

ESTADÍSTICA PARA INGENIERÍA Y CIENCIAS

PRÁCTICA 4: Análisis de varianza

Ivan Svetlich

```
#Librerias
library(IRdisplay)
library(formattable)
library(ggplot2)
library(cowplot)
library(dplyr)
library(stringr)
```

Ejercicio 1

Un fabricante está interesado en la resistencia a la tensión de una fibra sintética. Se sospecha que la resistencia está relacionada con el porcentaje de algodón en la fibra. Para ello se emplean cinco niveles de porcentaje de algodón, y se corren cinco réplicas en orden aleatorio obteniéndose los siguientes datos:

Porcentaje de algodón	Resistencia				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

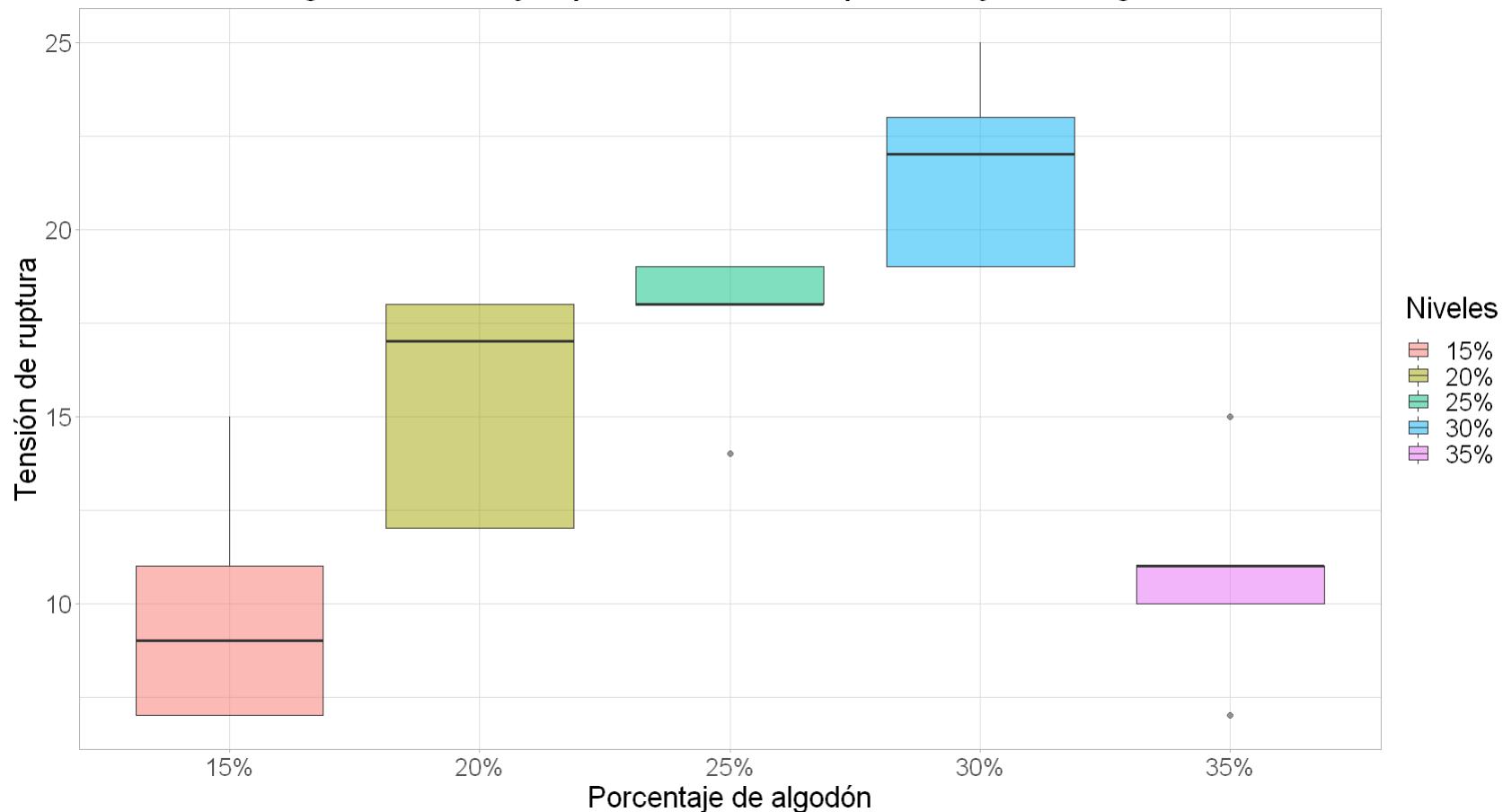
```
data_1 <- read.csv("./TP4_tables/data1.csv") # Leo los datos desde archivo .csv
```

a) ¿El porcentaje de algodón tiene algún efecto sobre la tensión de ruptura? Dibuje diagramas de caja comparativos y realice un análisis de varianza. Utilice $\alpha = 0.05$.

Averiguar si el porcentaje de algodón en la fibra tiene un efecto sobre la tensión de ruptura implica determinar si las medias poblacionales de cada tratamiento son iguales o no. Como primer indicio, los diagramas de cajas permiten hacer una comparación visual de las distintas poblaciones:

```
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_1, aes(x=cotton_percentage, y=tensile_strength, fill=cotton_percentage)) +
  labs(
    title="Diagramas de cajas para los distintos porcentajes de algodón",
    x="Porcentaje de algodón",
    y="Tensión de ruptura",
    fill="Niveles") +
  geom_boxplot(alpha=0.5) +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```

Diagramas de cajas para los distintos porcentajes de algodón



Un análisis visual de los diagramas de cajas sugiere que existen diferencias entre las medias de las poblaciones correspondientes a los distintos tratamientos.

La determinación formal respecto a si las medias del tratamiento son diferentes implica probar la hipótesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5, \quad \text{contra } H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

seiendo μ_i la media del tratamiento i . Para ello se utiliza el método de **análisis de varianza en un sentido**. El estadístico es:

$$F = \frac{MST_r}{MSE}$$

donde MST_r es la media cuadrática de tratamiento y MSE la media cuadrática del error.

```
data_1.aov <- aov(tensile_strength ~ cotton_percentage, data_1)
aov_test <- summary(data_1.aov)[[1]]
```

```
display_markdown('#### **ANOVA de un sentido:**')
aov_test <- cbind(c('Porcentaje de algodón', 'Residuos'), aov_test)
colnames(aov_test)[1] <- 'Source'
rownames(aov_test) <- c()
table <- formattable(aov_test, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style(
  'text-align'='left'))))
as.htmlwidget(table, width="70%", height=NULL)
```

ANOVA de un sentido:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Porcentaje de algodón	4	475.76	118.94	14.75682	9.127937e-06
Residuos	20	161.20	8.06	NA	NA

El resultado de la prueba es un p-valor = $9.127937e - 06 < 0.05$. Por lo tanto, hay evidencia significativa en contra de la hipótesis nula y se concluye que el porcentaje de algodón en la fibra tiene un efecto sobre la tensión de ruptura de la misma.

b) Use el método de Tukey para identificar diferencias específicas entre los porcentajes.

El método de Tukey-Kramer está basado en la distribución de rango studentizado y se utiliza para construir intervalos de confianza y realizar pruebas de hipótesis de forma simultánea para todas las diferencias entre las medias de los distintos tratamientos. Esta herramienta permite determinar cuáles pares de tratamientos difieren en su efecto sobre la variable respuesta.

Los intervalos de confianza simultáneos de Tukey-Kramer de nivel $100(1 - \alpha)$ para todas las diferencias $\mu_i - \mu_j$ son:

$$\bar{X}_i - \bar{X}_j \pm q_{I,N-I,\alpha} \sqrt{\frac{MSE}{2} \left(\frac{1}{J_i} + \frac{1}{J_j} \right)}$$

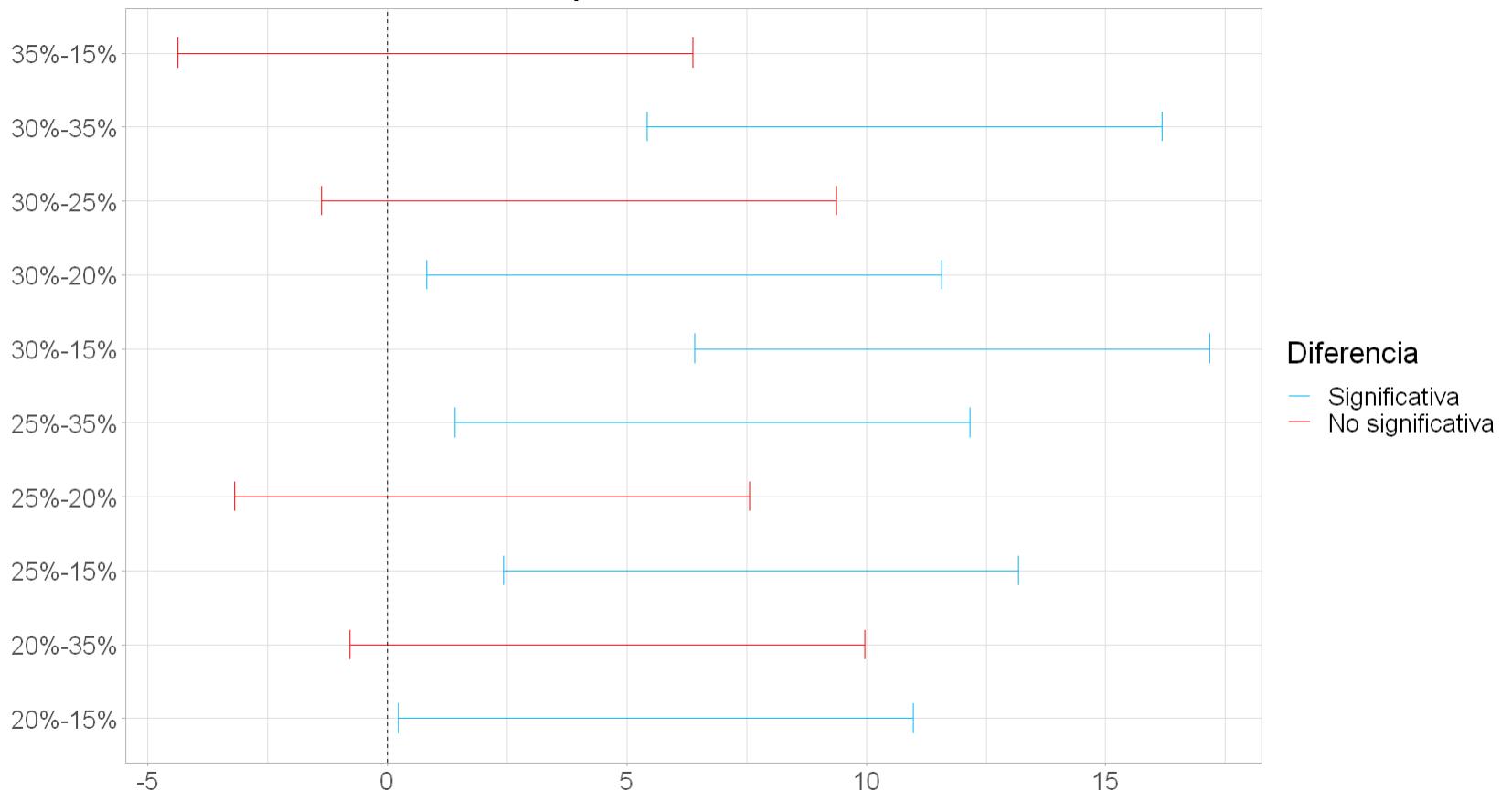
donde:

- \bar{X}_i y \bar{X}_j : medias muestrales en los niveles i y j ,
- I : número de tratamientos,
- J_i y J_j : tamaños de las muestras en los niveles i y j ,
- N : número total de observaciones,
- MSE : media cuadrática del error.

```
data_1.tukey <- as.data.frame(TukeyHSD(data_1.aov, ordered = TRUE, conf.level = 0.95)[1]$cotton_percentage)
```

```
data_1.tukey$names <- c(rownames(data_1.tukey))
# Gráfico de los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_1.tukey, aes(names, diff)) +
  labs(
    title="Intervalos de confianza de 95% para la diferencia de medias entre tratamientos",
    x="",
    y="",
    col="Diferencia") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(lwr*upr > 0,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5', '#f50505'), labels=c('Significativa', 'No significativa'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```

Intervalos de confianza de 95% para la diferencia de medias entre tratamientos



Los intervalos que no contienen al cero indican, con un nivel de confianza de 95%, que existe una diferencia en los efectos de los tratamientos considerados.

c) Encuentre los residuos y examínelos en lo que respecta a la insuficiencia del modelo.

Las pruebas de hipótesis usuales del ANOVA de un sentido son válidas si se cumplen las siguientes condiciones:

1. Las poblaciones en tratamiento deben ser normales.
2. Las poblaciones en tratamiento deben tener todas la misma varianza.

```
smoothed <- data.frame(with(data_1.aov, lowess(x = data_1.aov$fitted, y = data_1.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_1.aov) +
  geom_point(aes(x=data_1.aov$fitted, y=data_1.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
```

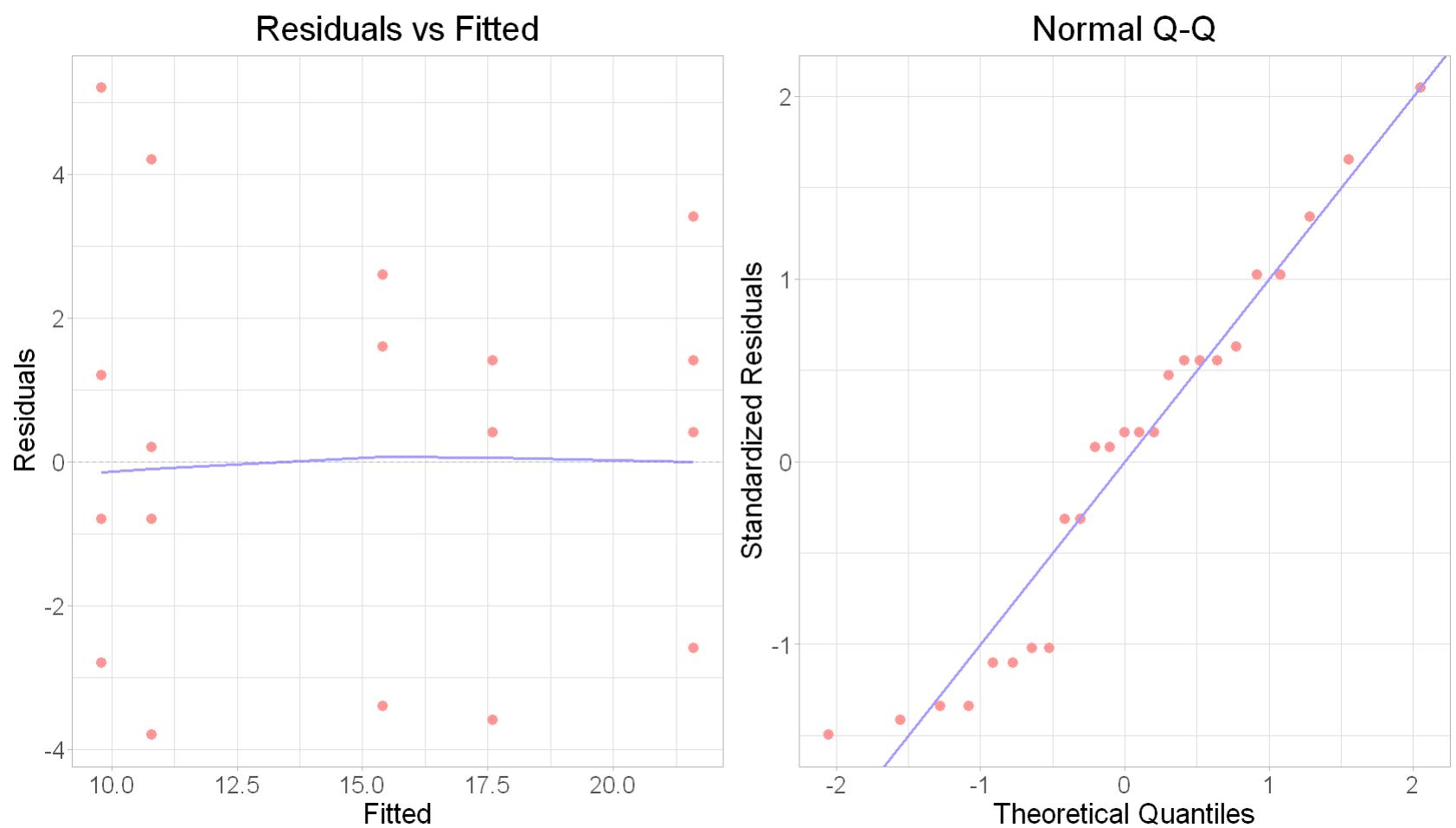
```

ylab("Residuals") +
theme_light() +
theme(text=element_text(size=20),
plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_1.aov) +
stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
geom_abline(col="#a399ff", size=1) +
xlab("Theoretical Quantiles") +
ylab("Standardized Residuals") +
ggttitle("Normal Q-Q") +
theme_light() +
theme(text=element_text(size=20),
plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



La gráfica de residuos vs. valores ajustados no presenta patrones apreciables, y por lo tanto es razonable asumir que la varianza es constante. En cuanto al gráfico Q-Q, si bien los valores no se ajustan estrictamente a una recta, no hay evidencias de una violación grave del principio de normalidad.

Ejercicio 2

Se estudia la resistencia a la compresión del concreto, así como cuatro técnicas de mezclado diferentes. Se obtienen los siguientes datos:

Técnica de mezclado	Resistencia a la compresión			
	1	2	3	4
1	3129	3200	2800	2600
2	3000	3300	2900	2700
3	2865	2975	2985	2600
4	2890	3150	3050	2765

```

data_2 <- read.csv("./TP4_tables/data2.csv") # Leo Los datos desde archivo .csv
data_2$mixing_method <- as.factor(data_2$mixing_method)

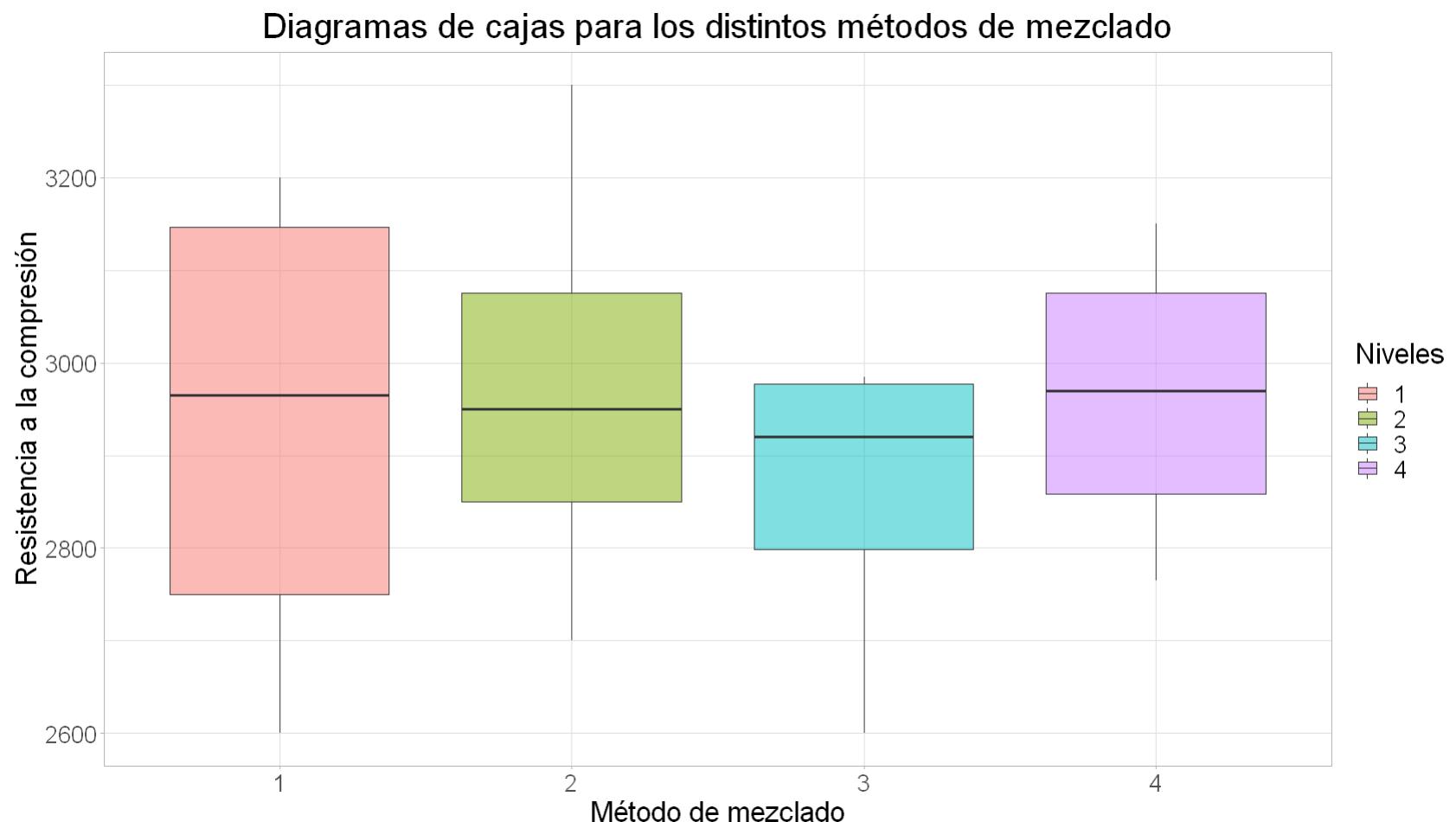
```

a) Pruebe la hipótesis de que las técnicas de mezclado afectan la resistencia del concreto. Utilice $\alpha = 0.05$.

Las hipótesis a probar son:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4, \quad \text{contra } H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

```
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_2, aes(x=mixing_method, y=compression_strength, fill=mixing_method)) +
  labs(
    title="Diagramas de cajas para los distintos métodos de mezclado",
    x="Método de mezclado",
    y="Resistencia a la compresión",
    fill="Niveles") +
  geom_boxplot(alpha=0.5) +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```



En primera instancia, los diagramas de cajas sugieren que prodía no existir una diferencia entre los efectos de los métodos de mezclado estudiados.

```
data_2.levels <- split(data_2, f=data_2$mixing_method)
```

```
# Cálculo de MStr
grand_mean <- mean(data_2$compression_strength) # media general
treatment_means <- sapply(data_2.levels, function(x) {
  mean(x$compression_strength)
}) # media de cada tratamiento
J <- sapply(data_2.levels, nrow) # cantidad de observaciones para cada tratamiento
SSTr <- sum(J * (treatment_means - grand_mean)^2)
I <- length(data_2.levels) # cantidad de niveles
MStr <- SSTr / (I - 1)
display_markdown(sprintf('$MStr = %.f$', MStr))
```

$$MStr = 11460$$

```
# Cálculo de MSE
samples <- t(sapply(data_2.levels, function(x) {x$compression_strength})) # matriz con todas las observaciones
residuals <- samples - treatment_means # residuos
SSE <- sum(residuals^2)
N <- sum(J) # cantidad de observaciones
MSE <- SSE / (N - I)
display_markdown(sprintf('$MSE = %.f$', MSE))
```

$$MSE = 50772$$

```
F <- MStr / MSE
display_markdown(sprintf('$F = %.4f$', F))
```

$$F = 0.2257$$

```
alpha <- 0.05
f_alpha <- qf(alpha, df1=I-1, df2=N-I, lower=FALSE)
display_markdown(sprintf('$f_{\alpha} = %.3f, f_{1-\alpha} = %.4f$', alpha, I-1, N-I, f_alpha))
```

$$f_{\alpha=0.05, 3, 12} = 3.4903$$

Como $F = 0.2257 < 3.4903$, no es posible rechazar la hipótesis nula y se concluye que la técnica de mezclado utilizada no afecta la resistencia del concreto.

b) Encuentre el p-valor para el estadístico F del inciso a).

```
p_value <- pf(F, df1=I-1, df2=N-I, lower=FALSE)
display_markdown(sprintf('$\\text{p-valor} = %.4f$', p_value))
```

p-valor = 0.8767

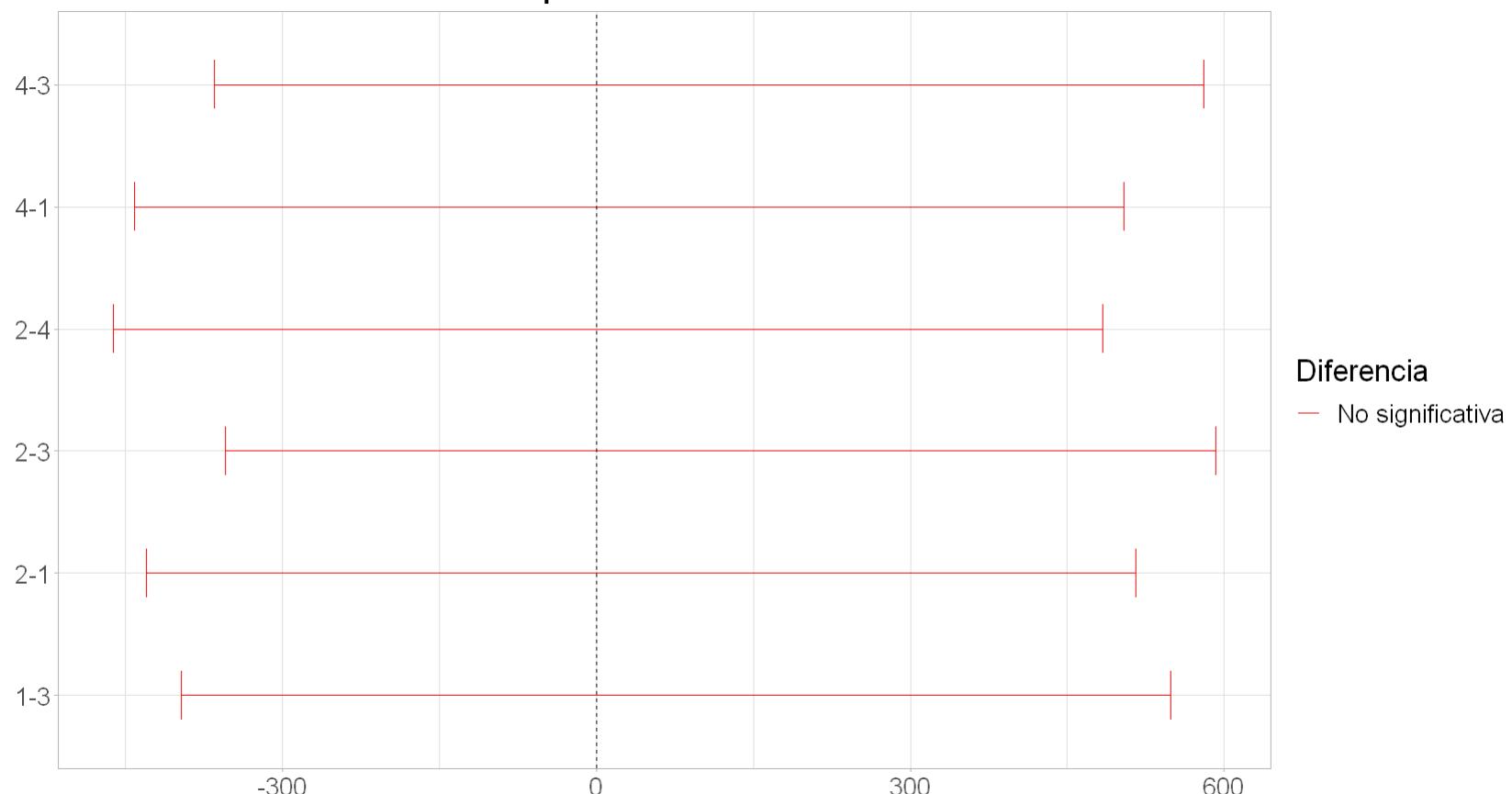
Como el p-valor = 0.8767 > 0.05, se llega a la misma conclusión que en el inciso anterior.

c) Use el método de Tukey para identificar diferencias específicas entre los porcentajes.

```
data_2.aov <- aov(compression_strength ~ mixing_method, data_2)
data_2.tukey <- as.data.frame(TukeyHSD(data_2.aov, ordered = TRUE, conf.level = 0.95)[1]$mixing_method)
```

```
data_2.tukey$names <- c(rownames(data_2.tukey))
# Gráfico de los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_2.tukey, aes(names, diff)) +
  labs(
    title="Intervalos de confianza de 95% para la diferencia de medias entre tratamientos",
    x="",
    y="",
    col="Diferencia") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(lwr*upr > 0,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5','#f50505'), labels=c('Significativa','No significativa'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```

Intervalos de confianza de 95% para la diferencia de medias entre tratamientos



Todos los intervalos de confianza contienen al cero, lo cual indica que ningún par de tratamientos presenta una diferencia significativa entre sus medias.

d) Analice los residuos de este experimento.

```
smoothed <- data.frame(with(data_2.aov, lowess(x = data_2.aov$fitted, y = data_2.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
```

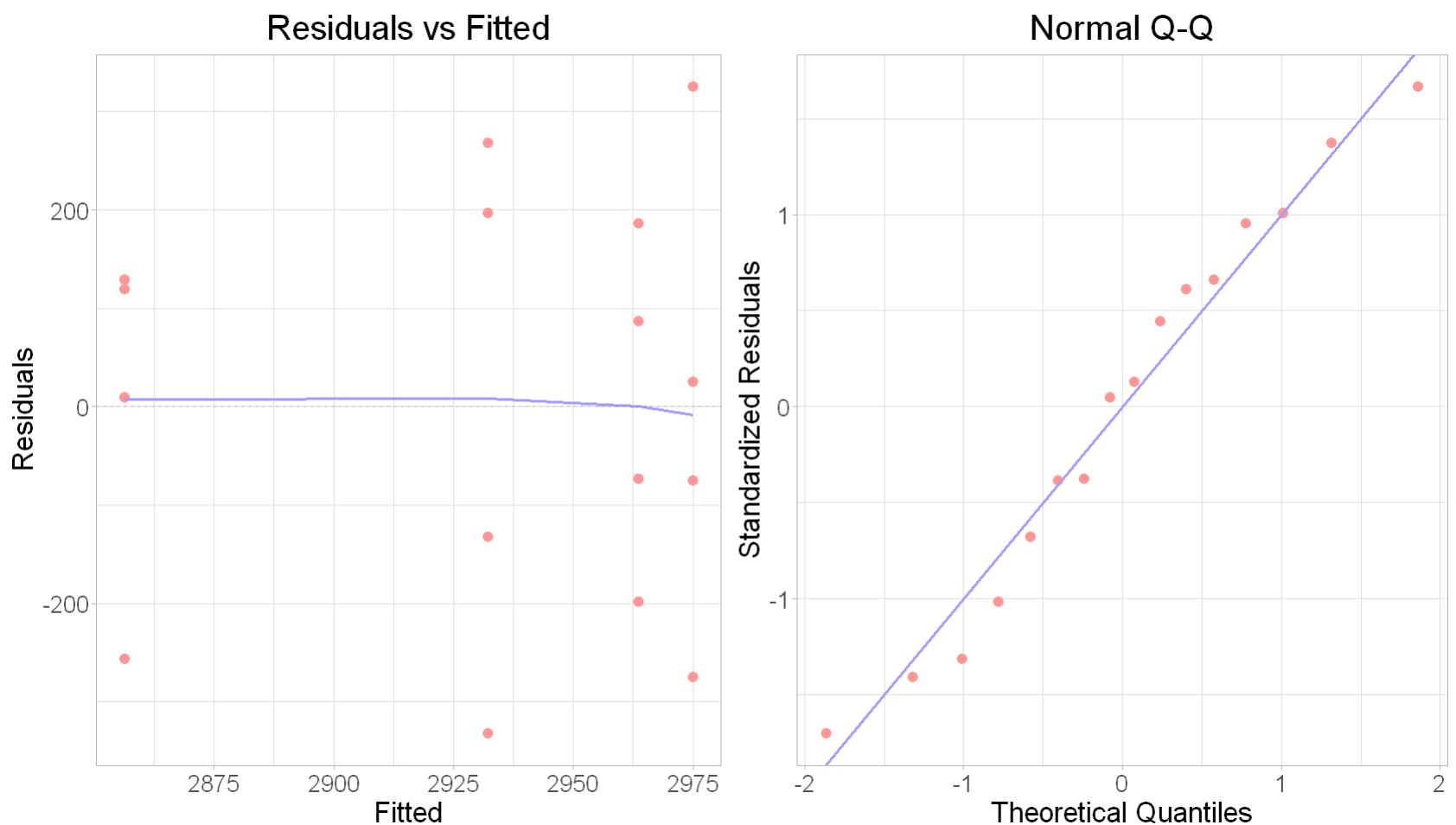
```

res_vs_fit <- ggplot(data_2.aov) +
  geom_point(aes(x=data_2.aov$fitted, y=data_2.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_2.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggtitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



Las gráficas de *Residuos vs. Valores ajustados* y *Residuos estandarizados vs. Cuantiles Teóricos* no sugieren violaciones notables de los supuestos de varianza constante y normalidad.

Ejercicio 3

Un ingeniero en electrónica está interesado en el efecto sobre la conductividad de una válvula electrónica que tiene cinco tipos diferentes de recubrimiento para los tubos de rayos catódicos utilizados en un dispositivo de visualización de un sistema de telecomunicaciones. Se obtienen los datos siguientes sobre la conductividad:

Conductividad					
Recubrimiento	1	2	3	4	
1	143	141	150	146	
2	152	149	137	143	
3	134	133	132	127	
4	147	148	144	142	

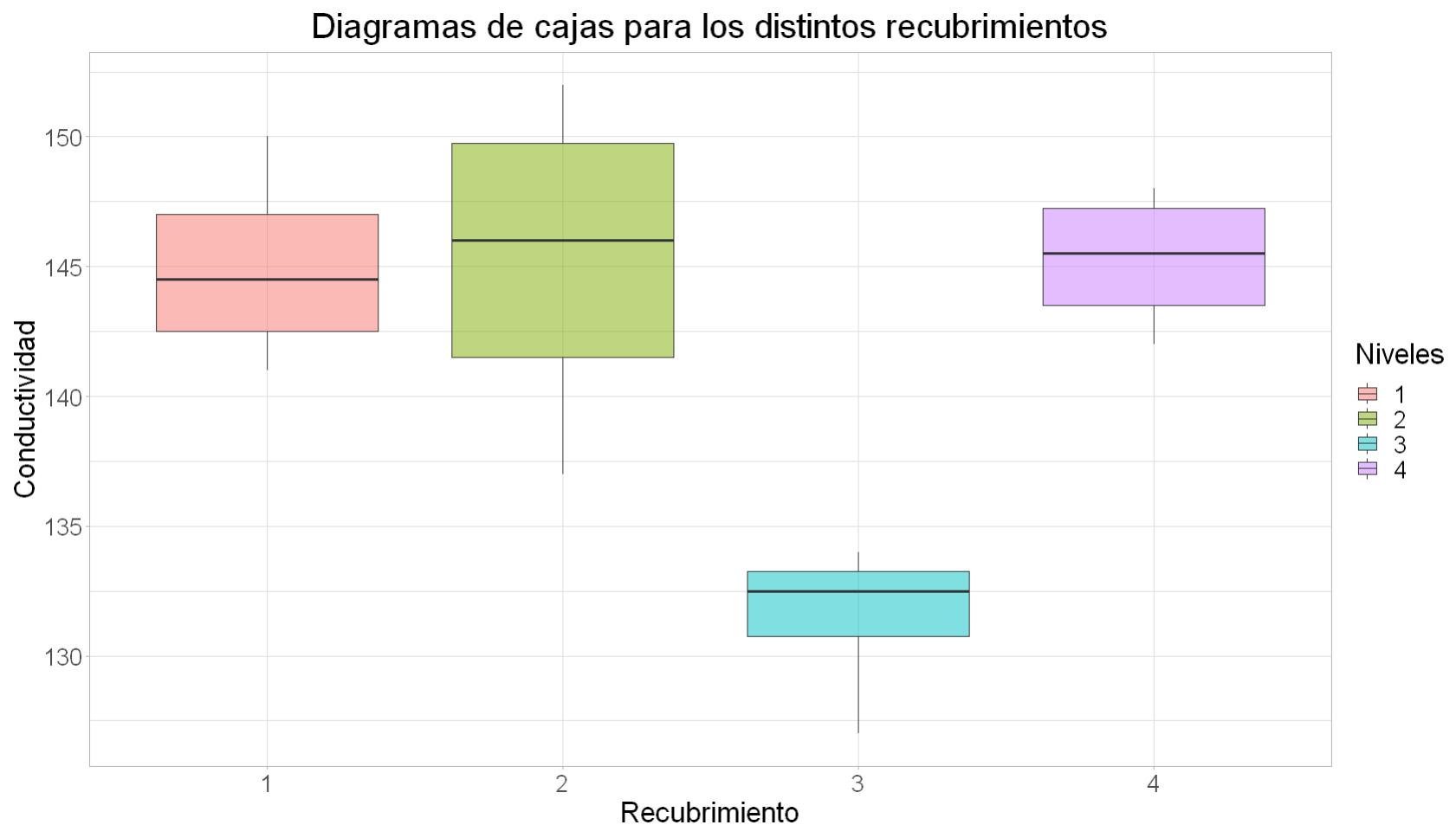
```

data_3 <- read.csv("./TP4_tables/data3.csv") # Leo Los datos desde archivo .csv
data_3$protection <- factor(data_3$protection)

```

a) ¿Existe alguna diferencia en la conductividad debida al tipo de recubrimiento? Utilice $\alpha = 0.05$.

```
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_3, aes(x=protection, y=conductivity, fill=protection)) +
  labs(
    title="Diagramas de cajas para los distintos recubrimientos",
    x="Recubrimiento",
    y="Conductividad",
    fill="Niveles") +
  geom_boxplot(alpha=0.5) +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```



Los gráficos de cajas sugieren que el recubrimiento tipo 3 produce un efecto diferente al del resto de los recubrimientos estudiados. Para determinar formalmente si existe alguna diferencia en la conductividad debida al tipo de recubrimiento se prueban las hipótesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4, \quad \text{contra} \quad H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

donde μ_i es la media poblacional del tratamiento i ($i = [1, 2, 3, 4]$).

```
data_3.aov <- aov(conductivity ~ protection, data_3)
aov_test3 <- summary(data_3.aov)[[1]]
```

```
display_markdown('#### **ANOVA de un sentido:**')
aov_test3 <- cbind(c('Porcentaje de algodón', 'Residuos'), aov_test3)
colnames(aov_test3)[1] <- 'Source'
rownames(aov_test3) <- c()
table <- formattable(aov_test3, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style(
  'text-align='left'))))
as.htmlwidget(table, width="70%", height=NULL)
```

ANOVA de un sentido:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Porcentaje de algodón	3	560.5	186.83333	9.726681	0.001554877
Residuos	12	230.5	19.20833	NA	NA

El p-valor de la prueba es $0.0016 < 0.05$. Por lo tanto, hay evidencia significativa en contra de la hipótesis nula y se puede concluir que al menos dos medias difieren entre sí. Esto significa que el recubrimiento utilizado afecta la conductividad.

b) Analice los residuos de este experimento.

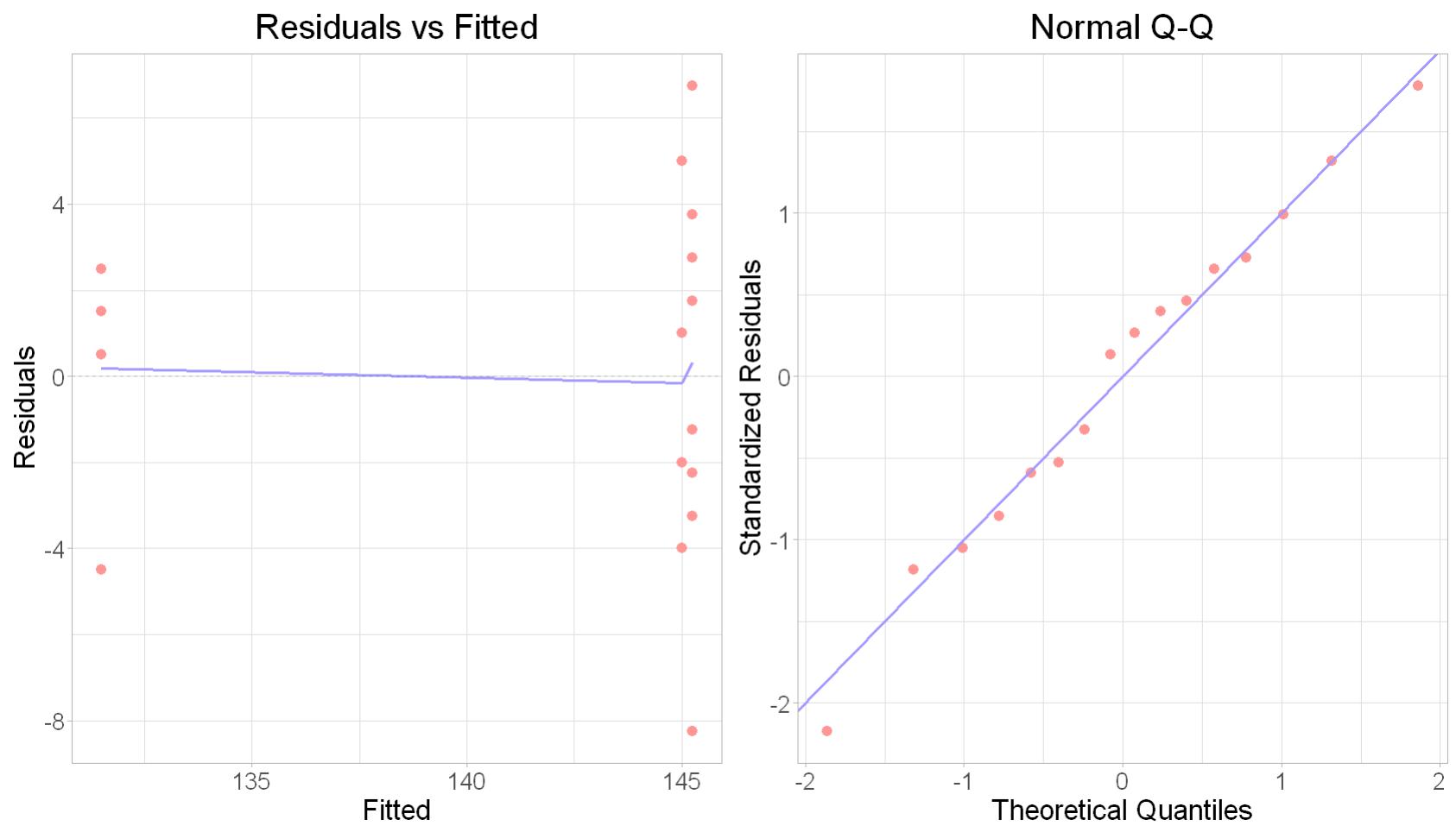
```

smoothed <- data.frame(with(data_3.aov, lowess(x = data_3.aov$fitted, y = data_3.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_3.aov) +
  geom_point(aes(x=data_3.aov$fitted, y=data_3.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggttitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
  plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_3.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggttitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
  plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



La distribución de los valores en el gráfico de *Residuos vs. Valores Ajustados* muestra una mayor dispersión para cuanto mayor es el valor ajustado, lo cual puede indicar que la varianza de los residuos no es constante. El gráfico de *Residuos estandarizados vs. Cuantiles teóricos* no exhibe desviaciones considerables de la recta, y por lo tanto la asunción de normalidad es razonable.

c) Construya un intervalo del 95% para la estimación de la media del recubrimiento de tipo 1. Construya un intervalo del 99% para la estimación de la diferencia de las medias entre los recubrimientos 1 y 4.

Un intervalo de confianza de nivel $(1 - \alpha)\%$ para la media del tratamiento i está dado por:

$$\bar{X}_i \pm t_{N-I, \alpha/2} \sqrt{\frac{MSE}{J_i}}$$

```

data_3.levels <- split(data_3, f=data_3$protection)
sample_mean <- mean(data_3.levels[[1]]$conductivity) # media muestral para el recubrimiento tipo 1
display_markdown(sprintf('$\overline{X}_1 = %.f$', sample_mean))

```

$$\overline{X}_1 = 145$$

```

J <- sapply(data_3.levels, nrow) # cantidad de observaciones para cada tratamiento
I <- length(data_3.levels) # cantidad de niveles
N <- sum(J) # número total de observaciones
samples <- t(sapply(data_3.levels, function(x) {x$conductivity})) # matriz con todas las observaciones
treatment_means <- sapply(data_3.levels, function(x) {

```

```

    mean(x$conductivity)
}) # media de cada tratamiento
residuals <- samples - treatment_means #residuos
SSE <- sum(residuals^2)
MSE <- SSE / (N - I)
display_markdown(sprintf('$MSE = %.4f$', MSE))

```

$$MSE = 19.2083$$

```

alpha <- 0.05
t <- qt(alpha/2, N-I, lower=FALSE) # distribución t de Student con alpha=0.05/2 y N-I grados de libertad
aux <- t * sqrt(MSE / J[1])
conf_int <- c(sample_mean - aux, sample_mean + aux) # intervalo de confianza
conf_int.df <- as.data.frame(cbind('Tratamiento 1', round(conf_int[1], 4), sample_mean, round(conf_int[2], 4)))
colnames(conf_int.df) <- c(' ', '2.5%', 'Media estimada', '97.5%')

```

```

display_markdown('#### **Intervalo de confianza del \textbf{95\%} para la estimación de la media del recubrimiento tipo 1:**')
rownames(conf_int.df) <- c()
as.htmlwidget(formattable(conf_int.df, align='c', list(' ' = formatter("span", style = ~ style(
  'font-weight'='bold', 'text-align'='left')))), width="50%")

```

Intervalo de confianza del 95% para la estimación de la media del recubrimiento tipo 1:

	2.5%	Media estimada	97.5%
Tratamiento 1	140.2254	145	149.7746

Para construir un intervalo de confianza para la diferencia entre las medias de dos tratamientos específicos, se utiliza el método de la diferencia significativa mínima de Fisher. El intervalo de nivel $(1 - \alpha)$ para la diferencia $\mu_i - \mu_j$ es:

$$\bar{X}_i - \bar{X}_j \pm t_{N-I, \alpha/2} \sqrt{\frac{MSE}{J_i} + \frac{MSE}{J_j}}$$

```

x_bar_1 <- mean(data_3.levels[[1]]$conductivity) # media muestral para el recubrimiento tipo 1
x_bar_4 <- mean(data_3.levels[[4]]$conductivity) # media muestral para el recubrimiento tipo 4
x_bar_diff <- x_bar_1 - x_bar_4
display_markdown(sprintf('$\overline{X}_1 - \overline{X}_4 = %.2f$', x_bar_diff))

```

$$\bar{X}_1 - \bar{X}_4 = -0.25$$

```

alpha <- 0.01
t <- qt(alpha/2, N-I, lower=FALSE) # distribución t de Student con alpha=0.05/2 y N-I grados de libertad
aux <- t * sqrt(MSE / J[1] + MSE / J[4])
conf_int <- c(x_bar_diff - aux, x_bar_diff + aux) # intervalo de confianza
conf_int.df <- as.data.frame(cbind('μ1 - μ4', round(conf_int[1], 4), x_bar_diff, round(conf_int[2], 4)))
colnames(conf_int.df) <- c(' ', '0.5%', 'Valor estimado', '99.5%')

```

```

display_markdown('#### **Intervalo de confianza del \textbf{99\%} para la estimación de $\mu_1 - \mu_4:$**')
rownames(conf_int.df) <- c()
as.htmlwidget(formattable(conf_int.df, align='c', list(' ' = formatter("span", style = ~ style(
  'font-weight'='bold', 'text-align'='left')))), width="50%")

```

Intervalo de confianza del 99% para la estimación de $\mu_1 - \mu_4$:

	0.5%	Valor estimado	99.5%
μ1 - μ4	-9.7162	-0.25	9.2162

Ejercicio 4

Se determina el tiempo de respuesta, en milisegundos, para tres tipos diferentes de circuitos en una calculadora electrónica. Los resultados son los siguientes:

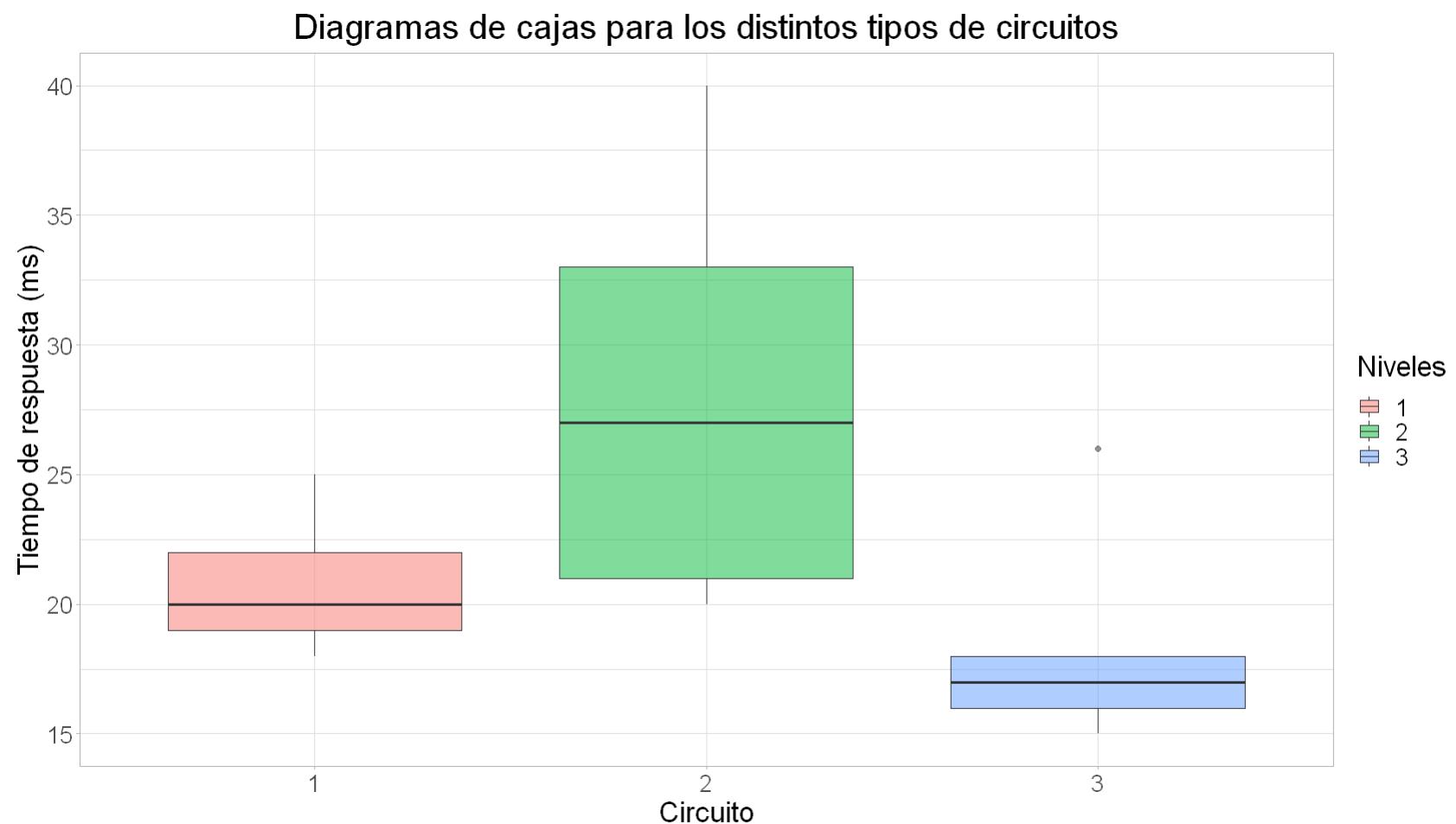
Tipo de circuito	Tiempo de respuesta (ms)				
	1	2	3	4	5
1	19	22	20	18	25

Tipo de circuito	Tiempo de respuesta (ms)				
	1	2	3	4	5
2	20	21	33	27	40
3	16	15	18	26	17

```
data_4 <- read.csv("./TP4_tables/data4.csv") # Leo los datos desde archivo .csv
data_4$circuit <- factor(data_4$circuit)
```

a) Utilice $\alpha = 0.01$ para probar la hipótesis de que el tiempo de respuesta de los tres circuitos es el mismo.

```
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_4, aes(x=circuit, y=response_time, fill=circuit)) +
  labs(
    title="Diagramas de cajas para los distintos tipos de circuitos",
    x="Circuito",
    y="Tiempo de respuesta (ms)",
    fill="Niveles") +
  geom_boxplot(alpha=0.5) +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```



Un análisis visual sugiere que el tipo de circuito utilizado afecta al tiempo de respuesta de la calculadora, y que los mejores resultados se obtienen con el circuito tipo 3. La hipótesis a probar es:

$$H_0: \mu_1 = \mu_2 = \mu_3, \quad \text{contra } H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

```
data_4.aov <- aov(response_time ~ circuit, data_4)
aov_test4 <- summary(data_4.aov)[[1]]
```

```
display_markdown('#### **ANOVA de un sentido:**')
aov_test4 <- cbind(c('Porcentaje de algodón', 'Residuos'), aov_test4)
colnames(aov_test4)[1] <- 'Source'
rownames(aov_test4) <- c()
table <- formattable(aov_test4, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style(
  'text-align='left'))))
as.htmlwidget(table, width="70%", height=NULL)
```

ANOVA de un sentido:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Porcentaje de algodón	2	260.9333	130.46667	4.006141	0.04648445
Residuos	12	390.8000	32.56667	NA	NA

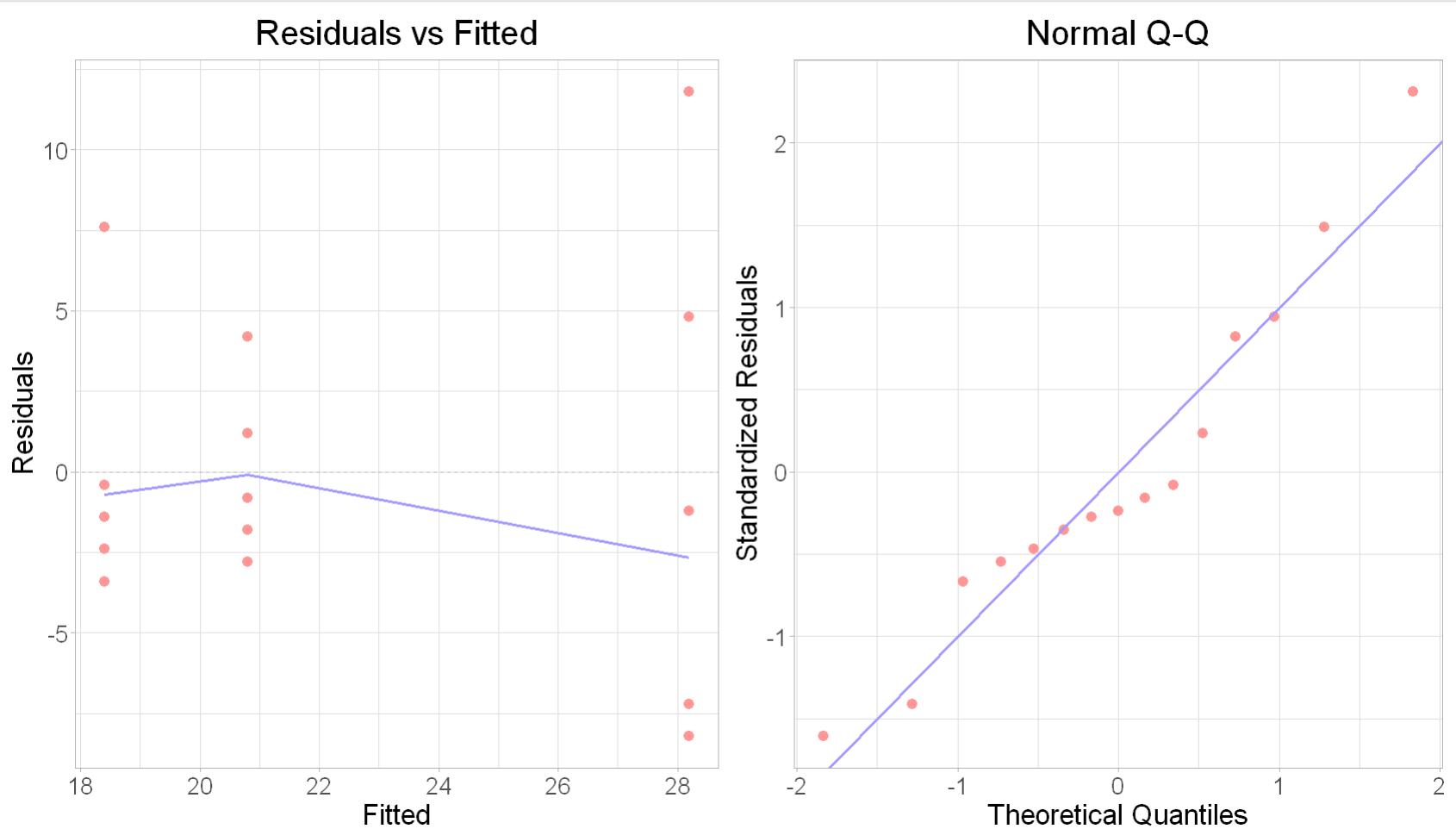
La prueba arroja un p-valor = 0.0465 > 0.01. Por lo tanto, no hay evidencia suficiente en contra de la hipótesis nula y se concluye que el tipo de circuito utilizado no afecta el tiempo de respuesta de la calculadora.

b) Analice los residuos de este experimento.

```
smoothed <- data.frame(with(data_4.aov, lowess(x = data_4.aov$fitted, y = data_4.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_4.aov) +
  geom_point(aes(x=data_4.aov$fitted, y=data_4.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_4.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggtitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)
```



Análisis de residuos

- Se observa una mayor dispersión de los residuos en el extremo derecho del gráfico, lo cual puede indicar que la varianza de los residuos no es constante.
- Las desviaciones respecto a la recta normal en el gráfico Q-Q no permiten asumir normalidad.

Un intervalo de confianza de nivel $(1 - \alpha)\%$ para la media del tratamiento i está dado por:

$$\bar{X}_i \pm t_{N-I, \alpha/2} \sqrt{\frac{MSE}{J_i}}$$

c) Encuentre un intervalo de confianza del 95% para el tiempo de respuesta del tercer circuito.

```
data_4.levels <- split(data_4, f=data_4$circuit)
x_bar_3 <- mean(data_4.levels[[3]]$response_time) # media muestral para el recubrimiento tipo 4
display_markdown(sprintf('$\overline{X}_3 = %.2f$', x_bar_3))
```

$$\overline{X}_3 = 18.40$$

```
J <- sapply(data_4.levels, nrow) # cantidad de observaciones para cada tratamiento
I <- length(data_4.levels) # cantidad de niveles
N <- sum(J) # número total de observaciones
samples <- t(sapply(data_4.levels, function(x) {x$response_time})) # matriz con todas las observaciones
treatment_means <- sapply(data_4.levels, function(x) {
  mean(x$response_time)
}) # media de cada tratamiento
residuals <- samples - treatment_means #residuos
SSE <- sum(residuals^2)
MSE <- SSE / (N - I)
display_markdown(sprintf('MSE = %.4f', MSE))
```

$$MSE = 32.5667$$

```
alpha <- 0.05
t <- qt(alpha/2, N-I, lower=FALSE) # distribución t de Student con alpha=0.05/2 y N-I grados de libertad
aux <- t * sqrt(MSE / J[1])
conf_int <- c(x_bar_3 - aux, x_bar_3 + aux) # intervalo de confianza
conf_int.df <- as.data.frame(cbind('Tratamiento 3', round(conf_int[1], 4), sample_mean, round(conf_int[2], 4)))
colnames(conf_int.df) <- c(' ', '2.5%', 'Media estimada', '97.5%)
```

```
display_markdown('#### **Intervalo de confianza del $\overline{X}$ para el tiempo de respuesta del tercer circuito:**')
rownames(conf_int.df) <- c()
as.htmlwidget(formattable(conf_int.df, align='c', list(' ' = formatter("span", style = ~ style(
  'font-weight'='bold', 'text-align'='left')))), width="50%")
```

Intervalo de confianza del 95% para el tiempo de respuesta del tercer circuito:

	2.5%	Media estimada	97.5%
Tratamiento 3	12.8394	145	23.9606

Ejercicio 5

Para investigar el efecto de la temperatura de secado del grano de trigo sobre la calidad del horneado del pan se realiza un experimento en donde se emplean tres niveles de temperatura, y la variable de respuesta medida es el volumen de la hogaza de pan producida. Los datos son los siguientes:

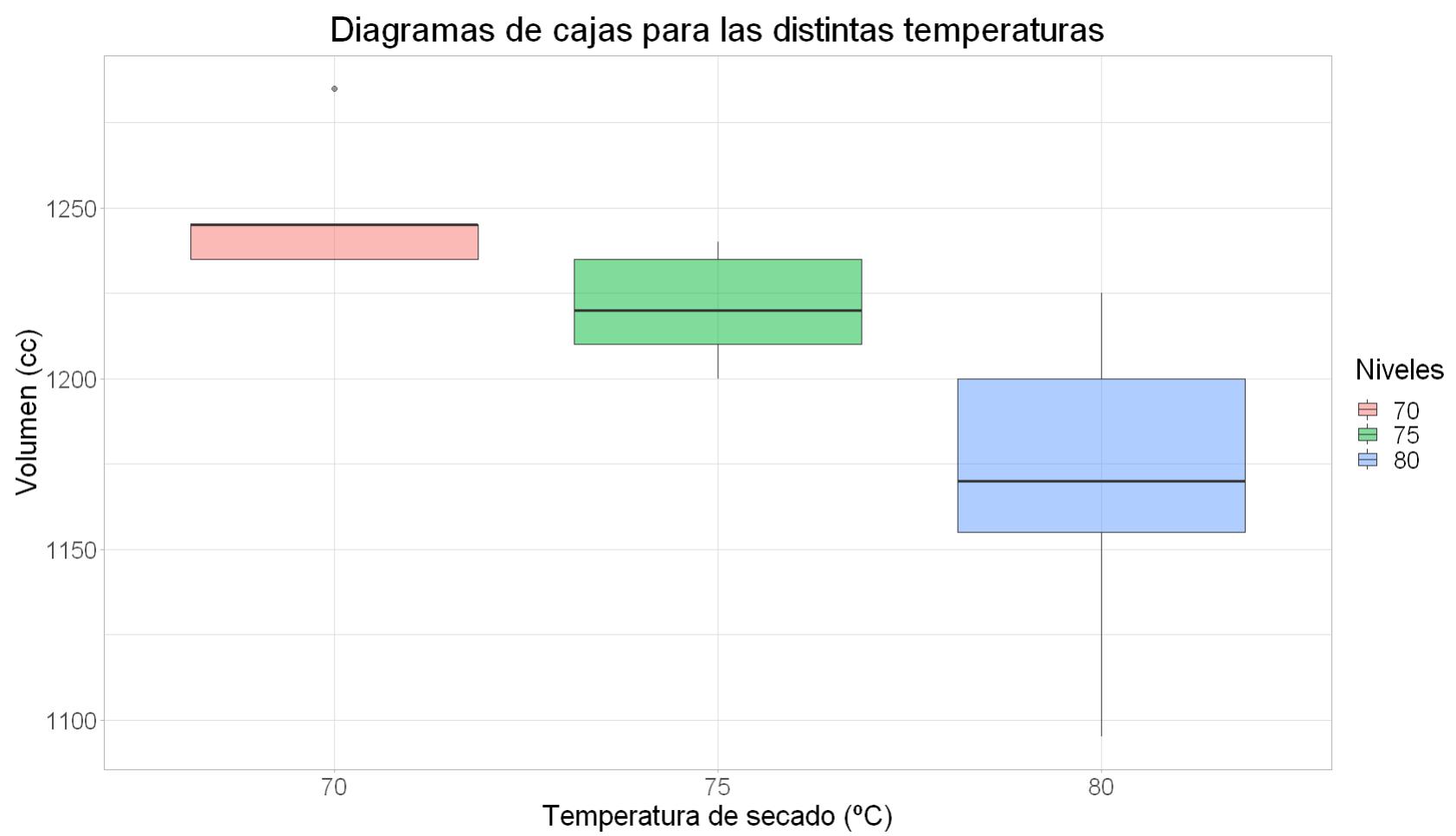
	Temperatura (°C)	Volumen (cc)				
	70.0	1245	1235	1285	1245	1235
	75.0	1235	1240	1200	1220	1210
	80.0	1225	1200	1170	1155	1095

```
data_5 <- read.csv("./TP4_tables/data5.csv") # Leo los datos desde archivo .csv
data_5$temperature <- factor(data_5$temperature)
```

a) ¿La temperatura de secado afecta el volumen promedio del pan? Utilice $\alpha = 0.01$.

```
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_5, aes(x=temperature, y=volume, fill=temperature)) +
  labs(
    title="Diagramas de cajas para las distintas temperaturas",
    x="Temperatura de secado (°C)",
    y="Volumen (cc)",
    fill="Niveles") +
  geom_boxplot(alpha=0.5) +
```

```
theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```



Los gráficos de cajas sugieren que la temperatura de secado afecta el volumen del pan. Las hipótesis de la prueba son:

$$H_0: \mu_1 = \mu_2 = \mu_3, \quad \text{contra} \quad H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

```
data_5.levels <- split(data_5 , f=data_5$temperature)
model_5 <- lm(volume ~ temperature, data_5)
```

```
# Cálculo de MStr
J <- sapply(data_5.levels, nrow) # cantidad de observaciones para cada tratamiento
I <- length(data_5.levels) # cantidad de niveles
SSTr <- sum((predict(model_5) - mean(data_5$volume))^2)
MStr <- SSTr / (I - 1)
display_markdown(sprintf('$MStr = %.2f$', MStr))
```

$$MStr = 8240.00$$

```
# Cálculo de MSE
N <- sum(J) # cantidad total de observaciones
SSE <- sum(model_5$residuals^2)
MSE <- SSE / (N - I)
display_markdown(sprintf('$MSE = %.2f$', MSE))
```

$$MSE = 1050.83$$

```
F <- MStr / MSE
display_markdown(sprintf('$F = %.4f$', F))
```

$$F = 7.8414$$

```
alpha <- 0.01
f_alpha <- qf(alpha, df1=I-1, df2=N-I, lower=FALSE)
display_markdown(sprintf('$f_{\alpha} = %.3f, f_{1-\alpha} = %.3f$', alpha, I-1, N-I, f_alpha))
```

$$f_{\alpha=0.010, 2, 12} = 6.9266$$

El valor del estadístico es $F = 7.8414 > 6.9266$. Por lo tanto, hay evidencia significativa en contra de la hipótesis nula y se concluye que la temperatura de secado del grano de trigo tiene un efecto sobre el volumen de la hogaza de pan.

b) Encuentre el p-valor de esta prueba.

```
p_value <- pf(F, df1=I-1, df2=N-I, lower=FALSE)
display_markdown(sprintf('$p-value = %.4f$', p_value))
```

$$p-value = 0.0066$$

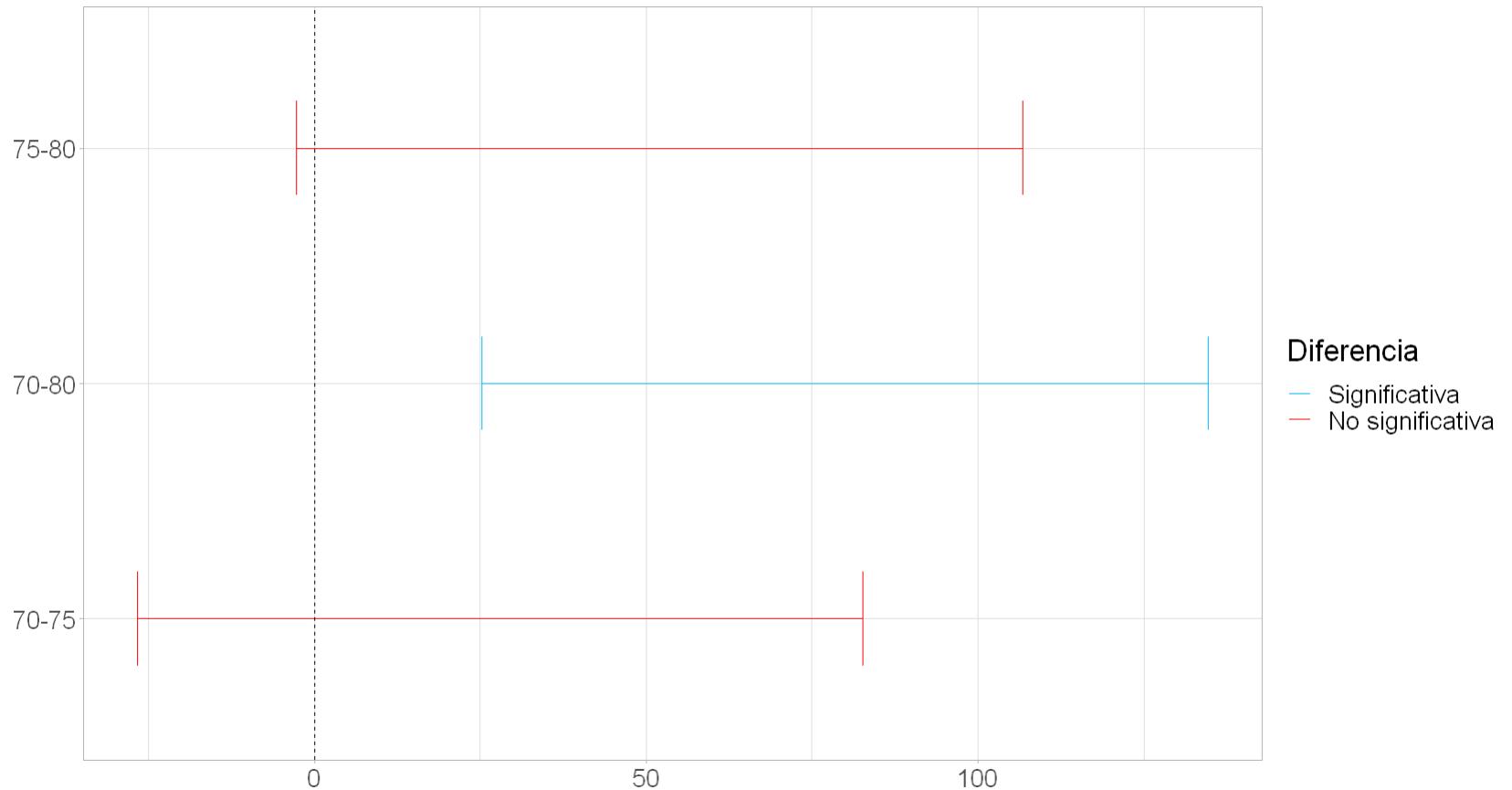
Como el p-valor = $0.0066 > 0.01$, se llega a la misma conclusión que en el inciso anterior.

c) Use el método de Tukey para identificar qué medias son diferentes.

```
data_5.aov <- aov(volume ~ temperature, data_5)
data_5.tukey <- as.data.frame(TukeyHSD(data_5.aov, ordered = TRUE, conf.level = 0.95)[1]$temperature)

data_5.tukey$names <- c(rownames(data_5.tukey))
# Gráfico de los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_5.tukey, aes(names, diff)) +
  labs(
    title="Intervalos de confianza de 95% para la diferencia de medias entre tratamientos",
    x="",
    y="",
    col="Diferencia") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(lwr*upr > 0,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5','#f50505'), labels=c('Significativa','No significativa'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```

Intervalos de confianza de 95% para la diferencia de medias entre tratamientos



El gráfico anterior indica que existe una diferencia entre las medias de los tratamientos correspondientes a las temperaturas 70°C y 80°C .

d) Analice los residuos de este experimento y haga un comentario sobre la adecuación del modelo.

```
smoothed <- data.frame(with(data_5.aov, lowess(x = data_5.aov$fitted, y = data_5.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_5.aov) +
  geom_point(aes(x=data_5.aov$fitted, y=data_5.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

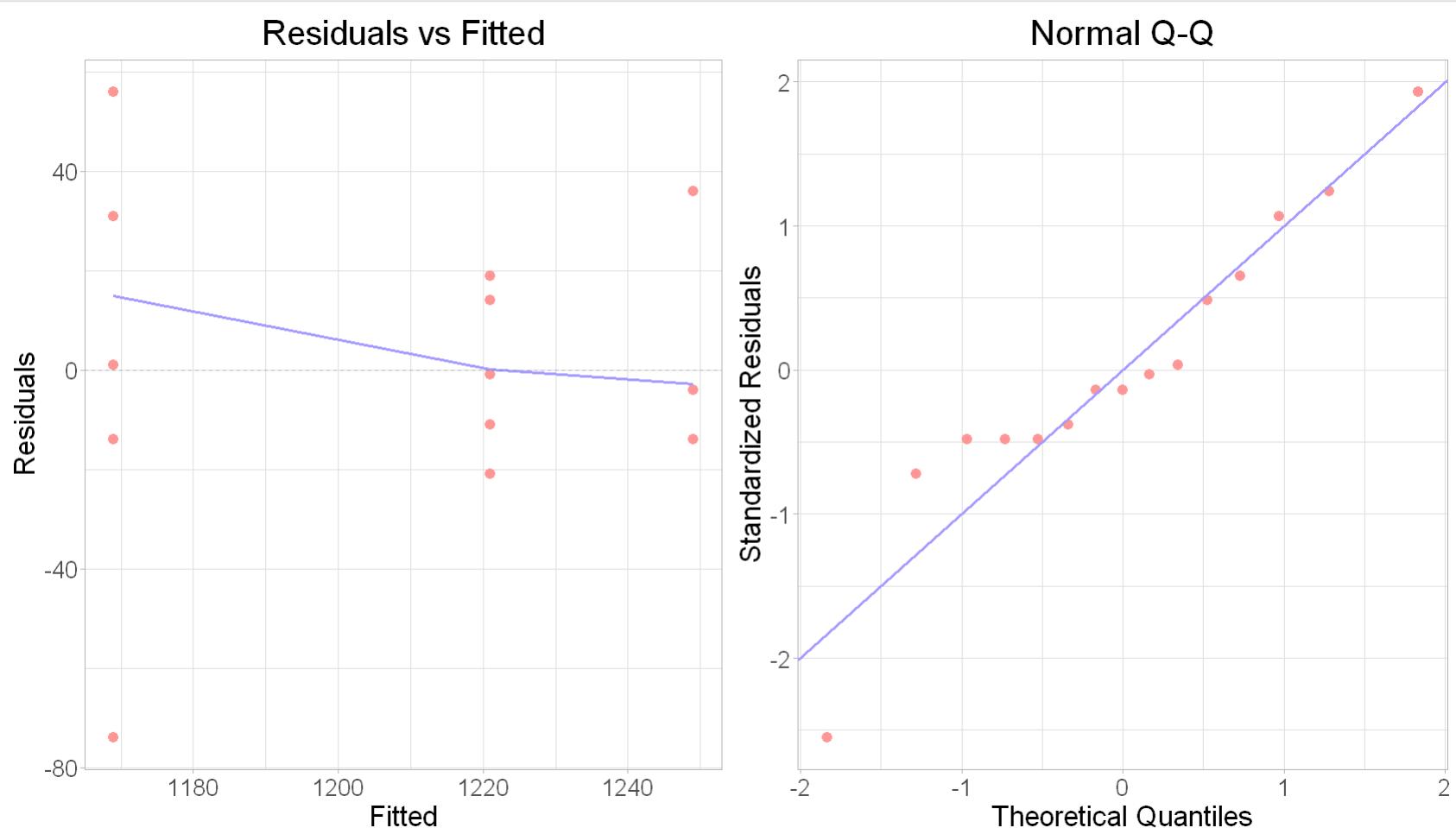
qq_plot <- ggplot(data_5.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
```

```

ggttitle("Normal Q-Q") +
theme_light() +
theme(text=element_text(size=20),
plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



La distribución de los valores en el gráfico de *Residuos vs. Valores Ajustados* no permite suponer que la varianza es constante. El gráfico de *Residuos estandarizados vs. Cuantiles teóricos* muestra desviaciones considerables respecto a la recta normal.

Ejercicio 6

Se realiza un experimento para determinar el efecto que tienen cuatro tipos diferentes de puntas de un probador de dureza sobre los valores de dureza observados de una aleación. Para ello se obtienen cuatro especímenes de aleación, y se prueba cada punta sobre cada uno de ellos. Los datos obtenidos son los siguientes:

Tipo de punta	Espécimen			
	1	2	3	4
1	9.3	9.4	9.6	10.0
2	9.4	9.3	9.8	9.9
3	9.2	9.4	9.5	9.7
4	9.7	9.6	10.0	10.2

```

data_6 <- read.csv("./TP4_tables/data6.csv") # Leo los datos desde archivo .csv
data_6$indenter <- factor(data_6$indenter)
data_6$specimen <- factor(data_6$specimen)

```

a) ¿Existe alguna diferencia en las mediciones de dureza entre las puntas? Utilice $\alpha = 0.01$.

Este experimento implica un diseño de bloques, donde el tipo de punta utilizada es el factor de interés y el espécimen sobre el cual se realiza la prueba es el factor bloqueado.

```

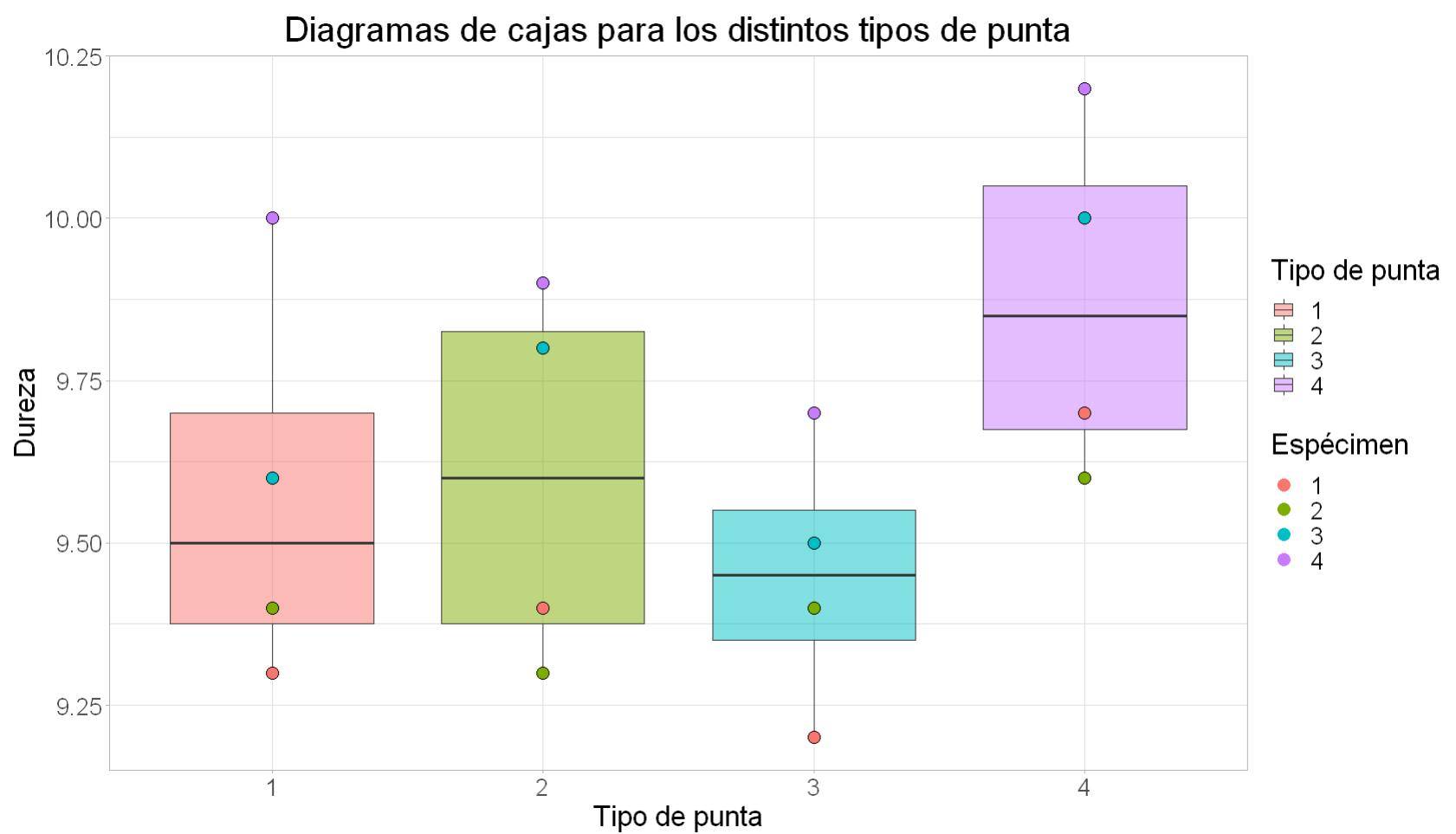
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_6, aes(x=indenter, y=hardness, fill=indenter)) +
  labs(
    title="Diagramas de cajas para los distintos tipos de punta",
    x="Tipo de punta",
    y="Dureza",
    fill="Punta")

```

```

    fill="Tipo de punta",
    col="Espécimen") +
  geom_boxplot(alpha=0.5, aes(fill=indenter)) +
  geom_point(aes(col=specimen), size=4) +
  geom_point(size=4, shape=1) +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5)) +
  guides(fill = guide_legend(override.aes = list(shape = NA), order = 1))

```



Los gráficos de cajas sugieren que el tipo de punta utilizado podría no afectar significativamente a la lectura de dureza del espécimen. Las hipótesis de la prueba son:

$$H_0: \mu_1 = \mu_2 = \mu_3, \quad \text{contra} \quad H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

```

data_6.aov <- aov(hardness ~ indenter + specimen, data_6)
aov_test6 <- summary(data_6.aov)[[1]]

```

```

display_markdown('#### **ANOVA de dos sentidos: Dureza vs Tipo de punta + Espécimen (bloqueado)**')
display_markdown('\n')
aov_test6 <- cbind(c('Tipo de punta', 'Espécimen', 'Residuos'), aov_test6)
colnames(aov_test6)[1] <- 'Source'
rownames(aov_test6) <- c()
table <- formattable(aov_test6, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style(
  'text-align='left'))))
as.htmlwidget(table, width="70%", height=NULL)

```

ANOVA de dos sentidos: Dureza vs Tipo de punta + Espécimen (bloqueado)

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tipo de punta	3	0.385	0.128333333	14.4375	0.0008712721
Espécimen	3	0.825	0.275000000	30.9375	0.0000452327
Residuos	9	0.080	0.008888889	NA	NA

El p-valor para el factor **Tipo de punta** es $0.0009 < 0.01$. Por lo tanto, se rechaza la hipótesis nula y se concluye que el tipo de punta utilizada afecta al valor obtenido en la prueba de dureza.

b) Use el método de Tukey para identificar diferencias específicas entre las puntas. Analice los residuos de este experimento.

```

data_6.tukey <- as.data.frame(TukeyHSD(data_6.aov, ordered = TRUE, conf.level = 0.95)[1]$indenter)

```

```

data_6.tukey$names <- c(rownames(data_6.tukey))
# Gráfico de Los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_6.tukey, aes(names, diff)) +

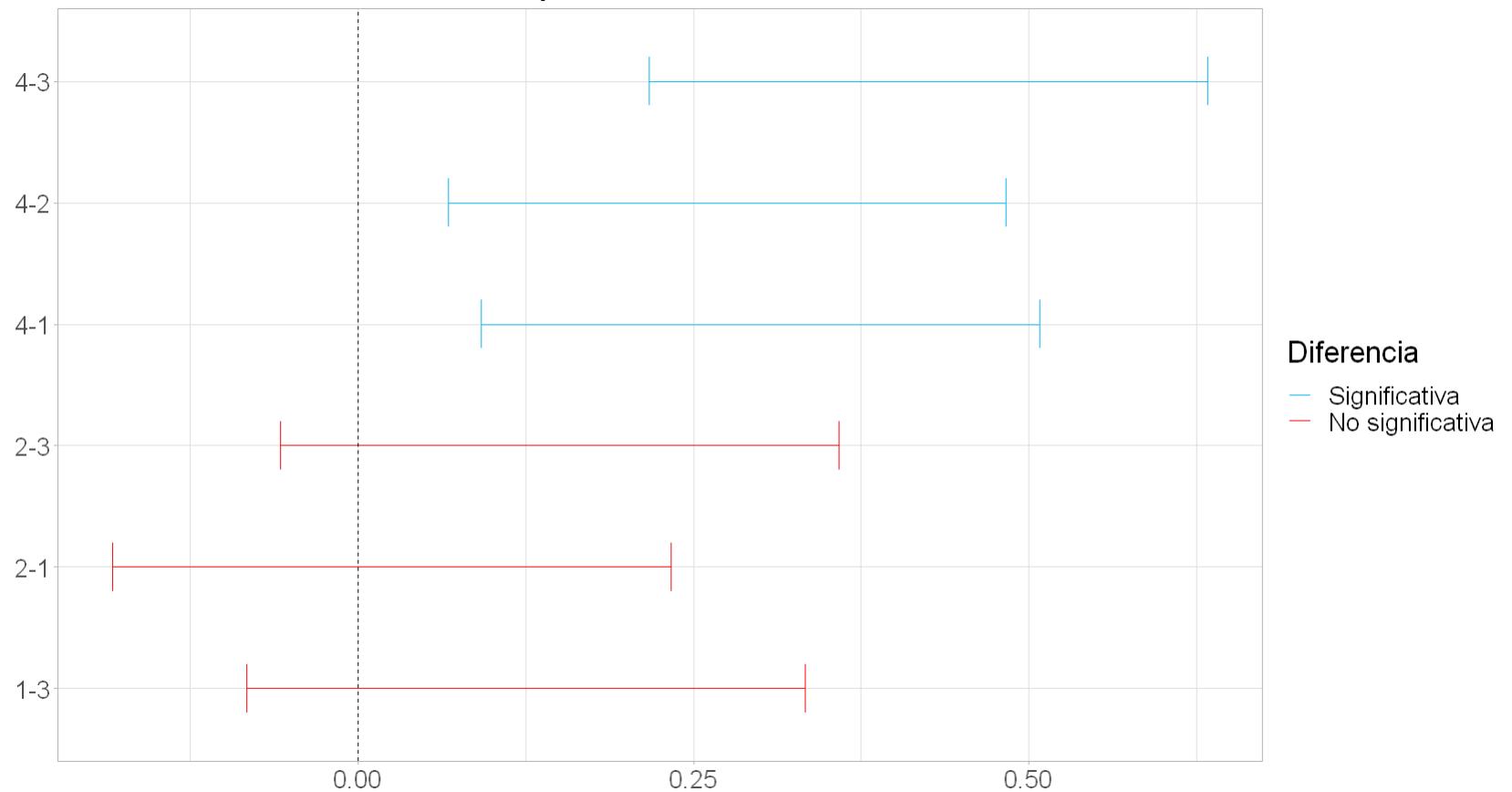
```

```

  labs(
    title="Intervalos de confianza de 95% para la diferencia de medias entre tratamientos",
    x="",
    y="",
    col="Diferencia") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(lwr*upr > 0,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5','#f50505'), labels=c('Significativa','No significativa'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

```

Intervalos de confianza de 95% para la diferencia de medias entre tratamientos



Los intervalos de confianza obtenidos indican que la punta tipo 4 difiere del resto de las puntas.

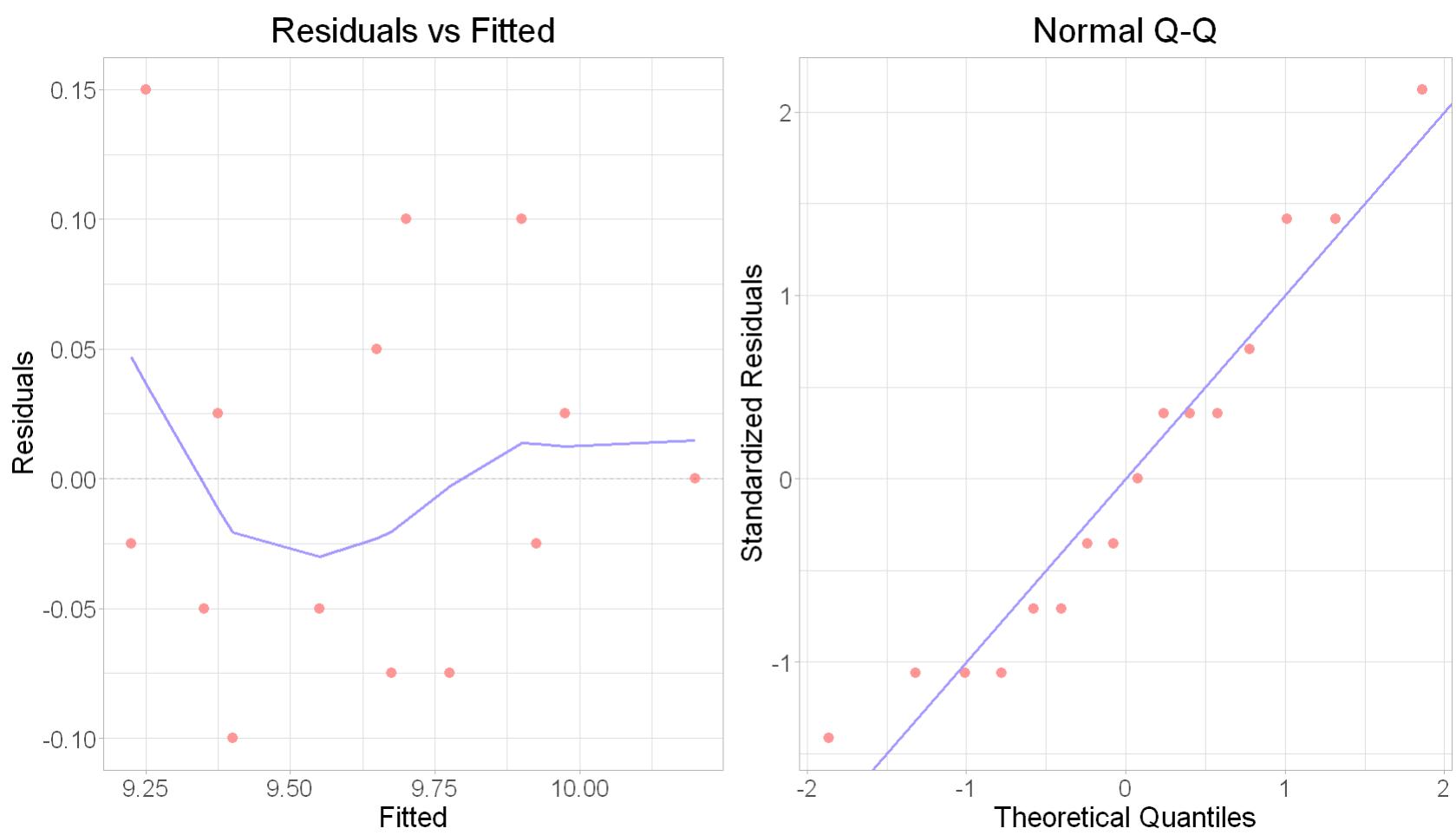
```

smoothed <- data.frame(with(data_6.aov, lowess(x = data_6.aov$fitted, y = data_6.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_6.aov) +
  geom_point(aes(x=data_6.aov$fitted, y=data_6.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_6.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggtitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



Los valores en el gráfico de *Residuos vs. Valores Ajustados* exhiben una dispersión relativamente uniforme, por lo que no hay evidencias fuertes en contra del supuesto de varianza constante. El gráfico de *Residuos estandarizados vs. Cuantiles teóricos* no muestra desviaciones notables en torno a la recta normal.

Ejercicio 7

Se estudian diferentes algoritmos para estimar los costos de desarrollo de software. Para esto se aplican seis algoritmos a ocho proyectos de desarrollo de software y se observa el porcentaje de error al estimar los costos de desarrollo. Los datos son los siguientes:

Algoritmo	Proyecto							
	1	2	3	4	5	6	7	8
1	1244	21	82	2221	905	839	527	122
2	281	129	396	1306	336	910	473	199
3	220	84	458	543	300	794	488	142
4	225	83	425	552	291	826	509	153
5	19	11	-34	121	15	103	87	-17
6	-20	35	-53	170	104	199	142	41

```
data_7 <- read.csv("./TP4_tables/data7.csv") # Leo los datos desde archivo .csv
data_7$algorithm <- factor(data_7$algorithm)
data_7$project <- factor(data_7$project)
```

a) ¿Existe alguna diferencia entre los algoritmos en cuanto a la exactitud de la estimación del costo promedio? Utilice $\alpha = 0.05$.

Este experimento requiere un diseño de bloques, donde el algoritmo utilizado es el factor de interés y el proyecto sobre el cual se aplica es el factor bloqueado.

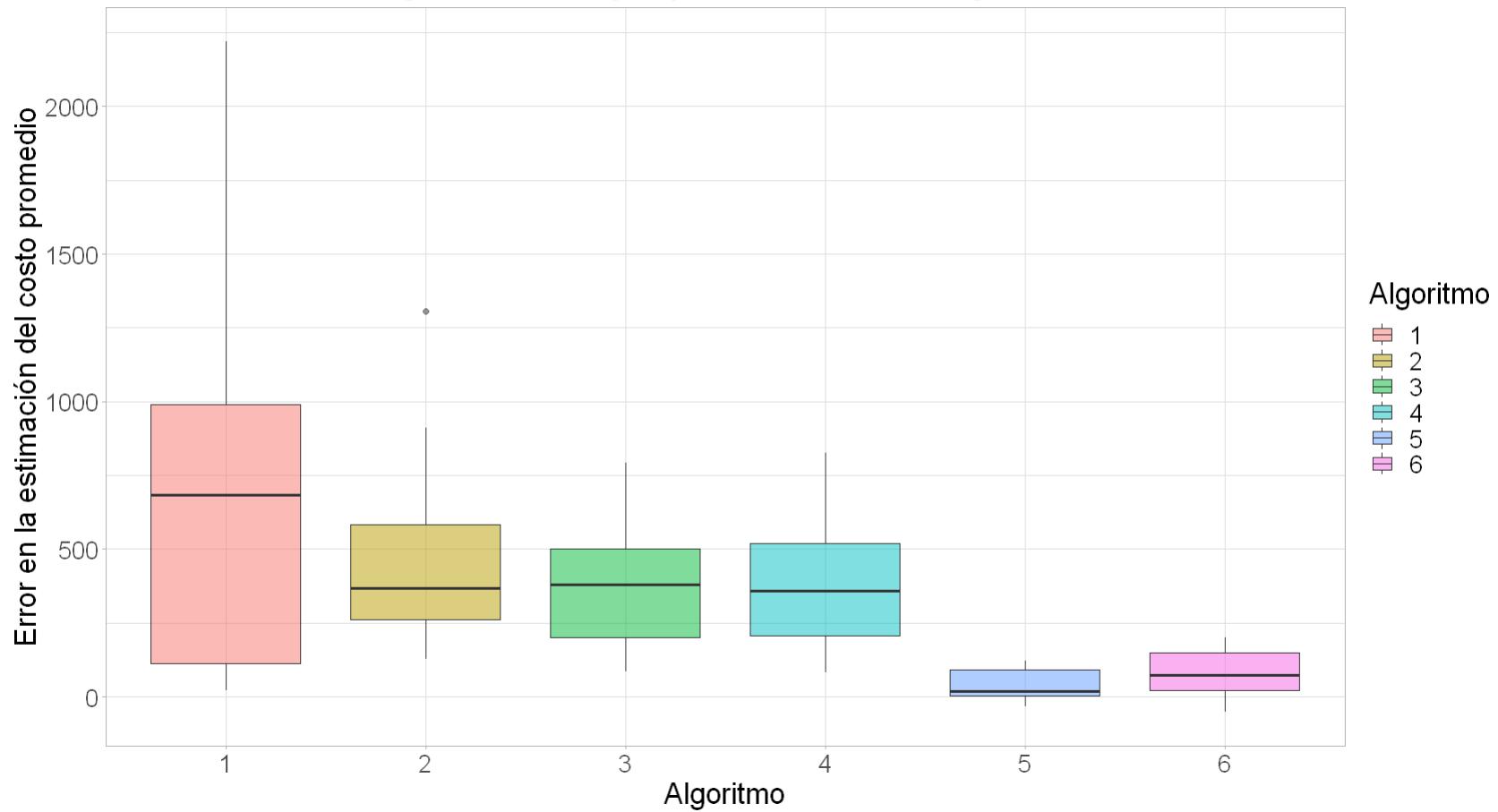
```
# Plot
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_7, aes(x=algorithm, y=estimation_error, fill=algorithm)) +
  labs(
    title="Diagramas de cajas para los distintos algoritmos",
    x="Algoritmo",
    y="Error en la estimación del costo promedio",
    fill="Algoritmo",
```

```

    col="Proyecto") +
  geom_boxplot(alpha=0.5, aes(fill=algorithm)) +
  #geom_point(aes(col=project), size=4, alpha=0.5) +
  #geom_point(shape=1, size=4) +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5)) +
  guides(fill = guide_legend(override.aes = list(shape = NA), order = 1))

```

Diagramas de cajas para los distintos algoritmos



Los gráficos de cajas sugieren que la exactitud de la estimación obtenida con los algoritmos 5 y 6 podría diferir de la del resto de los algoritmos. Las hipótesis de la prueba son:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6, \quad \text{contra } H_1: \text{dos o mas } \mu_i \text{ son diferentes}$$

```

data_7.aov <- aov(estimation_error ~ algorithm + project, data_7)
aov_test7 <- summary(data_7.aov)[[1]]

```

```

display_markdown('#### **ANOVA de dos sentidos: Error de estimación vs Algoritmo + Proyecto (bloqueado)**')
display_markdown('\n')
aov_test7 <- cbind(c('Algoritmo', 'Proyecto', 'Residuos'), aov_test7)
colnames(aov_test7)[1] <- 'Source'
rownames(aov_test7) <- c()
table <- formattable(aov_test7, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style(
  'text-align'='left'))))
as.htmlwidget(table, width="70%", height=NULL)

```

ANOVA de dos sentidos: Error de estimación vs Algoritmo + Proyecto (bloqueado)

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Algoritmo	5	2825746	565149.14	6.229422	0.0003135205
Proyecto	7	2710323	387188.97	4.267835	0.0016829800
Residuos	35	3175290	90722.57	NA	NA

El p-valor para el factor **Algoritmo** es $0.0003 < 0.05$. Por lo tanto, se rechaza la hipótesis nula y se concluye que el algoritmo aplicado afecta al error de estimación del costo de desarrollo.

b) Analice los residuos de este experimento.

```

smoothed <- data.frame(with(data_7.aov, lowess(x = data_7.aov$fitted, y = data_7.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_7.aov) +
  geom_point(aes(x=data_7.aov$fitted, y=data_7.aov$residuals), color="#ff9696", size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +

```

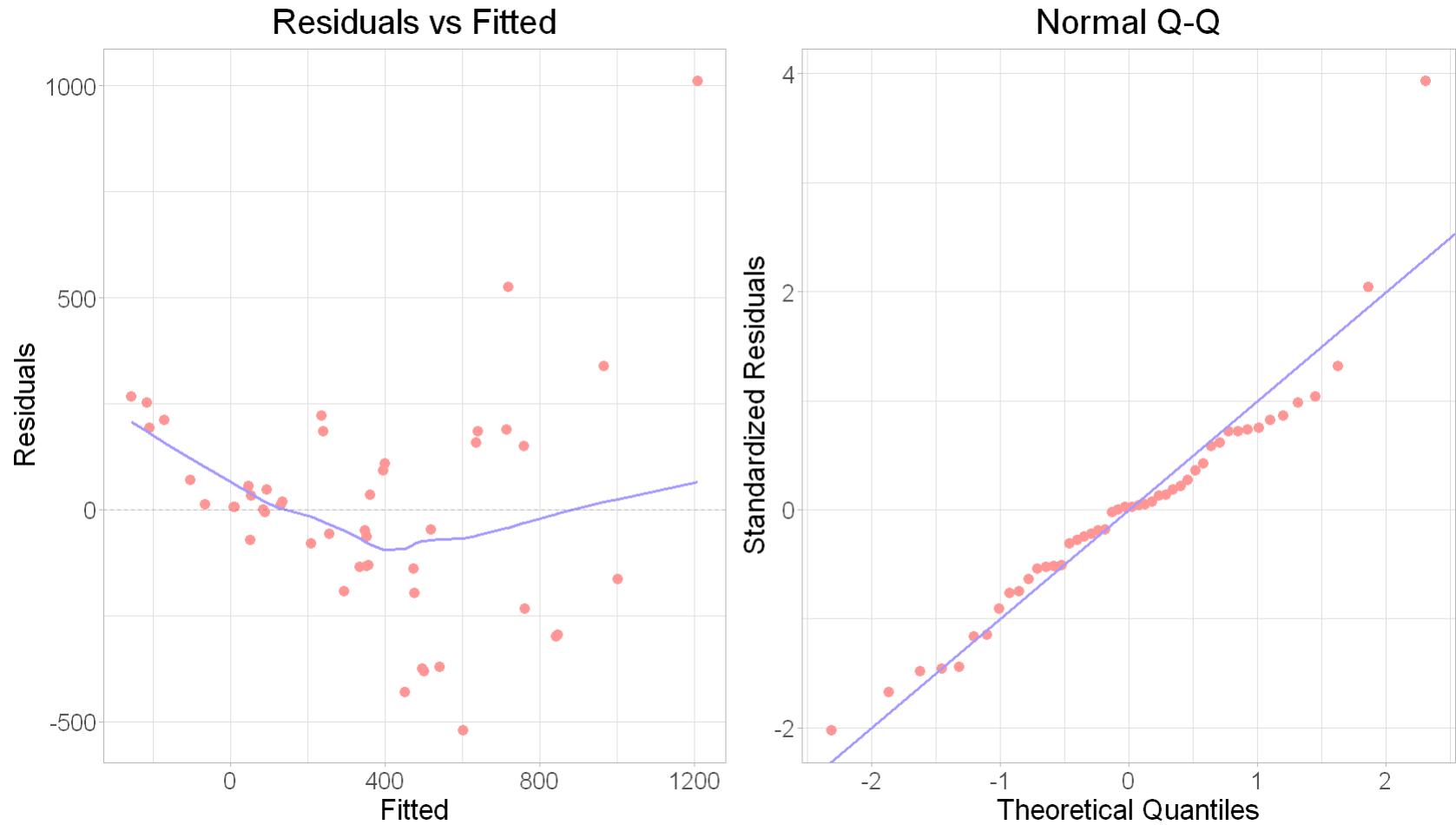
```

xlab("Fitted") +
ylab("Residuals") +
theme_light() +
theme(text=element_text(size=20),
plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_7.aov) +
stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
geom_abline(col="#a399ff", size=1) +
xlab("Theoretical Quantiles") +
ylab("Standardized Residuals") +
ggtitle("Normal Q-Q") +
theme_light() +
theme(text=element_text(size=20),
plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



El gráfico de *Residuos vs. Valores Ajustados* muestra una mayor dispersión de los valores en el lado derecho, lo cual sugiere que la varianza no es constante. El gráfico de *Residuos estandarizados vs. Cuantiles teóricos* no muestra evidencias importantes en contra del supuesto de normalidad.

c) ¿Qué algoritmo recomendaría para usarlo en la práctica?

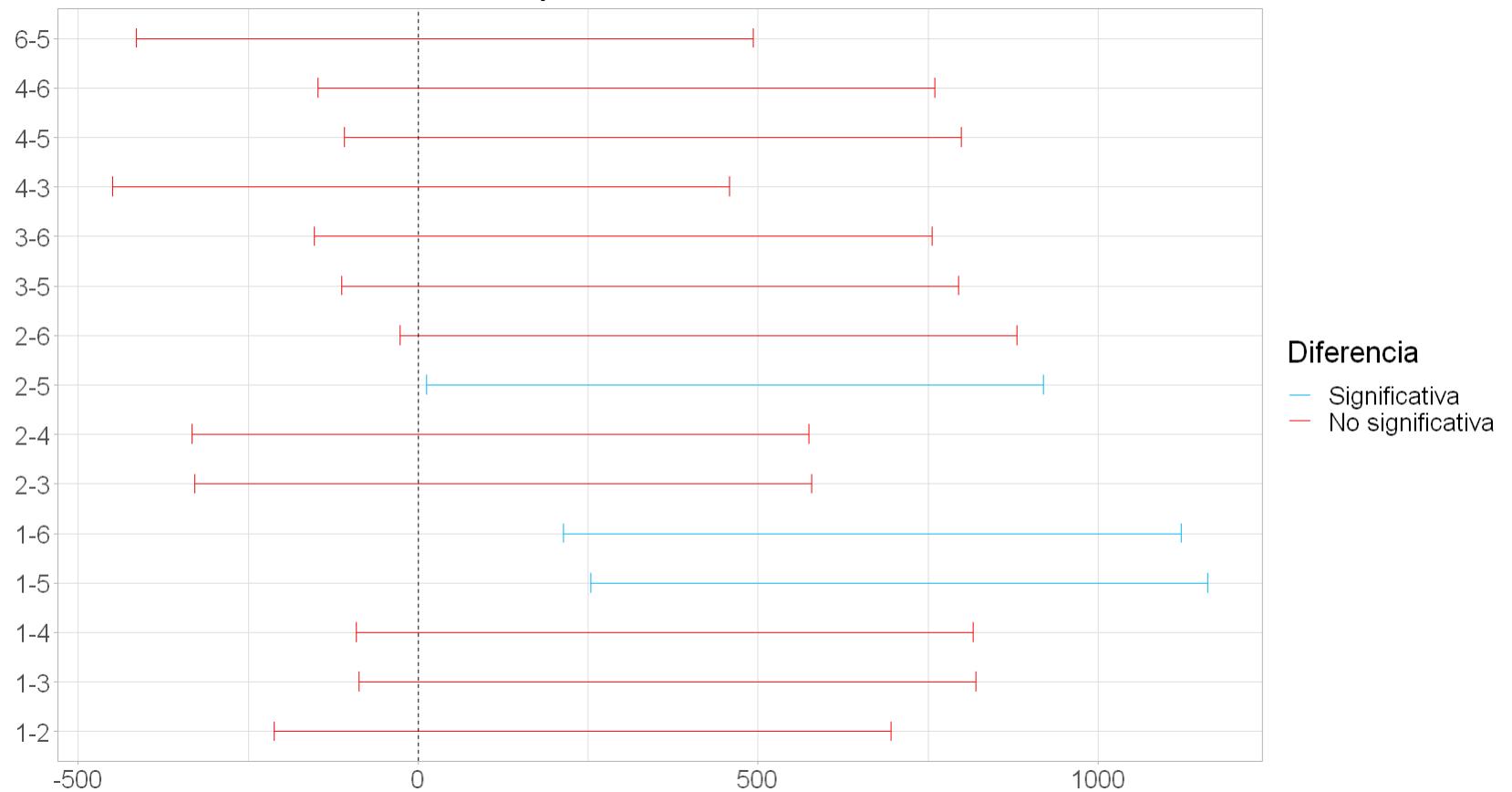
```

data_7.tukey <- as.data.frame(TukeyHSD(data_7.aov, ordered = TRUE, conf.level = 0.95)[1]$algorithm)

data_7.tukey$names <- c(rownames(data_7.tukey))
# Gráfico de los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_7.tukey, aes(names, diff)) +
  labs(
    title="Intervalos de confianza de 95% para la diferencia de medias entre tratamientos",
    x="",
    y="",
    col="Diferencia") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(lwr*upr > 0,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5','#f50505'), labels=c('Significativa','No significativa'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
plot.title = element_text(size=24, hjust = 0.5))

```

Intervalos de confianza de 95% para la diferencia de medias entre tratamientos



Una forma de determinar con cuáles de los algoritmos se obtienen mejores resultados es plantear la hipótesis:

$$H_0: \mu_i = 0 \quad \text{contra} \quad H_1: \mu_i \neq 0$$

donde $i = 1, \dots, 6$. Para ello se construyen intervalos de confianza con $\alpha = 0.05$ para las medias de cada tratamiento y se observa cuales contienen al cero. Un intervalo que incluye al cero implica que no se puede afirmar (con un nivel de significancia α) que existe un error en la estimación de los costos de desarrollo.

```
data_7.levels <- split(data_7, f=data_7$algorithm)
x_bar <- sapply(data_7.levels, function(x) {
  mean(x$estimation_error)
}) # media de cada tratamiento
```

```
# Cálculo de MSE
model_7 <- lm(estimation_error ~ algorithm + project, data_7)
J <- sapply(data_7.levels, nrow) # cantidad de observaciones para cada tratamiento
I <- length(data_7.levels) # cantidad de niveles
N <- sum(J) # cantidad total de observaciones
SSE <- sum(model_7$residuals^2)
MSE <- SSE / (N - I)
display_markdown(sprintf('$MSE = %.2f$', MSE))
```

$MSE = 75602.14$

```
alpha <- 0.05
t <- qt(alpha/2, N-I, lower=FALSE) # distribución t de Student con alpha=0.05/2 y N-I grados de libertad
aux <- t * sqrt(MSE / J)
conf_int <- matrix(c(x_bar - aux, x_bar + aux), ncol=2, byrow=FALSE) # intervalos de confianza
```

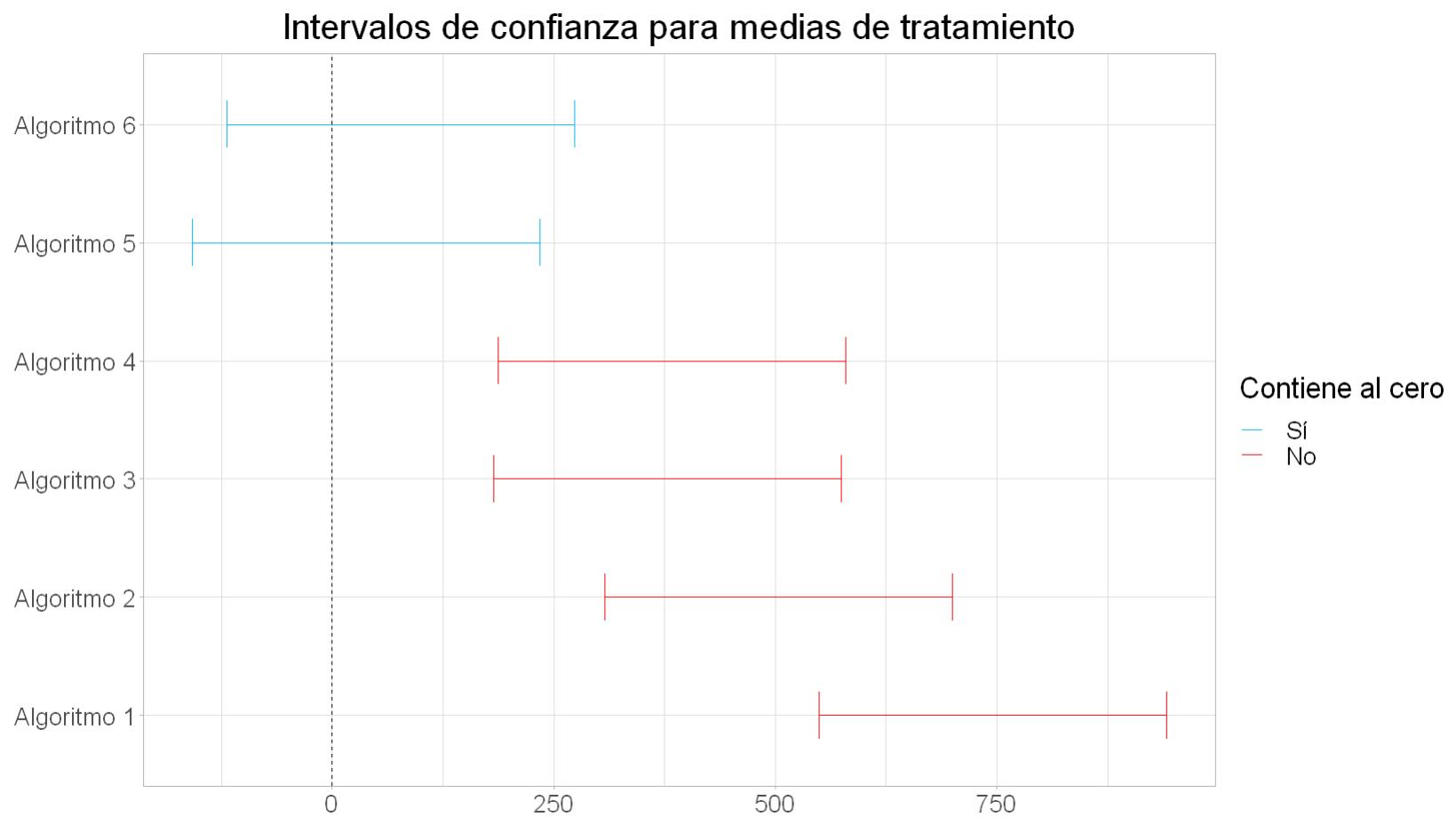
```
conf_int.df <- as.data.frame(data.frame(paste("Algoritmo", 1:6, sep=" "), conf_int[,1], x_bar, conf_int[,2]))
colnames(conf_int.df) = c(" ", "2.5%", "Media estimada", "97.5%")
contains_zero <- conf_int.df[, "2.5%"] * conf_int.df[, "97.5%"] <= 0 # TRUE si el intervalo contiene al cero
conf_int.df <- cbind(conf_int.df, contains_zero)
colnames(conf_int.df)[5] <- "Contiene al 0"
```

```
display_markdown('#### **Intervalos de confianza para medias de tratamiento**')
display_markdown('\n')
table <- formattable(conf_int.df, align=c('l', 'c', 'c', 'c', 'c'), list(`Contiene al 0` = formatter("span",
x ~ ifelse(x, "Sí", "No"),
style = x ~ style(color = ifelse(x, "green", "red"))), `Media estimada` = formatter("span", style = ~ style("font.weight" = "bold")),
` ` = formatter("span", style = ~ style("font.weight" = "bold"))))
#table <- format_table(table, list(area(1:2) ~ color_tile("transparent", "Lightgray")))
as.htmlwidget(table, width="50%", height=NULL)
```

Intervalos de confianza para medias de tratamiento

	2.5%	Media estimada	97.5%	Contiene al 0
Algoritmo 1	548.9423	745.125	941.3077	No
Algoritmo 2	307.5673	503.750	699.9327	No
Algoritmo 3	182.4423	378.625	574.8077	No
Algoritmo 4	186.8173	383.000	579.1827	No
Algoritmo 5	-158.0577	38.125	234.3077	Sí
Algoritmo 6	-118.9327	77.250	273.4327	Sí

```
colnames(conf_int.df) <- c("algorithm", "lwr", "mean", "upr", "contains_zero")
# Gráfico de los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(conf_int.df, aes(algorithm)) +
  labs(
    title="Intervalos de confianza para medias de tratamiento",
    x="",
    y="",
    col="Contiene al cero") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(contains_zero==TRUE,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5','#f50505'), labels=c(' Sí',' No'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))
```



Los resultados anteriores sugieren que los algoritmos 5 y 6 ofrecen mejor exactitud en la estimación del costo promedio que el resto. Entre ellos, el algoritmo 5 presenta el menor valor medio para el error de estimación.

Ejercicio 8

Se presentan los resultados de un experimento relacionado con la capacidad de una batería utilizada en el mecanismo de lanzamiento de un lanzacohetes tierra-aire. Las placas de la batería pueden fabricarse con tres materiales. El objetivo es diseñar una batería que no se vea afectada por la temperatura ambiente. La respuesta de la batería es la vida efectiva de ésta en horas. Para esto se fijan tres niveles de temperatura y se realiza un experimento factorial con cuatro réplicas.

Material	Temperatura (°F)					
	baja		media		alta	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	136	122	25	70

Material	Temperatura (°F)					
	baja	media	alta			
3	159	126	106	115	58	45
	138	110	174	120	96	104
	168	160	150	139	82	60

```
data_8 <- read.csv("./TP4_tables/data8.csv") # Leo Los datos desde archivo .csv
data_8$material <- factor(data_8$material)
data_8$temperature <- factor(data_8$temperature)
```

Se trata de un experimento de dos factores, donde el material es el factor fila y la temperatura es el factor columna (siguiendo la distribución de la tabla provista), ambos con 3 niveles. Cada uno de los 9 tratamientos cuenta con 4 réplicas.

a) Pruebe las hipótesis apropiadas y obtenga conclusiones mediante el empleo del análisis de varianza con $\alpha = 0.05$.

```
options(dplyr.summarise.inform = FALSE)
data_8.mean <- data_8 %>%
  group_by(material, temperature) %>%
  summarise(cell_mean=mean(lifespan)) # media de cada tratamiento

data_8.mean <- data_8.mean %>%
  group_by(material) %>%
  summarise(material_mean=mean(cell_mean), across()) # media de cada material

data_8.mean <- data_8.mean %>%
  group_by(temperature) %>%
  summarise(temperature_mean=mean(cell_mean), across()) # media de cada nivel de temperatura
```

```
data_8.mean <- data_8.mean %>%
  arrange(match(temperature, c("low", "mid", "high"))) %>%
  arrange(match(material, c(1, 2, 3))) %>%
  select(material, temperature, cell_mean, material_mean, temperature_mean)
```

```
eng_colnames <- colnames(data_8.mean)
colnames(data_8.mean) <- c("material", "temperatura", "media de la celda", "media de la fila", "media de la columna")
as.htmlwidget(formattable(data_8.mean, align="c"), width="80%", height=NULL)
```

material	temperatura	media de la celda	media de la fila	media de la columna
1	low	134.75	83.16667	144.83333
1	mid	57.25	83.16667	107.58333
1	high	57.50	83.16667	64.16667
2	low	155.75	108.33333	144.83333
2	mid	119.75	108.33333	107.58333
2	high	49.50	108.33333	64.16667
3	low	144.00	125.08333	144.83333
3	mid	145.75	125.08333	107.58333
3	high	85.50	125.08333	64.16667

Un análisis de varianza de dos sentidos está diseñado para responder tres preguntas principales:

1. ¿El modelo aditivo vale?
2. ¿Si es así, la media del resultado es la misma para todos los niveles del factor fila?
3. ¿Si es así, la media del resultado es la misma para todos los niveles del factor columna?

1. Para probar si el modelo aditivo vale se prueba la hipótesis nula de que todas las interacciones son iguales a 0:

$$H_0: \gamma_{11} = \gamma_{12} = \dots = \gamma_{IJ} = 0$$

2. Para probar si la media del resultado es igual para todos los niveles del factor renglón, se prueba la hipótesis nula de que todos los efectos renglón son iguales a 0:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_I = 0$$

3. Para probar si la media del resultado es igual para todos los niveles del factor columna, se prueba la hipótesis nula de que todos los

efectos columna son iguales a 0:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_J = 0$$

```
data_8.aov <- summary(aov(lifespan ~ material * temperature, data_8))[[1]]
```

```
display_markdown('#### **ANOVA de dos sentidos**')
display_markdown('\n')
data_8.aov <- cbind(c('Material', 'Temperatura', 'Interacción', 'Residuos'), data_8.aov)
colnames(data_8.aov)[1] <- 'Source'
rownames(data_8.aov) <- c()
data_8.aov["Pr(>F)"] <- round(data_8.aov["Pr(>F)"], 4)
table <- formattable(data_8.aov, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style
('text-align'='left'))))
as.htmlwidget(table, width="70%", height=NULL)
```

ANOVA de dos sentidos

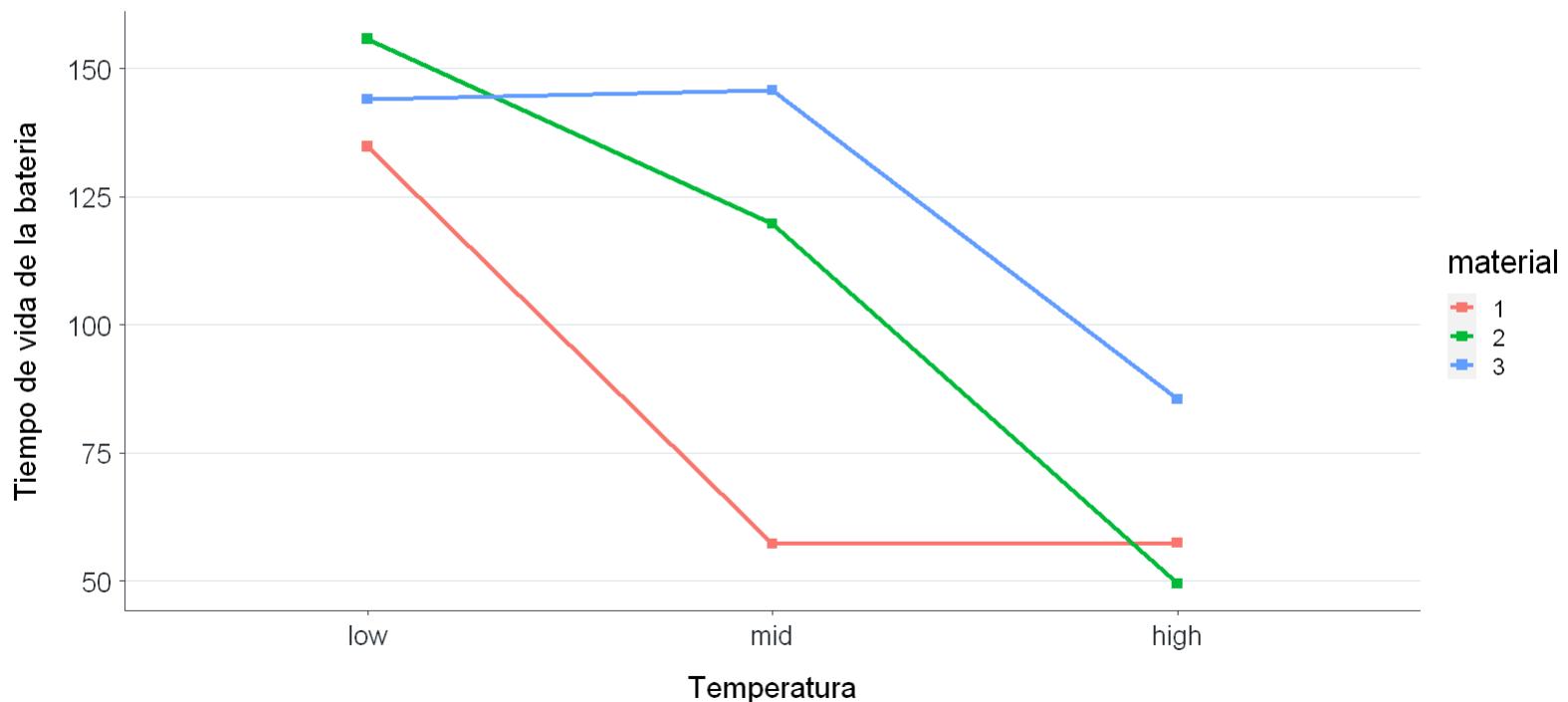
Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Material	2	10683.722	5341.861	7.911372	0.0020
Temperatura	2	39118.722	19559.361	28.967692	0.0000
Interacción	4	9613.778	2403.444	3.559535	0.0186
Residuos	27	18230.750	675.213	NA	NA

El p-valor para la interacción es 0.00152 < 0.05. Por lo tanto, se rechaza la hipótesis nula de que todas las interacciones son iguales a 0 y se concluye que el modelo no es aditivo. Este resultado implica que no se pueden utilizar las pruebas de hipótesis para los efectos principales.

b) Analice de manera gráfica la interacción.

```
data.frame(data_8) %>%
  group_by(material, temperature) %>%
  summarise(cell_mean=mean(lifespan)) %>%
  arrange(match(temperature, c("low", "mid", "high"))) %>%
  mutate(temperature=factor(temperature, levels=temperature)) %>%
  ggplot(aes(temperature, cell_mean)) +
  geom_line(size = 1.2, aes(group = material, color = material)) +
  geom_point(size = 2.6, aes(color = material), shape = 15) +
  labs(
    title = "Interacción entre nivel de temperatura y material",
    subtitle = "\n",
    x = "Temperatura",
    y = "Tiempo de vida de la batería") +
  theme(
    #text = element_text(family = "Roboto"),
    plot.margin = unit(rep(1.2, 4), "cm"),
    plot.title = element_text(size = 22, color = "#22292F", face = "bold", margin = margin(b = 5)),
    plot.subtitle = element_text(size = 15, margin = margin(b = 35)),
    plot.caption = element_text(size = 16, margin = margin(t = 25), color = "#606F7B"),
    panel.background = element_blank(),
    axis.text = element_text(size = 16, color = "#22292F"),
    axis.text.x = element_text(margin = margin(t = 5)),
    axis.text.y = element_text(margin = margin(r = 5)),
    axis.line = element_line(color = "#3D4852"),
    axis.title = element_text(size = 18),
    axis.title.y = element_text(margin = margin(r = 15),
                                hjust = 0.5),
    axis.title.x = element_text(margin = margin(t = 15),
                                hjust = 0.5),
    panel.grid.major = element_line(color = "#DAE1E7"),
    panel.grid.major.x = element_blank(),
    legend.title = element_text(size=20), #change legend title font size
    legend.text = element_text(size=14)) #change legend text font size
```

Interacción entre nivel de temperatura y material



El gráfico de interacción sugiere que el modelo no es aditivo. Algunas observaciones:

- Las baterías construidas con el **material 1** muestran un mayor tiempo de vida a temperaturas medias, mientras que el resto de los materiales presentan mayor duración a temperaturas bajas.
- El tiempo de vida de las baterías construidas con el **material 2** desciende a medida que aumenta la temperatura.
- No hay diferencias significativas entre la duración a temperaturas medias y altas para las baterías construidas con el **material 3**.

c) Analice los residuos de este experimento.

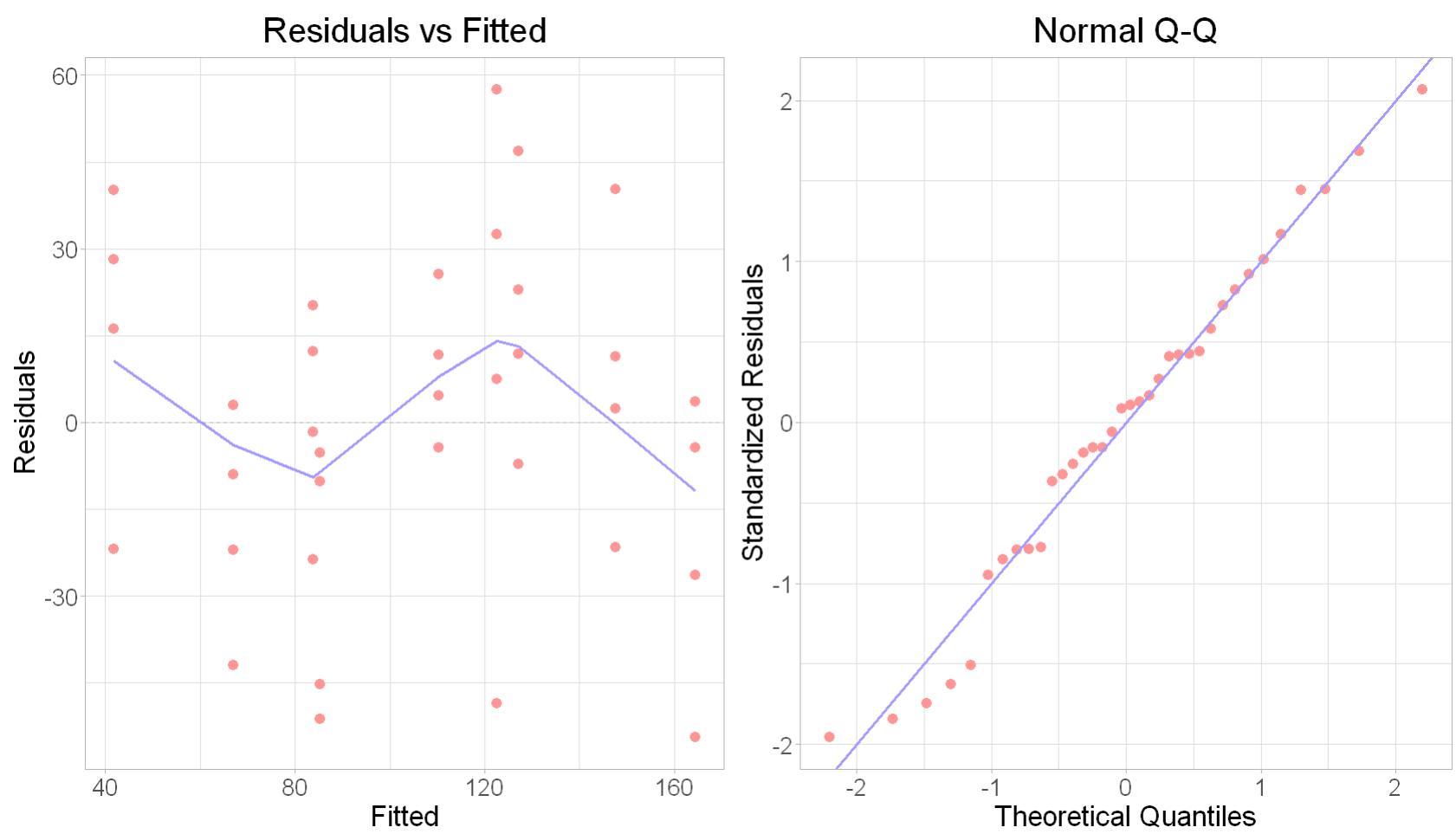
```

data_8.aov <- aov(lifespan ~ material + temperature, data_8)
smoothed <- data.frame(with(data_8.aov, lowess(x = data_8.aov$fitted, y = data_8.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_8.aov) +
  geom_point(aes(x=data_8.aov$fitted, y=data_8.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
  plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_8.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggtitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
  plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



La gráfica de *Residuos vs. Valores Ajustados* no presenta evidencias fuertes contra el supuesto de las variancias iguales, y la gráfica de *Probabilidad Normal* no exhibe desviaciones notables respecto de la recta. Por lo tanto, es razonable asumir que el modelo es válido.

Ejercicio 9

Se efectúa un experimento para determinar si la temperatura de encendido o la posición de la caldera tienen efecto sobre la densidad de un ánodo de carbón.

Posición	Temperatura (°C)		
	800	825	850
1	570	1063	565
	565	1080	510
	583	1043	590
2	528	988	526
	547	1026	538
	521	1004	532

```
data_9 <- read.csv("./TP4_tables/data9.csv") # Leo Los datos desde archivo .csv
data_9$position <- factor(data_9$position)
data_9$temperature <- factor(data_9$temperature)
```

Se trata de un experimento de dos factores, donde la posición es el factor fila con 2 niveles y la temperatura es el factor columna con 3 niveles. Cada uno de los 6 tratamientos cuenta con 3 réplicas.

a) Establezca las hipótesis de interés y

b) Pruebe las hipótesis anteriores utilizando análisis de varianza con $\alpha = 0.05$.

1. Para probar si el modelo aditivo vale se prueba la hipótesis nula de que todas las interacciones son iguales a 0:

$$H_0: \gamma_{11} = \gamma_{12} = \dots = \gamma_{IJ} = 0$$

```
data_9.aov <- summary(aov(anode_density ~ position * temperature, data_9))[[1]]
```

```
display_markdown('#### **ANOVA de dos sentidos**')
display_markdown('\n')
data_9.aov <- cbind(c('Posición', 'Temperatura', 'Interacción', 'Residuos'), data_9.aov)
```

```

colnames(data_9.aov)[1] <- 'Source'
rownames(data_9.aov) <- c()
data_9.aov["Pr(>F)"] <- round(data_9.aov["Pr(>F)"], 4)
table <- formattable(data_9.aov, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style
('text-align'='left'))))
as.htmlwidget(table, width="70%", height=NULL)

```

ANOVA de dos sentidos

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Posición	1	7160.0556	7160.0556	15.9981380	0.0018
Temperatura	2	945342.1111	472671.0556	1056.1170556	0.0000
Interacción	2	818.1111	409.0556	0.9139772	0.4271
Residuos	12	5370.6667	447.5556	NA	NA

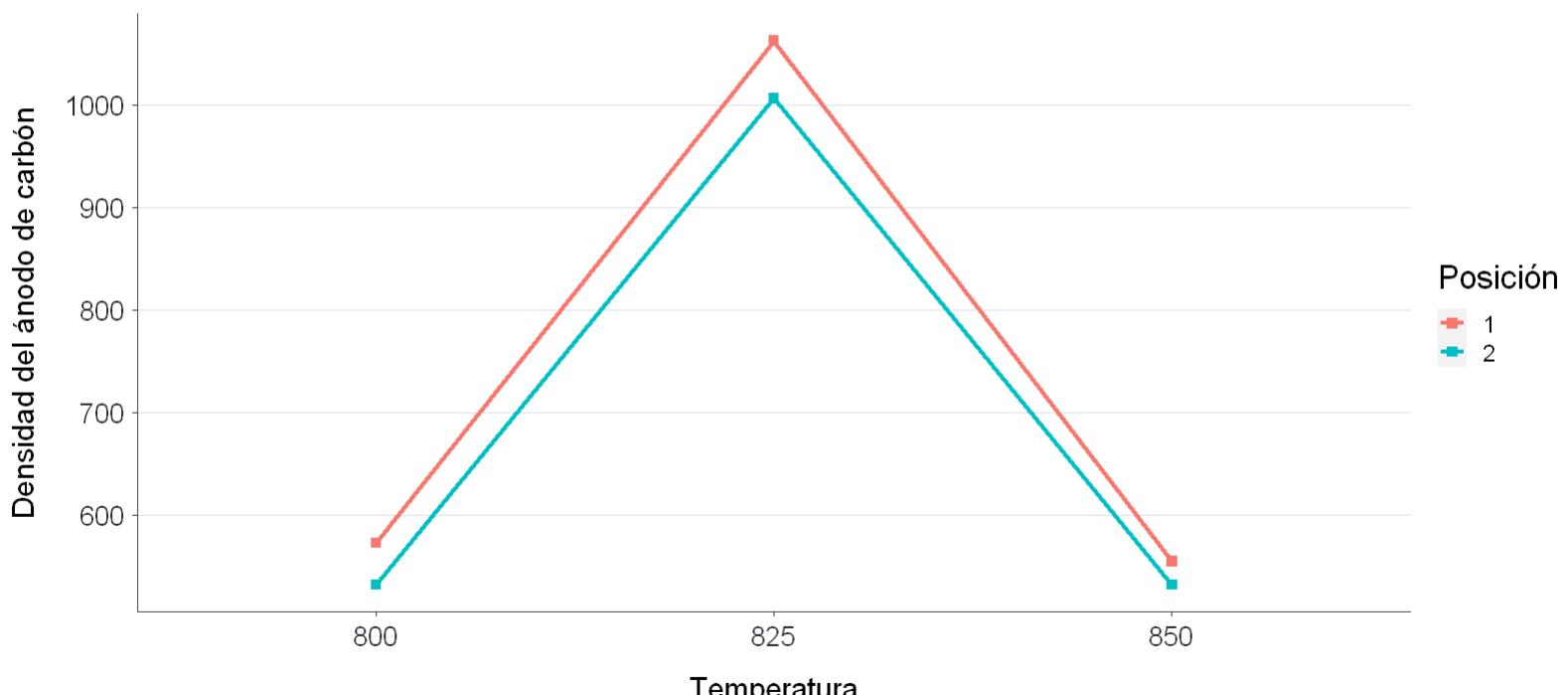
El p-valor para la interacción es 0.9731. Utilizando un nivel de significancia $\alpha = 0.05$, se acepta la hipótesis nula de que todas las interacciones son iguales a 0 y se concluye que el modelo es aditivo.

```

data.frame(data_9) %>%
  group_by(position, temperature) %>%
  summarise(cell_mean=mean(anode_density)) %>%
  arrange(match(temperature, c("low", "mid", "high"))) %>%
  mutate(temperature=factor(temperature, levels=temperature)) %>%
  ggplot(aes(temperature, cell_mean)) +
  geom_line(size = 1.2, aes(group = position, color = position)) +
  geom_point(size = 2.6, aes(color = position), shape = 15) +
  labs(
    title = "Interacción entre nivel de temperatura y posición",
    subtitle = "\n",
    x = "Temperatura",
    y = "Densidad del ánodo de carbón",
    color = "Posición") +
  theme(
    #text = element_text(family = "Roboto"),
    plot.margin = unit(rep(1.2, 4), "cm"),
    plot.title = element_text(size = 22, color = "#22292F", face = "bold", margin = margin(b = 5)),
    plot.subtitle = element_text(size = 15, margin = margin(b = 35)),
    plot.caption = element_text(size = 16, margin = margin(t = 25), color = "#606F7B"),
    panel.background = element_blank(),
    axis.text = element_text(size = 16, color = "#22292F"),
    axis.text.x = element_text(margin = margin(t = 5)),
    axis.text.y = element_text(margin = margin(r = 5)),
    axis.line = element_line(color = "#3D4852"),
    axis.title = element_text(size = 18),
    axis.title.y = element_text(margin = margin(r = 15),
                                hjust = 0.5),
    axis.title.x = element_text(margin = margin(t = 15),
                                hjust = 0.5),
    panel.grid.major = element_line(color = "#DAE1E7"),
    panel.grid.major.x = element_blank(),
    legend.title = element_text(size=20), #change Legend title font size
    legend.text = element_text(size=14)) #change Legend text font size

```

Interacción entre nivel de temperatura y posición



El gráfico indica que la interacción entre los factores posición y temperatura es insignificante, lo cual reafirma la validez del modelo aditivo.

2. Para probar si la media del resultado es igual para todos los niveles del factor **Posición**, se prueba la hipótesis nula de que todos los efectos fila son iguales a 0:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_I = 0$$

El p-valor para el factor **Posición** es 0.6149. Utilizando un nivel de significancia $\alpha = 0.05$, no es posible rechazar la hipótesis nula y por lo tanto se deduce que la posición de la caldera no afecta a la densidad del ánodo.

3. Para probar si la media del resultado es igual para todos los niveles del factor **Temperatura**, se prueba la hipótesis nula de que todos los efectos columna son iguales a 0:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_J = 0$$

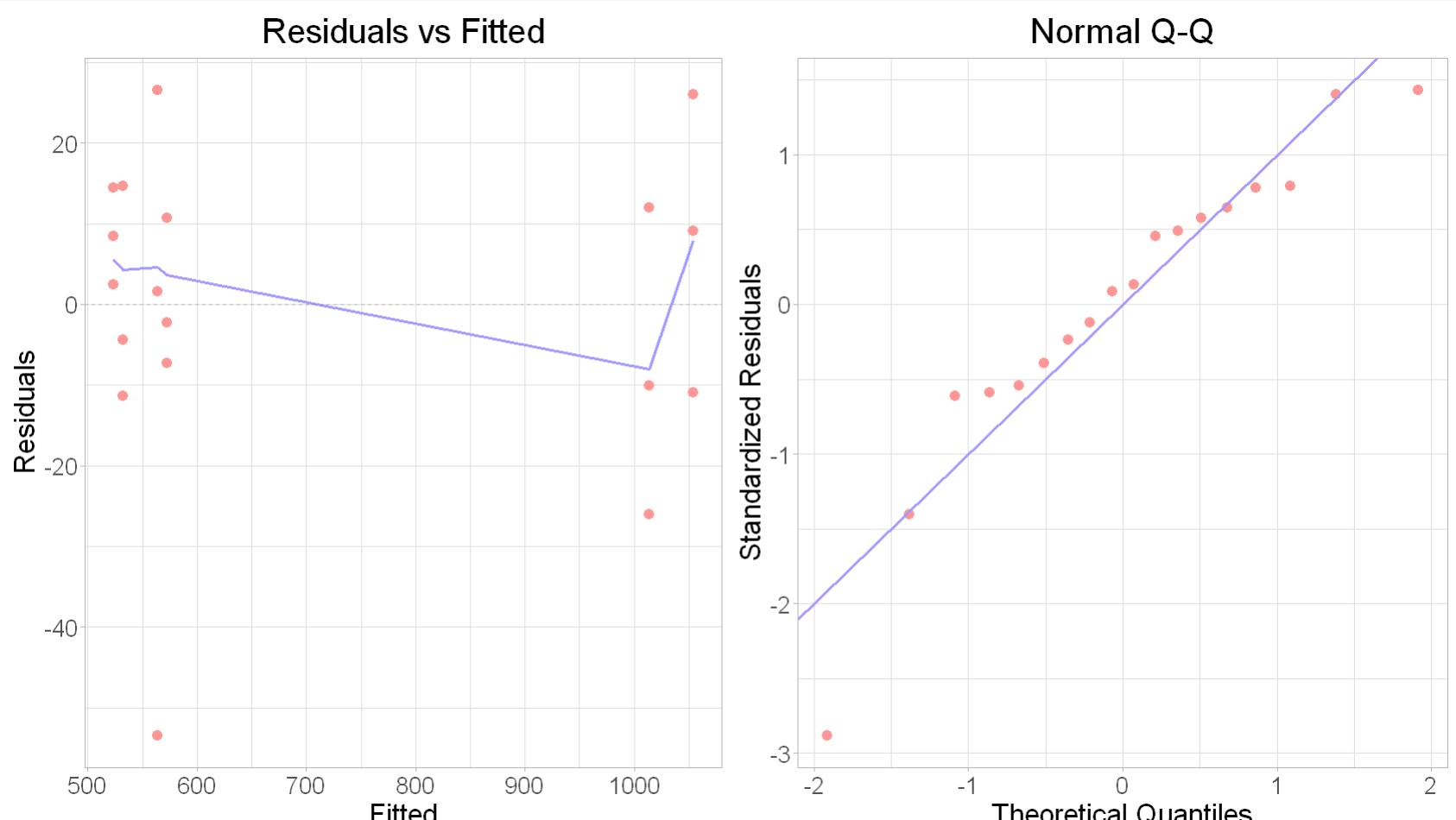
El p-valor para el factor **Temperatura** es 0.0015 < 0.05. Por lo tanto, se rechaza la hipótesis nula y se concluye que la temperatura de encendido afecta a la densidad del ánodo.

c) Analice los residuos de este experimento.

```
data_9.aov <- aov(anode_density ~ position + temperature, data_9)
smoothed <- data.frame(with(data_9.aov, lowess(x = data_9.aov$fitted, y = data_9.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_9.aov) +
  geom_point(aes(x=data_9.aov$fitted, y=data_9.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggttitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_9.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggttitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)
```



En el gráfico de *Residuos vs. Valores Ajustados* los valores parecen estar más dispersos en el lado derecho, poniendo en duda la supuesta de varianza constante. Los valores en la gráfica de *Probabilidad Normal* presentan cierta curvatura, aunque las desviaciones respecto a la recta normal no son notables y se podría asumir la normalidad de los residuos.

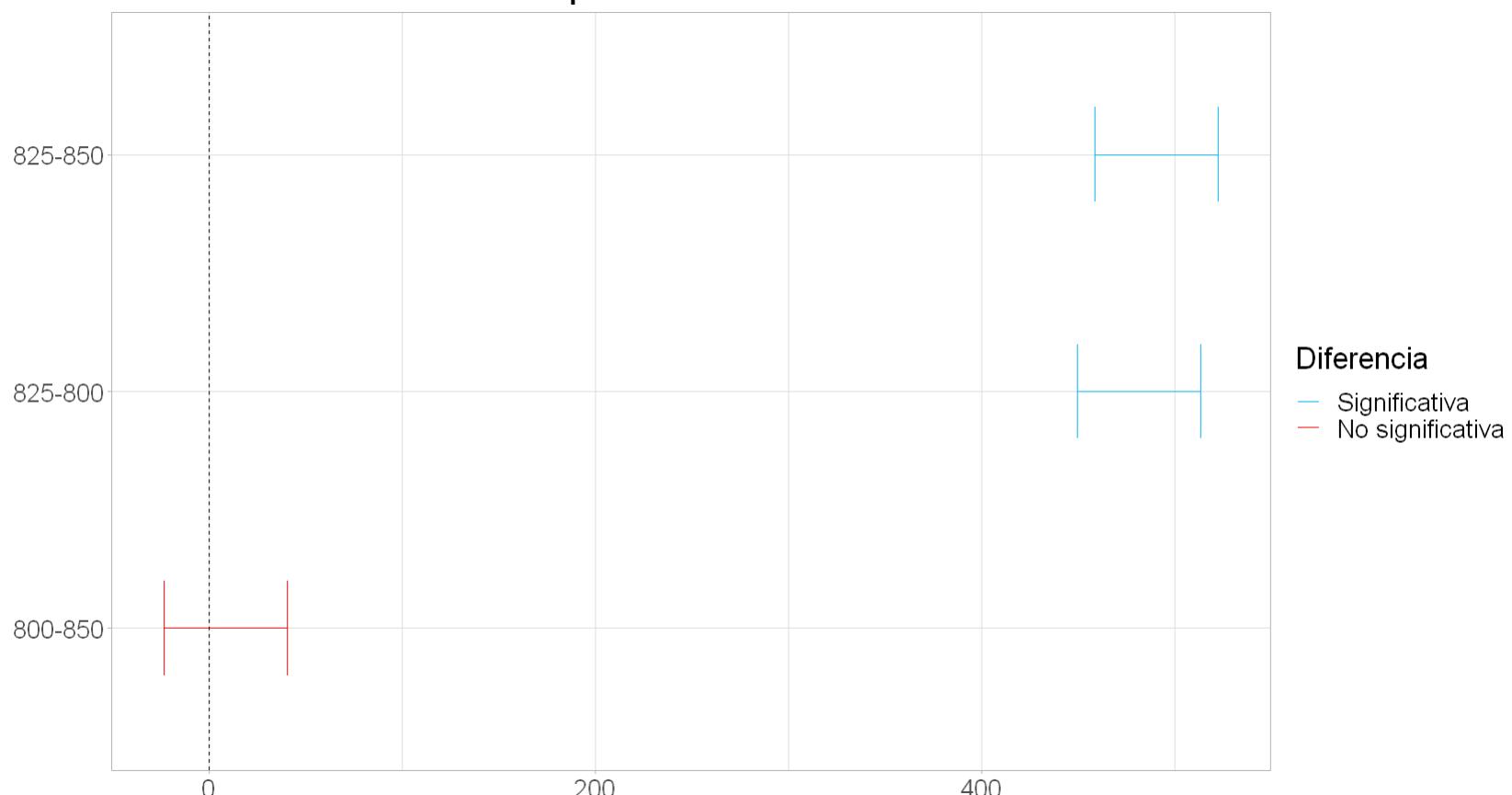
d) Investigue las diferencias entre la densidad promedio del ánodo para los tres niveles diferentes de temperatura.

Se puede aplicar el método de comparaciones múltiples de Tukey para determinar cuáles niveles de temperatura difieren entre sí.

```
data_9.aov <- aov(anode_density ~ position + temperature, data_9)
data_9.tukey <- as.data.frame(TukeyHSD(data_9.aov, ordered = TRUE, conf.level = 0.95)$temperature)
```

```
data_9.tukey$names <- c(rownames(data_9.tukey))
# Gráfico de los intervalos de confianza
options(repr.plot.width=14, repr.plot.height=8)
ggplot(data_9.tukey, aes(names, diff)) +
  labs(
    title="Intervalos de confianza de 95% para la diferencia de medias entre tratamientos",
    x="",
    y="",
    col="Diferencia") +
  geom_errorbar(aes(ymin=lwr, ymax=upr, col=ifelse(lwr*upr > 0,'1','2')), width = 0.4, alpha=1) +
  scale_color_manual(values=c('#05b5f5','#f50505'), labels=c('Significativa','No significativa'), breaks=c('1','2')) +
  geom_hline(yintercept=0, linetype="dashed", col="black") +
  theme_light() +
  coord_flip(expand = TRUE) +
  theme(text=element_text(size=20),
  plot.title = element_text(size=24, hjust = 0.5))
```

Intervalos de confianza de 95% para la diferencia de medias entre tratamientos



Los intervalos de confianza construidos indican que el efecto de la temperatura de encendido 825°C difiere del efecto de las otras dos temperaturas consideradas en el análisis.

Ejercicio 10

Se investigan los efectos del porcentaje de madera dura en la pulpa, la tersura y el tiempo de cocción sobre la resistencia del papel. La tabla siguiente presenta los datos provenientes de un experimento con tres factores:

Porcentaje de concentración de madera dura	Tiempo de cocción 1.5 horas			Tiempo de cocción 2.0 horas		
	Tersura			Tersura		
	350	500	650	350	500	650

Porcentaje de concentración de madera dura	Tiempo de cocción 1.5 horas			Tiempo de cocción 2.0 horas		
	Tersura			Tersura		
	350	500	650	350	500	650
10	96.6	97.7	99.4	98.4	99.6	100.6
	96.6	96.6	99.8	98.6	100.4	100.9
15	98.5	96.0	98.4	97.5	98.7	99.6
	97.2	96.9	97.6	98.1	96.0	99.0
20	97.5	95.6	97.4	97.6	97.0	98.5
	96.6	96.2	98.1	98.4	97.8	99.8

```
data_10 <- read.csv("./TP4_tables/data10.csv") # Leo los datos desde archivo .csv
data_10$hardwood_percentage <- factor(data_10$hardwood_percentage) # porcentaje de madera dura
data_10$cooking_time <- factor(data_10$cooking_time) # tiempo de cocción
data_10$smoothness <- factor(data_10$smoothness) # tersura
```

En este caso se trata de un experimento de tres factores:

- **Porcentaje de concentración de madera dura:** 3 niveles (10%, 15% y 20%)
- **Tiempo de cocción:** 2 niveles (1.5 horas y 2.0 horas)
- **Tersura:** 3 niveles (350, 500 y 650)

Cada uno de los 18 tratamientos tiene 2 réplicas.

a) Analice los datos usando análisis de varianza y suponiendo que los tres factores son fijos. Utilice $\alpha = 0.05$.

```
data_10.aov <- summary(aov(resistance ~ hardwood_percentage * cooking_time * smoothness, data_10))[[1]]
```

```
display_markdown('#### **ANOVA de tres sentidos**')
display_markdown('\n')
data_10.aov <- cbind(c('Porcentaje de madera dura', 'Tiempo de cocción', 'Tersura', 'Porcentaje de madera dura : Tiempo de cocción', 'Porcentaje de madera dura : Tersura', 'Tiempo de cocción : Tersura ', 'Porcentaje de madera dura : Tiempo de cocción : Tersura', 'Residuos'), data_10.aov)
colnames(data_10.aov)[1] <- 'Source'
rownames(data_10.aov) <- c()
data_10.aov["Pr(>F)"] <- round(data_10.aov["Pr(>F)", 4]
table <- formattable(data_10.aov, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(`Source` = formatter("span", style = ~ style('text-align'='left'))))
as.htmlwidget(table, width="80%", height=NULL)
```

ANOVA de tres sentidos

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Porcentaje de madera dura	2	10.055000	5.0275000	10.0998884	0.0011
Tiempo de cocción	1	15.734444	15.734444	31.6093750	0.0000
Tersura	2	20.561667	10.2808333	20.6534598	0.0000
Porcentaje de madera dura : Tiempo de cocción	2	2.350556	1.1752778	2.3610491	0.1229
Porcentaje de madera dura : Tersura	4	6.053333	1.5133333	3.0401786	0.0445
Tiempo de cocción : Tersura	2	1.007222	0.5036111	1.0117187	0.3834
Porcentaje de madera dura : Tiempo de cocción : Tersura	4	1.577778	0.3944444	0.7924107	0.5453
Residuos	18	8.960000	0.4977778	NA	NA

El p-valor es mayor a 0.05 para todas las interacciones. Por lo tanto, no es posible rechazar la hipótesis nula de que todas las interacciones son iguales a 0 y se asume que el modelo es aditivo.

```
display_markdown('#### **Interacciones**')
```

```
interacion_1 <- data.frame(data_10) %>%
```

```

group_by(cooking_time, hardwood_percentage) %>%
summarise(cell_mean=mean(resistance)) %>%
ggplot(aes(hardwood_percentage, cell_mean)) +
geom_line(size = 1.2, aes(group = cooking_time, color = cooking_time)) +
geom_point(size = 2.6, aes(color = cooking_time), shape = 15) +
labs(
  title = "Porcentaje de madera dura y Tiempo de cocción",
  #subtitle = "\n",
  x = "Porcentaje de madera dura [%]",
  y = "Resistencia del papel",
  color = "Tiempo de cocción") +
theme(
  plot.margin = unit(rep(1.2, 4), "cm"),
  plot.title = element_text(size = 16, color = "#22292F", face = "bold", margin = margin(b = 5)),
  plot.subtitle = element_text(size = 10, margin = margin(b = 35)),
  plot.caption = element_text(size = 14, margin = margin(t = 25), color = "#606F7B"),
  panel.background = element_blank(),
  axis.text = element_text(size = 14, color = "#22292F"),
  axis.text.x = element_text(margin = margin(t = 5)),
  axis.text.y = element_text(margin = margin(r = 5)),
  axis.line = element_line(color = "#3D4852"),
  axis.title = element_text(size = 14),
  axis.title.y = element_text(margin = margin(r = 15),
                             hjust = 0.5),
  axis.title.x = element_text(margin = margin(t = 15),
                             hjust = 0.5),
  panel.grid.major = element_line(color = "#DAE1E7"),
  panel.grid.major.x = element_blank(),
  legend.title = element_text(size=14), #change Legend title font size
  legend.text = element_text(size=12)) #change Legend text font size

interacion_2 <- data.frame(data_10) %>%
group_by(smoothness, hardwood_percentage) %>%
summarise(cell_mean=mean(resistance)) %>%
ggplot(aes(hardwood_percentage, cell_mean)) +
geom_line(size = 1.2, aes(group = smoothness, color = smoothness)) +
geom_point(size = 2.6, aes(color = smoothness), shape = 15) +
labs(
  title = "Porcentaje de madera dura y Tersura",
  #subtitle = "\n",
  x = "Porcentaje de madera dura [%]",
  y = "Resistencia del papel",
  color = "Tersura") +
theme(
  plot.margin = unit(rep(1.2, 4), "cm"),
  plot.title = element_text(size = 16, color = "#22292F", face = "bold", margin = margin(b = 5)),
  plot.subtitle = element_text(size = 10, margin = margin(b = 35)),
  plot.caption = element_text(size = 14, margin = margin(t = 25), color = "#606F7B"),
  panel.background = element_blank(),
  axis.text = element_text(size = 14, color = "#22292F"),
  axis.text.x = element_text(margin = margin(t = 5)),
  axis.text.y = element_text(margin = margin(r = 5)),
  axis.line = element_line(color = "#3D4852"),
  axis.title = element_text(size = 14),
  axis.title.y = element_text(margin = margin(r = 15),
                             hjust = 0.5),
  axis.title.x = element_text(margin = margin(t = 15),
                             hjust = 0.5),
  panel.grid.major = element_line(color = "#DAE1E7"),
  panel.grid.major.x = element_blank(),
  legend.title = element_text(size=14), #change Legend title font size
  legend.text = element_text(size=12)) #change Legend text font size

interacion_3 <- data.frame(data_10) %>%
group_by(cooking_time, smoothness) %>%
summarise(cell_mean=mean(resistance)) %>%
ggplot(aes(smoothness, cell_mean)) +
geom_line(size = 1.2, aes(group = cooking_time, color = cooking_time)) +
geom_point(size = 2.6, aes(color = cooking_time), shape = 15) +
labs(
  title = "Tiempo de cocción y Tersura",
  #subtitle = "\n",
  x = "Tersura",
  y = "Resistencia del papel",
  color = "Tiempo de cocción") +
theme(
  plot.margin = unit(rep(1.2, 4), "cm"),
  plot.title = element_text(size = 16, color = "#22292F", face = "bold", margin = margin(b = 5)),
  plot.subtitle = element_text(size = 10, margin = margin(b = 35)),
  plot.caption = element_text(size = 14, margin = margin(t = 25), color = "#606F7B"),
  panel.background = element_blank(),
  axis.text = element_text(size = 14, color = "#22292F"),
  axis.text.x = element_text(margin = margin(t = 5)),

```

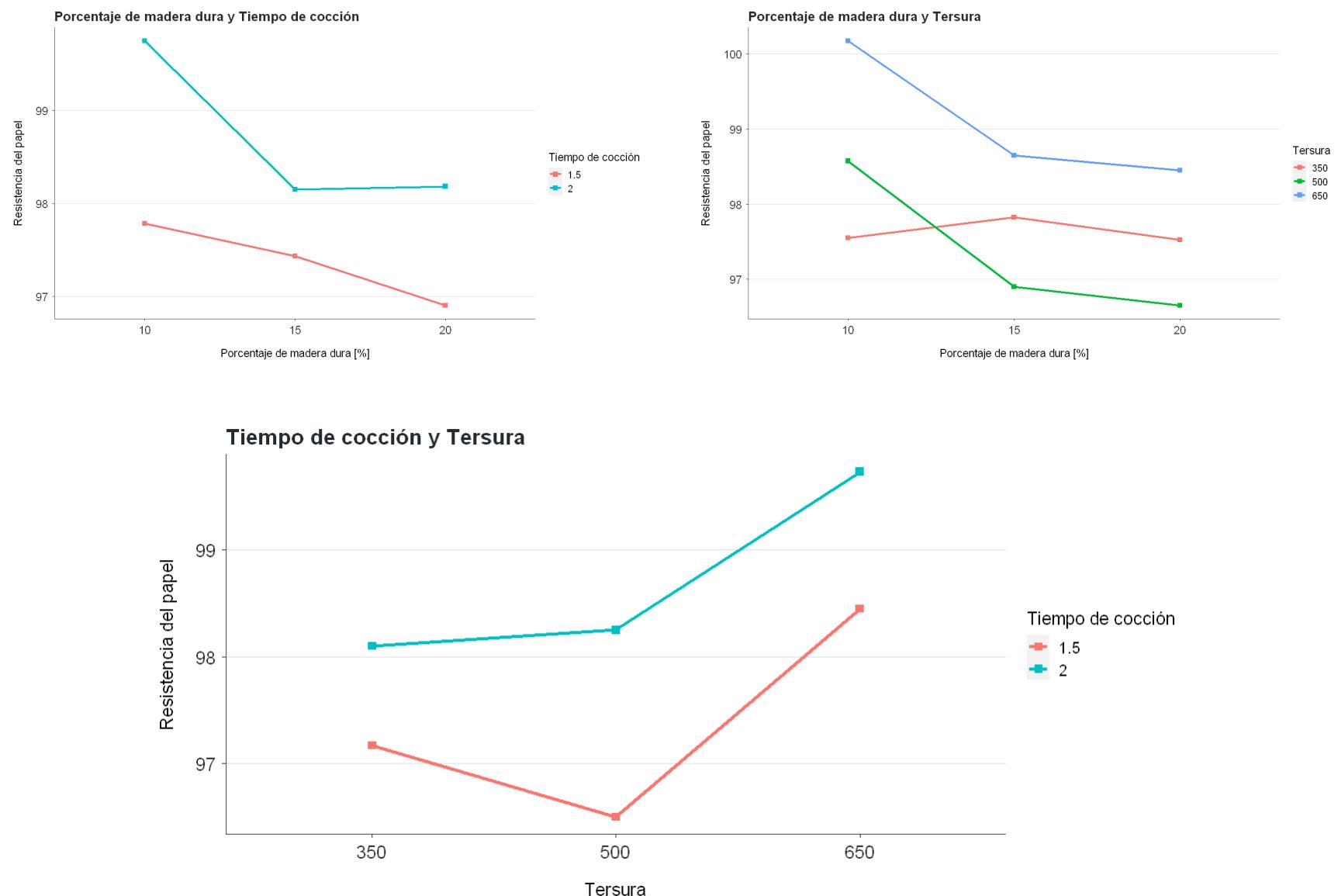
```

axis.text.y = element_text(margin = margin(r = 5)),
axis.line = element_line(color = "#3D4852"),
axis.title = element_text(size = 14),
axis.title.y = element_text(margin = margin(r = 15),
                           hjust = 0.5),
axis.title.x = element_text(margin = margin(t = 15),
                           hjust = 0.5),
panel.grid.major = element_line(color = "#DAE1E7"),
panel.grid.major.x = element_blank(),
legend.title = element_text(size=14), #change Legend title font size
legend.text = element_text(size=12)) #change Legend text font size

options(repr.plot.width=24, repr.plot.height=7)
plot_grid(interacion_1, interacion_2, ncol = 2)
options(repr.plot.width=12, repr.plot.height=6)
interacion_3

```

Interacciones



```

I[1] <- data_10 %>% distinct(hardwood_percentage) %>% pull() %>% length() # cantidad de niveles de porcentaje de madera dura
I[2] <- data_10 %>% distinct(cooking_time) %>% pull() %>% length() # cantidad de niveles de tiempo de coccion
I[3] <- data_10 %>% distinct(smoothness) %>% pull() %>% length() # cantidad de niveles de tersura

K <- nrow(data_10) / (prod(I)) # cantidad de réplicas por cada tratamiento

df1 <- I - 1
df2 <- prod(I) * (K - 1)

alpha <- 0.05
f_alpha <- qf(alpha, df1=df1, df2=df2, lower=FALSE)

```

```

data_10.f_values <- summary(aov(resistance ~ hardwood_percentage * cooking_time * smoothness, data_10))[[1]][1:3,]
colnames(data_10.f_values)[1] <- 'df1'
data_10.f_values['df2'] <- df2
data_10.f_values['F (α=0.05)'] <- f_alpha
data_10.f_values[' '] <- data_10.f_values['F value'] > data_10.f_values['F (α=0.05)'] # TRUE si se rechaza H0
data_10.f_values
data_10.f_values <- data_10.f_values[,c(1,6,4,8,7)]

```

A anova: 3 × 8

	df1	Sum Sq	Mean Sq	F value	Pr(>F)	df2	F (a=0.05)	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<lgI[,1]>
hardwood_percentage	2	10.05500	5.02750	10.09989	1.145263e-03	18	3.554557	TRUE
cooking_time	1	15.73444	15.73444	31.60938	2.465291e-05	18	4.413873	TRUE
smoothness	2	20.56167	10.28083	20.65346	2.185266e-05	18	3.554557	TRUE

```
display_markdown('#### **Estadísticos F**')
display_markdown('\n')
data_10.f_values <- cbind(c('Porcentaje de madera dura', 'Tiempo de cocción', 'tersura'), data_10.f_values)
colnames(data_10.f_values)[1] <- 'Factor'
rownames(data_10.f_values) <- c()
table <- formattable(data_10.f_values, align=c('l', 'c', 'c', 'c', 'c', 'c'), list(` ` = formatter("span",
x ~ ifelse(x, ">", "<"),
style = x ~ style(color = ifelse(x, "green", "red")))))
#table <- format_table(table, list(area(1:2) ~ color_tile("transparent", "Lightgray")))
as.htmlwidget(table, width="60%", height=NULL)
```

Estadísticos F

Factor	df1	df2	F value	F (a=0.05)
Porcentaje de madera dura	2	18	10.09989	> 3.554557
Tiempo de cocción	1	18	31.60938	> 4.413873
tersura	2	18	20.65346	> 3.554557

Para todos los factores considerados, el valor del estadístico de prueba es mayor al valor correspondiente a un nivel de significancia $\alpha = 0.05$. Por lo tanto, se concluye que los tres factores afectan a la resistencia del papel.

b) Encuentre los valores P de los cocientes F del inciso a).

```
f_values <- data_10.f_values[, 'F value']
p_values <- pf(f_values, df1=df1, df2=df2, lower=FALSE)
data_10.p_values <- data_10.f_values %>% select('df1', 'df2', 'F value') %>% mutate('p-valor'=round(p_values, 4))

display_markdown('#### **Valores P**')
display_markdown('\n')
data_10.p_values <- cbind(c('Porcentaje de madera dura', 'Tiempo de cocción', 'tersura'), data_10.p_values)
colnames(data_10.p_values)[1] <- 'Factor'
rownames(data_10.p_values) <- c()
table <- formattable(data_10.p_values, align=c('l', 'c', 'c', 'c', 'c', 'c'))
#table <- format_table(table, list(area(1:2) ~ color_tile("transparent", "Lightgray")))
as.htmlwidget(table, width="50%", height=NULL)
```

Valores P

Factor	df1	df2	F value	p-valor
Porcentaje de madera dura	2	18	10.09989	0.0011
Tiempo de cocción	1	18	31.60938	0.0000
tersura	2	18	20.65346	0.0000

En los tres casos, el *p-valor* es menor a $\alpha = 0.05$.

c) Analice los residuos de este experimento.

```
data_10.aov <- aov(resistance ~ hardwood_percentage * cooking_time * smoothness, data_10)
smoothed <- data.frame(with(data_10.aov, lowess(x = data_10.aov$fitted, y = data_10.aov$residuals)))
# Gráficos
options(repr.plot.width=14, repr.plot.height=8)
res_vs_fit <- ggplot(data_10.aov) +
```

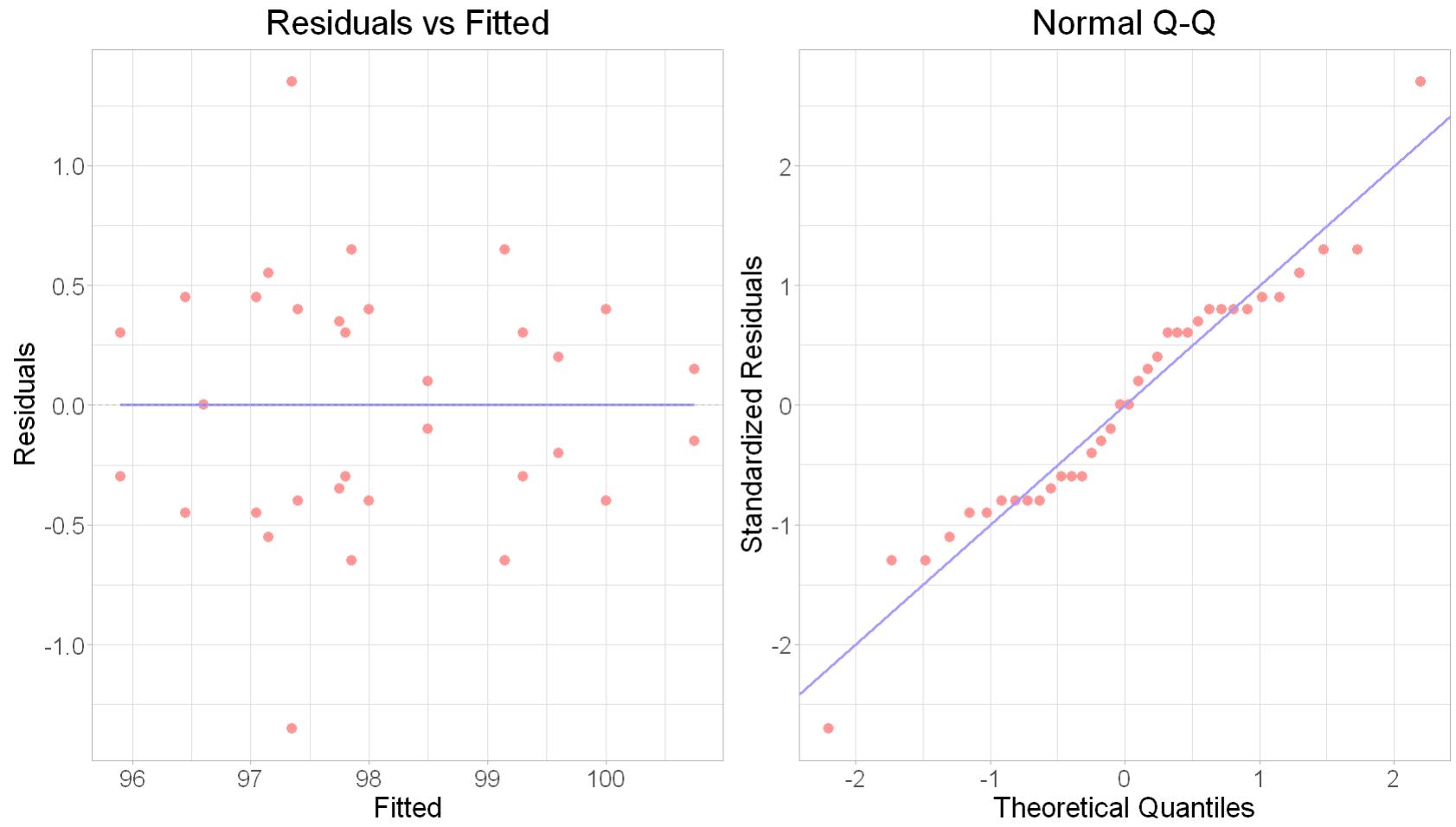
```

geom_point(aes(x=data_10.aov$fitted, y=data_10.aov$residuals), color= '#ff9696', size=3) +
  geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff", size=1) +
  geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
  ggtitle("Residuals vs Fitted") +
  xlab("Fitted") +
  ylab("Residuals") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

qq_plot <- ggplot(data_10.aov) +
  stat_qq(aes(sample = .stdresid), color= '#ff9696', size=3) +
  geom_abline(col="#a399ff", size=1) +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  ggtitle("Normal Q-Q") +
  theme_light() +
  theme(text=element_text(size=20),
        plot.title = element_text(size=24, hjust = 0.5))

plot_grid(res_vs_fit, qq_plot, ncol = 2)

```



Tanto la gráfica de *Residuos vs. Valores Ajustados* como la gráfica de *Probabilidad Normal* no sugieren violaciones significativas de los supuestos de igualdad de varianza y normalidad. Entonces, es razonable asumir la validez del modelo.