQ-Q (qqplot) los residuos se distribuyen muy cercanos a la linea de la normal, lo que indica un comportamiento normal. Estos valores nos permiten confirmar la validez del modelo.

Tema 3. Regresión Lineal

Planteamiento Actividad 3.1

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

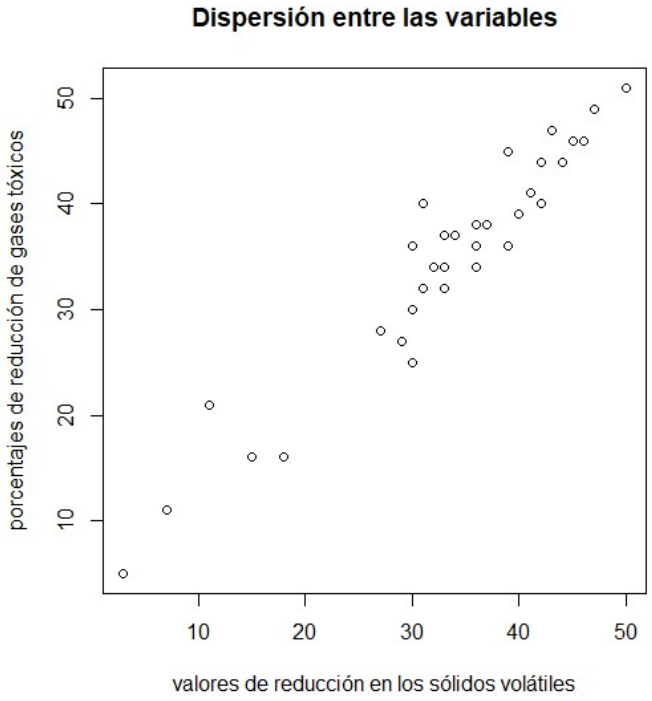
Realizar un análisis de regresión simple sobre los datos, incluyendo el contraste de linealidad. Estudiar posibles datos atípicos y si afectan al ajuste realizado. ¿Afectan a alguna de las hipótesis del modelo? ¿Es admisible que la variabilidad de los residuos aumente o disminuya con la propia variable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

Solución

Para realizar el análisis se utilizará 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.

```
# grafico de las variables

plot(Y31~X31,dat_31, 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 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(Y31~X31,dat_31)
summary(anova)
coefficients(anova)
```

Resultado:

```
            Df Sum Sq Mean Sq F value Pr(>F)
X31          1   3564    3564   358.2 <2e-16 ***
Residuals   31    308      10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> coefficients(anova)
(Intercept)         X31
  3.6265034   0.9314912
```

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.

```
#modelo de regresion
modelo31 <- lm(Y31~X31,dat_31)
summary(modelo31)
Corr <- cor(X31,Y31)
```

Resultado:

```
Call:
lm(formula = Y31 ~ X31, data = dat_31)
Residuals:
    Min      1Q  Median      3Q     Max
-6.5712 -1.5989 -0.4751  1.2509  7.4973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

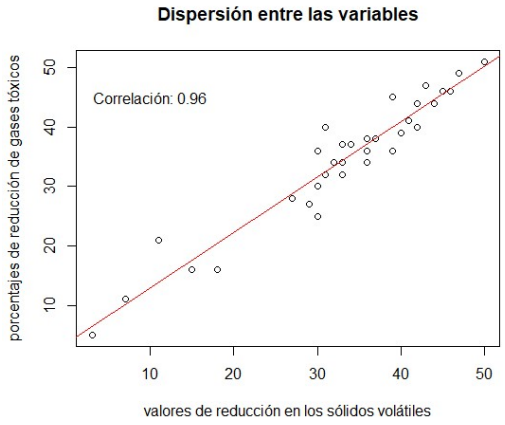
```
(Intercept)  3.62650    1.71576   2.114   0.0427 *
X31          0.93149    0.04921  18.927   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

$$Y = 3.62650+0.93149*X$$

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.

A continuación, el gráfico de dispersión de las variables incluyendo el modelo teórico de regresión:



Validación del modelo:

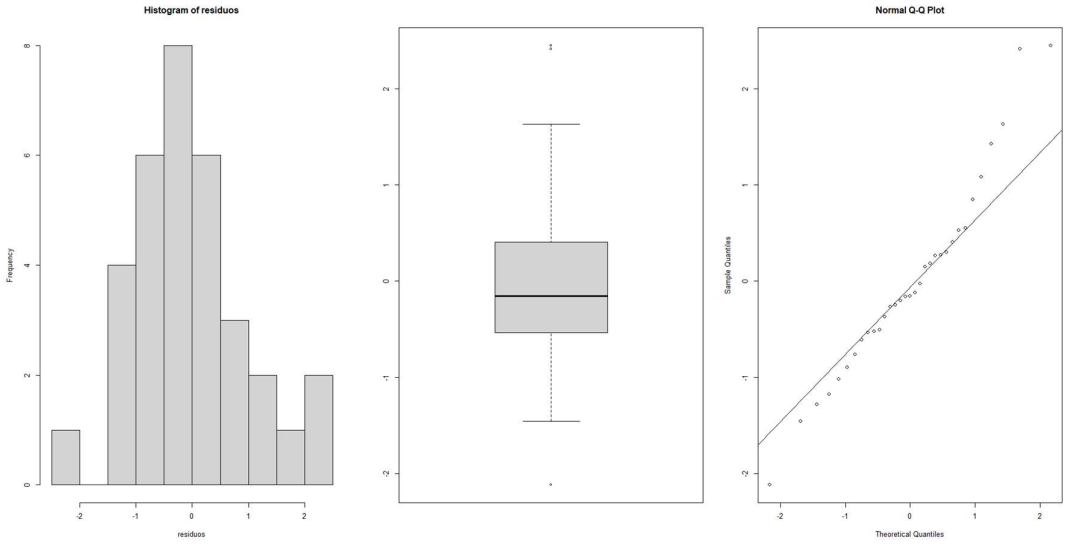
Normalidad:

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.

```
res<-rstandard(modelo31) # residuos estándar modelo ajustado
hist(res) # histograma residuos estandar
boxplot(res) # diagrama de cajas residuos estandar
qqnorm(res) # gráfico de cuantiles residuos estandar
qqline(res) # Linea de la distribución teórica Normal
```

Resultado:



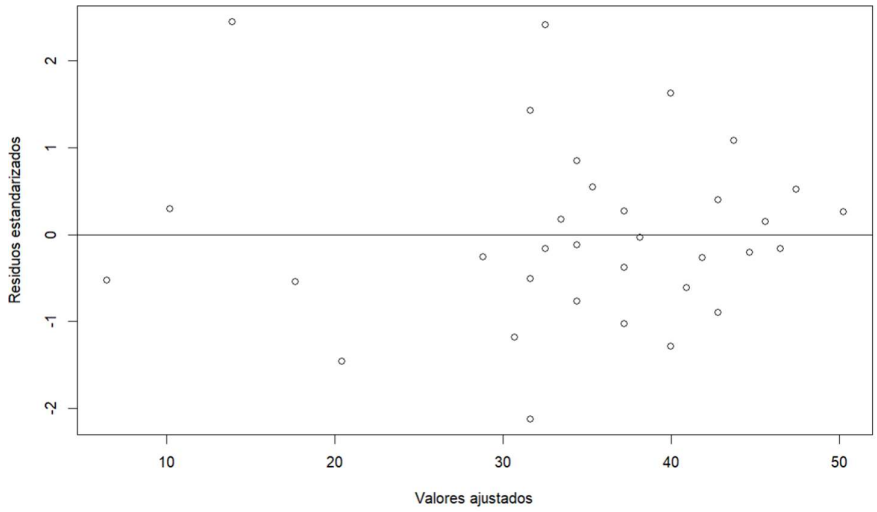
Tanto el histograma como el gráfico de cajas sugieren un comportamiento de distribución que pudiera ajustar con la distribución normal, unimodal, forma de campana. En gráfico QQ desviaciones de los puntos del gráfico respecto de la diagonal indican alteraciones/desviaciones de la normalidad. En el gráfico, 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 gráfico de dispersión con ambas variables.

```
#Valores atípicos - La independencia de los errores
plot(dat_31$X31,rstandard(modelo31),xlab="varX",ylab="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 auto correlacionados

En nuestro caso ejecutamos la prueba Durbin-Watson a continuación.

```
#prueba Durbin-Watson
library(car)
durbinWatsonTest(modelo31)
```

Resultados:

```
lag Autocorrelation D-W Statistic p-value
```

```
  1       -0.186225      2.363832   0.392
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 3.2

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.

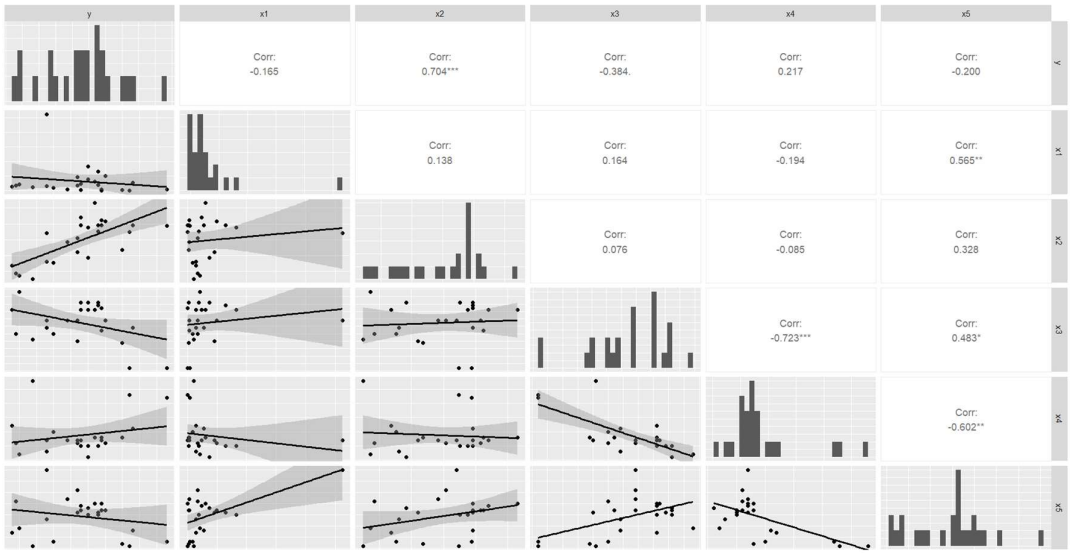
Solución

Lo primero que haremos es estudiar la relación que existe entre las variables. Esta información nos servirá para identificar cuáles pueden ser los mejores predictores en 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 histogramas de las variables en estudio.

```
library(GGally)
ggpairs(dat_32, lower = list(continuous = "smooth"),
        diag = list(continuous = "barDiag"), axisLabels = "none")
```

Resultado:



Del examen preliminar podemos observar que las variables que tienen una mayor relación lineal con el 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).

```
#modelo de regresion
modelo32 <- lm(y~x1+x2+x3+x4+x5,dat_32)
summary(modelo32)
```

Resultado:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = dat_32)

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).

```
# seleccion de predictores
step(object = modelo32, 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 = dat_32)

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:

```
#modelo de regresion v2
modelo321 <- lm(y~x2+x3+x4+x5,dat_32)
summary(modelo321)
coefficients(modelo321)
```

Resultados:

```
Call:
lm(formula = y ~ x2 + x3 + x4 + x5, data = dat_32)

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)

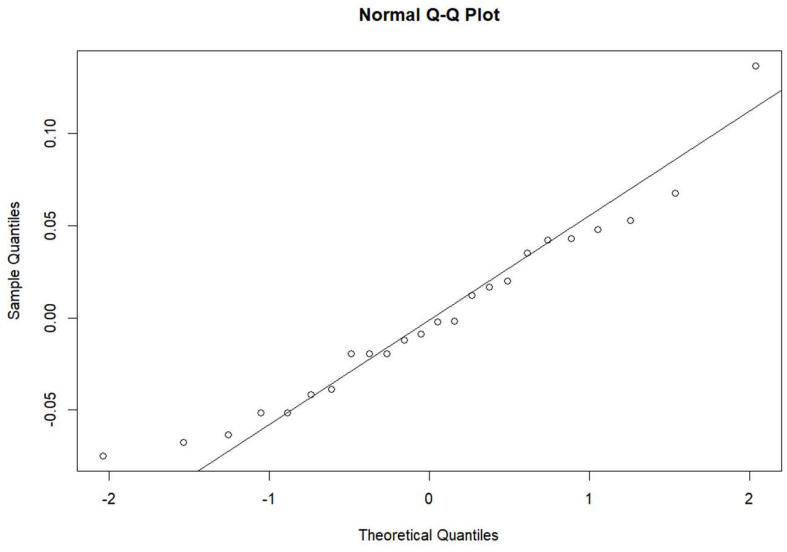
Validación del modelo

Normalidad:

Nos basamos en la forma del gráfico QQ.

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

Resultado:



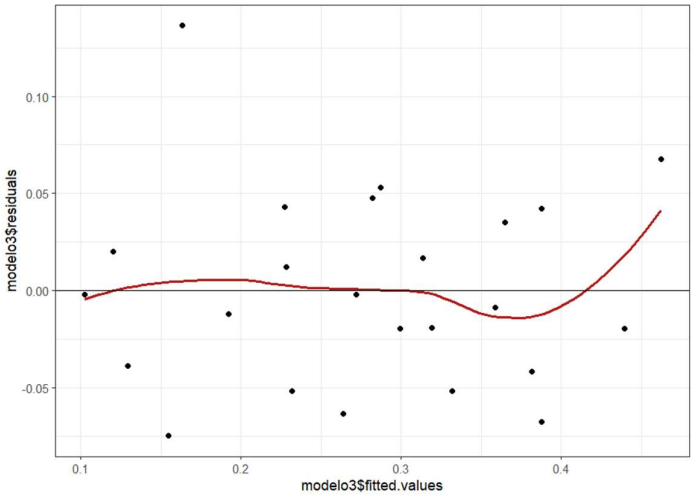
Se puede ver claramente como la mayoría de los puntos se ubican próximos a la línea de la distribución teórica. Por lo que se valida el supuesto.

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.

```
#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:



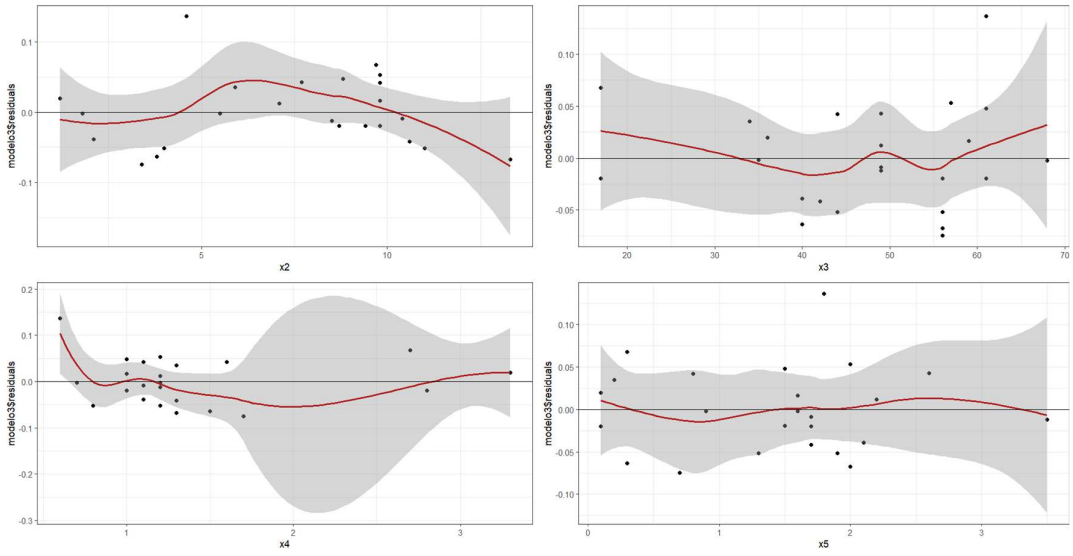
La forma como se distribuyen los valores a lo largo del eje, sin un patrón aparente, indica que el supuesto se cumple para el modelo.

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.

```
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.

Tema 4. Diseño por bloques aleatorizado

Planteamiento Actividad 4.1

Se realiza un experimento para analizar la influencia de la dosis de paracetamol ingerida sobre los niveles de toxicidad hepática bajo diferentes niveles de actividad renal. En la tabla siguiente se reflejan los niveles de toxicidad observados para diferentes niveles de insuficiencia renal, cuando se administran diferentes dosis de paracetamol.

Dosis	paracetamol	15	20	25	30	35
Actividad renal	$0-25\,\%$	7	7	15	11	18
	$25-50\,\%$	12	17	12	18	19
	$50-75\,\%$	14	18	18	19	23
	$75\,\%-1$	19	25	22	19	11

Tarea

Contrastar la significación de los bloques (niveles de actividad renal) y los tratamientos (dosis de paracetamol) en el estudio de los niveles de toxicidad hepática (α= 0;05).

Solución

En primer lugar creamos el conjunto de datos de forma de poder analizarlos adecuadamente, para la solución de esta actividad usaremos el software R. Principalmente hay que considerar: La variable respuesta, los tratamientos aplicados y los bloques en los que fueron organizados los tratamientos.

Empezamos estructurando un data frame y gráficos con la variable respuesta los tratamientos y bloques, cabe destacar que los tratamientos son los distintos niveles de paracetamol, los bloques la actividad renal en sus diversos niveles (los cuales hay que transformar a factor) y la variable respuesta de interés que es el nivel de toxicidad hepática producto de la influencia de los niveles de paracetamol.

La sintaxis aplicada fue la siguiente:

```
#Creando el Data Frame
#Variable respuesta: Niveles de Toxicidad hepática
```

```
Respuesta<-c(7,7,15,11,18,12,17,12,18,19,14,18,18,19,23,19,25,22,19,11)

Tratamiento<-c("0-25%","0-25%","0-25%","0-25%","0-25%","25-50%","25-50%","25-50%","25-
50%","25-50%","50-75%","50-75%","50-75%","50-75%","50-75%","75-1%","75-1%","75-1%","75-
1%","75-1%")

Bloques<-
c("15","20","25","30","35","15","20","25","30","35","15","20","25","30","35","15","20","
25","30","35")


## ***Se estructuro el data frame***.


datos_parac<-data.frame(Respuesta,Tratamiento,Bloques)

tratam_parac<-factor(Tratamiento)

bloque_act_ren<-factor(Bloques)


#Determinacion de los factores y bloques

tratamientof<-factor(tratam_parac)

bloquef<-factor(bloque_act_ren)
```

Modelo de Diseño por Bloques al Azar

```
modelo<-lm(Respuesta~tratam_parac+bloque_act_ren)

ANOVA<-aov(modelo)

RESUMEN_ANOVA<-summary(ANOVA)

RESUMEN_ANOVA

summary(modelo)
```

Resultado

```
# RESUMEN_ANOVA
               Df Sum Sq Mean Sq F value Pr(>F)
tratam_parac    3  176.8   58.93   2.994 0.0731 .
bloque_act_ren  4   54.2   13.55   0.688 0.6138
Residuals      12  236.2   19.68
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


# summary(modelo)
Call:
lm(formula = Respuesta ~ tratam_parac + bloque_act_ren)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-9.750 -1.163 -0.175  2.400  5.250

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.400      2.806   2.994   0.0112 *
tratam_parac25-50%    4.000      2.806   1.426   0.1795
tratam_parac50-75%    6.800      2.806   2.423   0.0321 *
tratam_parac75-1%     7.600      2.806   2.709   0.0190 *
bloque_act_ren20      3.750      3.137   1.195   0.2550
bloque_act_ren25      3.750      3.137   1.195   0.2550
bloque_act_ren30      3.750      3.137   1.195   0.2550
bloque_act_ren35      4.750      3.137   1.514   0.1559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.437 on 12 degrees of freedom
Multiple R-squared:  0.4944,  Adjusted R-squared:  0.1995
F-statistic: 1.677 on 7 and 12 DF,  p-value: 0.2059
```

Se observa que el modelo propuesto segùn las condiciones y datos recolectados es y= 8.400+4(tratam_parac 50-75%)+7.60(tratam_parac 75-1%) + el termino e (error).El modelo propuesto solo representa el 19.95% de la variabilidad total de la variable respuesta nivel de toxicidad hèpatica.

En la tabla ANOVA se va a determinar si existen diferencias significativas de los efectos tanto para los niveles de paracetamol (tratamientos) y niveles de actividad renal (bloques) destacando que en estos no se realiza la aleatorizaciòn sino en las dosis de paracetamol.

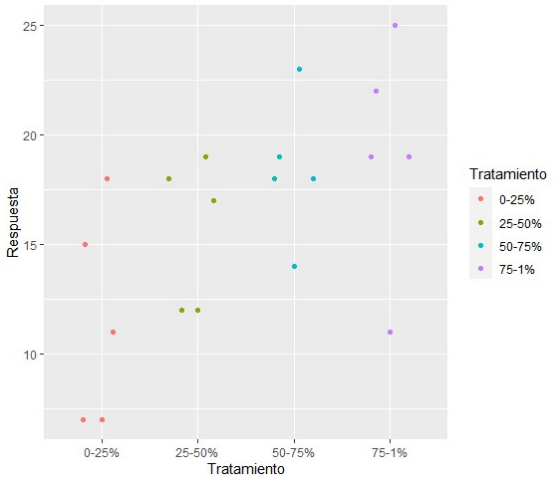
Visualización y Análisis exploratorio

Se elaboró un grafico para inspeccionar visualmente si puede inferirse una asociación entre el tratamiento analizado y la respuesta obtenida. Para ello usamos el siguiente gráfico de dispersión.

```
#  Gràfico 1
```

```
ggplot2(datos_parac, aes(x = Tratamiento, y = Respuesta, color = Tratamiento)) +
  geom_quasirandom()
```

Resultado

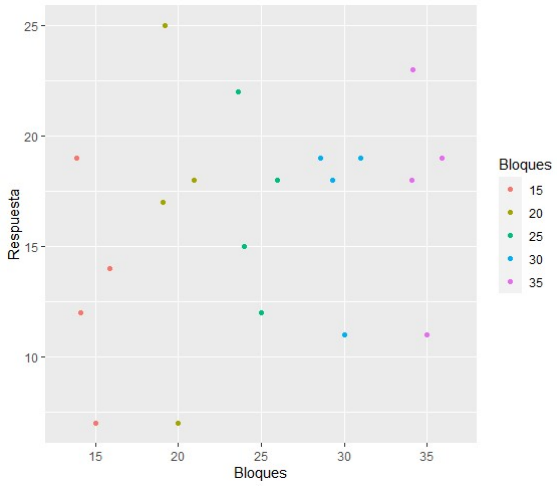


En el gráfico 1 se observa el Efecto de dosis de paracetamol en la toxicidad hepática, en líneas generales se observa que a mayor dosis de paracetamol mayor es la toxicidad hepática.

Sin embargo, por el hecho de que cada unidad experimental tiene una diversa actividad renal se debe analizar el problema como un Diseño por Bloque Completo al Azar.

```
# Gràfico 2

ggplot(datos_parac, aes(x = Bloques, y = Respuesta, color = Bloques)) +
  geom_quasirandom()
```

Resultado



El gráfico 2 muestra la relación entre el efecto en la toxicidad hepática por cada actividad renal una vez asignada la dosis de paracetamol según el nivel.

Comprobación de los supuestos (Análisis Residual)

Normalidad

Debemos comprobar que los residuos tienen un comportamiento que ajusta en una distribución normal. Para ello usamos el test de Shapiro-Wilk. En este test Se plantea como hipótesis nula que una muestra proviene de una población normalmente distribuida. Ejecutamos el test con la siguiente sintaxis:

```
normalidad=shapiro.test(resid(modelo))
print(normalidad)
```

Resultado

```
Shapiro-Wilk normality test
data:  resid(modelo)
W = 0.92362, p-value = 0.1163
```

Dado el p-valor obtenido, no tenemos evidencia suficiente para rechazar la hipótesis nula, por tanto, se acepta que los residuos se distribuyen de forma normal.

Homocedasticidad Tratamientos

Debemos comprobar la igualdad de varianza de los residuos para todos los niveles de tratamiento. En este caso aplicamos el test de Barlett, en el cual se contrasta la hipótesis nula de igualdad de varianzas entre grupos, que en este caso los grupos son los tratamientos. Aplicamos el siguiente código:

```
homocedasticidad_tratamiento=bartlett.test(resid(modelo)~Tratamiento)
print(homocedasticidad_tratamiento)
```

Resultado

```
Bartlett test of homogeneity of variances
data:  resid(modelo) by Tratamiento
Bartlett's K-squared = 5.5315, df = 3, p-value = 0.1368
```

El resultado de la prueba indica que la homogeneidad de las varianzas entre los tratamientos.

Homocedasticidad en Bloques

Ahora, comprobamos la igualdad de varianza de los residuos para los bloques. En este caso aplicamos nuevamente el test de Barlett, en el cual se contrasta la hipótesis nula de igualdad de varianzas entre grupos, que en este son los bloques. Aplicamos el siguiente código:

```
homocedasticidad_bloques=bartlett.test(resid(modelo)~Bloques)
print(homocedasticidad_bloques)
```

Resultado

```
Bartlett test of homogeneity of variances
data:  resid(modelo) by Bloques
Bartlett's K-squared = 7.2382, df = 4, p-value = 0.1238
```

El resultado indica nuevamente la homogeneidad de varianzas entre los bloques.

Conclusión

Se puede observar que los tratamientos y bloques no son significativos es decir el efecto promedio de toxicidad hepática tanto para los niveles de paracetamol y actividad renal no difieren a un nivel de significación del 5%. El modelo propuesto representa una variabilidad total del 19.95 % se cumplen los supuestos de normalidad de los residuos y varianza para los tratamientos y bloques ya que p-valor >0.05. Por consiguiente, el modelo es adecuado.

Sin embargo, con un nivel de significación del 10% se presenta una diferencia significativa entre los niveles de paracetamol y el efecto de toxicidad hepática. En este caso si se puede continuar con las pruebas de rango múltiple.