ESTADÍSTICA PARA INGENIERÍA Y CIENCIAS

PRÁCTICA 3: Regresión lineal múltiple

Ivan Svetlich

```
#Librerias
library(IRdisplay)
library(ggplot2)
library(latex2exp)
library(data.table)
library(dplyr)
library(formattable)
```

Ejercicio 1

Se realiza un estudio para investigar la resistencia al esfuerzo cortante de un suelo (y) y la relación que tiene ésta con la profundidad en pies (x_1) y con el contenido de humedad (x_2). Para ello se recopilan 10 observaciones, a partir de las cuales se obtiene lo siguiente:

n	$\sum_{i=1}^{n} x_{i1}$	$\sum_{i=1}^{n} x_{i2}$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_{i1}^2$	$\sum_{i=1}^{n} x_{i2}^2$	$\sum_{i=1}^{n} x_{i1}x_{i2}$	$\sum_{i=1}^{n} x_{i1}y_i$	$\sum_{i=1}^{n} x_{i2}y_i$
10	223	553	1916	371595.6	5200	31729	12352	43550	104736.8

Valores observados

a) Establezca las ecuaciones normales de mínimos cuadrados para el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Las ecuaciones normales son:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i$$

b) Estime los parámetros del modelo del inciso $a)$

```r
n <- 10
sum_x1 <- 223; sum_x2 <- 553
sum_y <- 1916; sum_y_sq <- 371595.6
sum_x1_sq <- 5200; sum_x2_sq <- 31729; sum_x1_x2 <- 12352
sum_x1_y <- 43550; sum_x2_y <- 104736.8

A <- as.matrix(rbind(
    c(n, sum_x1, sum_x1_x2),
    c(sum_x1, sum_x1_sq, sum_x1_x2),
    c(sum_x2, sum_x1_x2, sum_x2_sq)))
B <- as.matrix(c(sum_y, sum_x1_y, sum_x2_y))

beta <- solve(A, B)
```

```r
display_markdown(sprintf("$$\\begin{bmatrix} \\hat{\\beta_0} \\\\ \\hat{\\beta_1} \\\\ \\hat{\\beta_2}\\end{bmatrix} = \\begin{bmatrix} %f \\\\ %f \\\\ %f\\end{bmatrix}$$", beta[1], beta[2], beta[3]))
```

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 170.816315 \\ 3.724502 \\ -1.126089 \end{bmatrix}$$

c) ¿Cuál es la resistencia al esfuerzo cortante predicha cuando $x_1 = 18$ pies y $x_2 = 43\%$?

```r
x <- as.matrix(c(1, 18, 43))
y <- sum(beta * x)
```

```r
display_markdown(sprintf("$\\hat{y} = \\hat{\\beta_0} + \\hat{\\beta_1} x_1 + \\hat{\\beta_2} x_2 = %.4f$", y))
```

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 189.4355$

Ejercicio 2

Se piensa que la potencia eléctrica consumida al mes por una planta química está relacionada con la temperatura ambiente promedio (x_1), el número de días del mes (x_2), la pureza promedio del producto (x_3) y las toneladas de producto producidas (x_4). Los datos correspondientes al año pasado son los siguientes:

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

```r
df <- data.frame(read.table("./TP3_tables/data2.txt", header = TRUE))
names(df) <- c("y", "x1", "x2", "x3", "x4")
```

```
display_markdown("#### **Resumen de los datos**")
df_summary <- as.data.frame(apply(df, 2, summary))
df_summary$values <- rownames(df_summary)
df_summary <- df_summary[,c(6,1,2,3,4,5)]
colnames(df_summary)[1] <- " "
rownames(df_summary) <- c()
table <- formattable(df_summary, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL)
```

Resumen de los datos

	y	x1	x2	x3	x4
Min.	236.00	25.00000	21.00000	86.00000	88.00000
1st Qu.	265.50	43.25000	24.00000	87.75000	95.75000
Median	282.00	60.00000	25.00000	89.50000	98.50000
Mean	278.75	57.08333	24.33333	89.33333	99.33333
3rd Qu.	297.00	72.75000	25.00000	91.00000	101.25000
Max.	316.00	84.00000	26.00000	94.00000	110.00000

a) Ajuste un modelo de regresión lineal múltiple a los datos contenidos en la tabla anterior.

El modelo propuesto es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

```
model <- lm(y ~ x1 + x2 + x3 + x4, data=df)
```

```
coef <- as.data.frame(x=model$coefficients)
colnames(coef) <- c('')
rownames(coef) <- c('β0', 'β1', 'β2', 'β3', 'β4')
coef <- as.data.frame(t(coef))
table <- formattable(coef, align = c("c", "c", "c", "c", "c"))
display_markdown("#### **Coeficientes**")
as.htmlwidget(table, width="50%", height=NULL)
```

Coeficientes

β0	β1	β2	β3	β4
-102.7132	0.6053705	8.923644	1.437457	0.01360931

b) Prediga el consumo de potencia para un mes en el que $x_1 = 75^{o}F$, $x_2 = 24$ días, $x_3 = 90\%$ y $x_4 = 98$ toneladas.

```
x_test <- data.frame(x1=75, x2=24, x3=90, x4=98)
y_test <- predict(model, x_test, interval = "none", type = "response")
```

```
display_markdown(paste("$\\hat{y} = \\hat{\\beta}_0 + \\hat{\\beta}_1 x_1 + \\hat{\\beta}_2 x_2 + \\hat
{\\beta}_3 x_3 + \\hat{\\beta}_4 x_4 = $", round(y_test, 4)))
```

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 = 287.5618$

c) Pruebe la significancia de la regresión utilizando $\alpha = 0.01$. ¿Cuál es el p-valor de la prueba?

Para determinar la significancia de la regresión se prueba la hipótesis nula:

$$H_{0_1} : \beta_1 = \beta_2 = 0$$

Esta hipótesis establece que ninguna de las variables independientes tiene alguna relación lineal con la variable dependiente. El estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE}$$

Si la hipótesis nula es rechazada, al menos una de las variables regresoras esta linealmente relacionada con la variable respuesta.

```
f_stat<- as.data.frame(t(summary(model)$fstatistic))
f_stat$p_value <- c(pf(f_stat[,1], f_stat[,2], f_stat[,3], lower.tail=FALSE))
f_stat <- as.data.frame(cbind('F-statistic', f_stat))
colnames(f_stat) <- c(' ', 'value', 'df1', 'df2', 'p-value')
table <- formattable(f_stat, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL, )
```

	value	df1	df2	p-value
F-statistic	5.106018	4	7	0.03030277

El p-valor de la regresión es $0.0303 > 0.01$, así que no es posible rechazar la hipótesis nula de que al menos uno de los parámetros β_j del modelo es 0.

d) Estime σ^2.

La estimación de la varianza del error para un modelo de regresión múltiple con p parámetros y n observaciones es:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p} = \frac{SSE}{n-p} = MSE$$

```
y_pred <- predict(model) # valores predichos
y_true <- df['y'] # valores verdaderos
SSE <- sum((y_true - y_pred)^2)
n <- nrow(df) # n=12 observaciones
p <- ncol(coef) # p=5 coeficientes de regresión
MSE <- SSE / (n - p)
```

```
display_markdown(paste("$\\hat{\\sigma^2} = \\text{MSE} =$", round(MSE, 4)))
```

$\hat{\sigma^2} = \text{MSE} = 242.7156$

```
# Otra forma
residuals <- anova(model)['Residuals',]
residuals$values <- rownames(residuals)
residuals <- residuals[,c(6,1,2,3)]
colnames(residuals)[1] <- " "
rownames(residuals) <- c()
table <- formattable(residuals, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL, )
```

	Df	Sum Sq	Mean Sq
Residuals	7	1699.009	242.7156

La columna **Mean Sq** es el MSE, y su valor coincide con el calculado previamente.

e) Utilice la prueba t para evaluar la contribución al modelo de cada variable de regresión. Si se emplea $\alpha = 0.01$, ¿qué conclusiones pueden obtenerse?

El estadístico t de Student se utiliza para probar la hipótesis nula de que el valor verdadero del coeficiente β_j es igual a 0. Este estadístico es igual al cociente del estimador del coeficiente y su desviación estándar:

$$T_0 = \frac{\hat{\beta}_j}{\text{se}\left(\hat{\beta}_j\right)}$$

```
model_coef <- format(round(model_summary$coefficients, 5), nsmall=5)
model_coef <- cbind(coef=c('β0', 'β1', 'β2', 'β3', 'β4'), model_coef)
colnames(model_coef)[1] <- "Coefficient"
rownames(model_coef) <- c()
table <- formattable(as.data.frame(model_coef, 5), align='c', list('Coefficient' = formatter("span",style
= ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL)
```

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β0	-102.71324	207.85885	-0.49415	0.63633
β1	0.60537	0.36890	1.64103	0.14480
β2	8.92364	5.30052	1.68354	0.13615
β3	1.43746	2.39162	0.60104	0.56676
β4	0.01361	0.73382	0.01855	0.98572

El p-valor del test $H_0 : \beta_j = 0$ es mayor que $\alpha = 0.01$ para todos los coeficientes del modelo. Este resultado indica que la contribución de cada variable regresora, dado que el resto de los regresores también se incluyen en el modelo, es poco significativa.

f) Encuentre un intervalo de confianza del 95% para los coeficientes de regresión β_1, β_2, β_3 y β_4.

Dado un modelo con p parámetros y n observaciones, el intervalo de confianza de $100(1-\alpha)$ para un coeficiente de regresión β_j individual está dado por:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p} se(\hat{\beta}_j)$$

donde $\hat{\beta}_j$ es la estimación del coeficiente de regresión, $t_{\alpha/2,n-p}$ es el valor de la distribución t de Student con $(\alpha/2, n-p)$ grados de libertad y $se(\hat{\beta}_j)$ es la desviación standard de la estimación del coeficiente.

```
alpha <- 0.05
beta <- as.data.frame(model_summary$coefficients)['Estimate']
std_dev <- as.data.frame(model_summary$coefficients)['Std. Error']
t <- qt(alpha/2, 8, lower.tail=FALSE) # 7 grados de libertad
conf_int <- as.data.frame(c(round(beta - t * std_dev, 4), round(beta + t * std_dev, 4)), col.names=c('lwr
', 'upr'), row.names=c('$\\beta_0$', '$\\beta_1$', '$\\beta_2$', '$\\beta_3$', '$\\beta_4$'))
conf_int <- conf_int[-1,]
display_markdown("Intervalos de confianza del 95% para los coeficientes de regresión:")
for (row in 1:nrow(conf_int))
{
    display_markdown(glue::glue("{rownames(conf_int)[row]}: ({conf_int$lwr[row]}, {conf_int$upr[row]})"))
}
```

Intervalos de confianza del 95% para los coeficientes de regresión:

β_1: (-0.2453, 1.456)

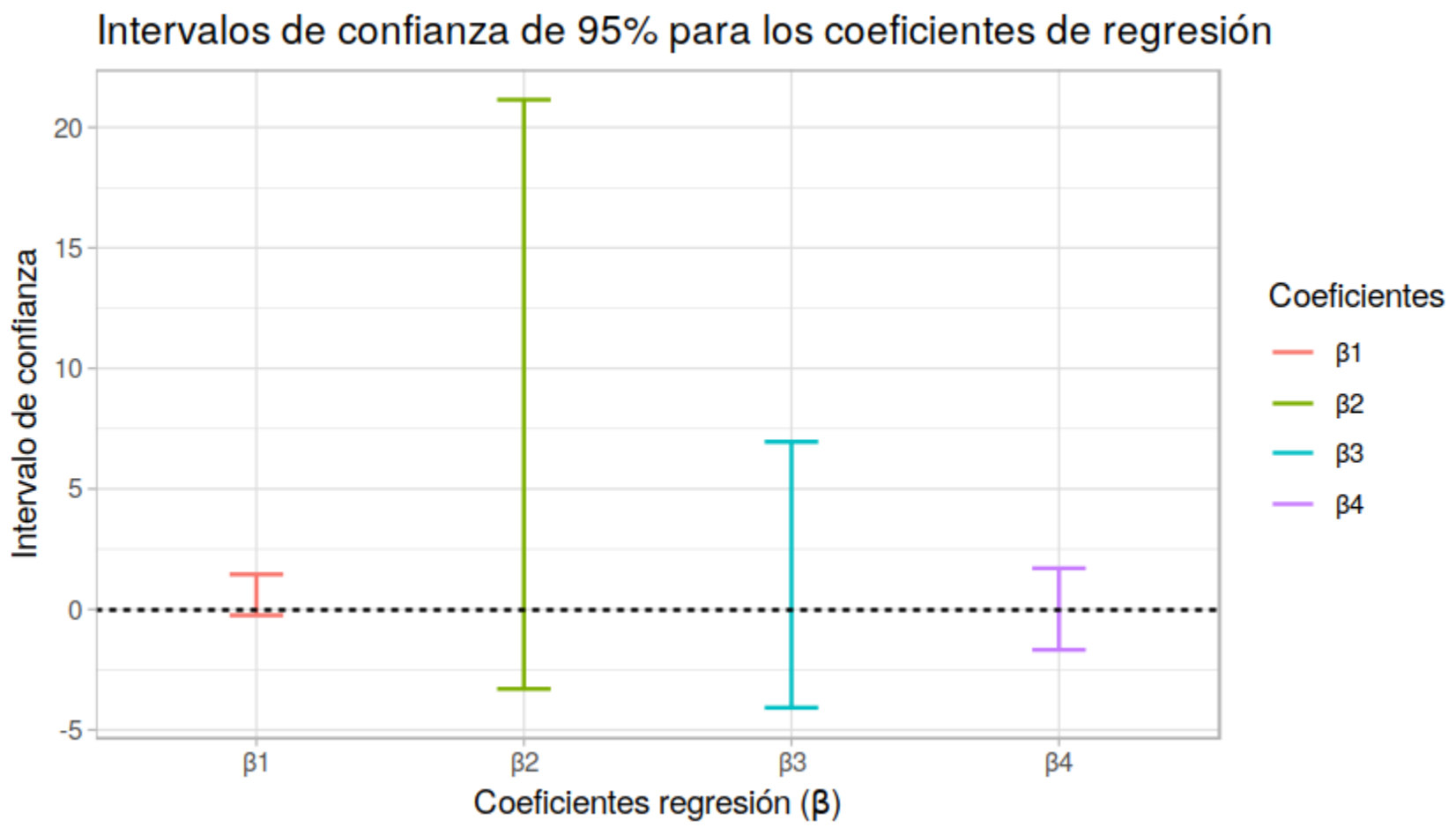
β_2: (-3.2994, 21.1467)

β_3: (-4.0776, 6.9525)

β_4: (-1.6786, 1.7058)

```
names <- c('β1', 'β2', 'β3', 'β4')
conf_int$avg <- apply(conf_int, 1, mean)

# Gráfico de los intervalos de confianza
options(repr.plot.width=7, repr.plot.height=4)
ggplot(conf_int, aes(names, avg, colour=names)) +
    ggtitle("Intervalos de confianza de 95% para los coeficientes de regresión") +
    geom_errorbar(aes(ymin=lwr, ymax=upr), width = 0.2) +
    labs(x=TeX('Coeficientes regresión ($\\beta$ )'), y='Intervalo de confianza', color='Coeficientes') +
    geom_hline(yintercept=0, linetype="dashed", col="black") +
    theme_light()
```



Todos los intervalos incluyen al cero, y por lo tanto no es posible rechazar la hipótesis $H_0 : \beta_j = 0$.

g) Encuentre un intervalo de confianza del 95% para la media de Y cuando $x_1 = 75$, $x_2 = 24$, $x_3 = 90$ y $x_4 = 98$.

El intervalo de confianza de $100(1 - \alpha)$ para la respuesta media en el punto $(x_1 = x_{01}, x_2 = x_{02}, x_3 = x_{03}, x_4 = x_{04})$ está dado por:

$$\hat{\mu} \pm t_{\alpha/2,n-p} se(\hat{\mu})$$

donde:

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \hat{\beta}_3 x_{03} + \hat{\beta}_4 x_{04}$$

```
x_test <- data.frame(x1=75, x2=24, x3=90, x4=98)
y_conf <- as.data.frame(predict(model, x_test, interval="confidence", level=0.95, type="response"))
colnames(y_conf) <- c('ajuste', 'cota inferiror', 'cota superior')
y_conf <- select(y_conf, 2, 1, 3)
display_markdown('#### **Intervalo de confianza:**')
as.htmlwidget(formattable(as.data.frame(y_conf), align="c"), width="50%")
display_markdown(sprintf("Intervalo de confianza del 95%% para la media de la potencia mensual consumida:
$\\left(%.4f, %.4f\\right)$", y_conf[1], y_conf[3]))
```

Intervalo de confianza:

cota inferiror	ajuste	cota superior
263.7879	287.5618	311.3357

Intervalo de confianza del 95% para la media de la potencia mensual consumida: $(263.7879, 311.3357)$

h) Encuentre un intervalo de predicción del 95% para la media de Y cuando $x_1 = 75$, $x_2 = 24$, $x_3 = 90$ y $x_4 = 98$.

El intervalo de predicción de $100(1-\alpha)$ para una observación futura en el punto $(x_1 = x_{01}, x_2 = x_{02}, x_3 = x_{03}, x_4 = x_{04})$ está dado por:

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 + \left[se\left(\hat{\mu}\right)\right]^2}$$

donde:

$$\hat{y}_0 = \hat{\mu} = \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \hat{\beta}_3 x_{03} + \hat{\beta}_4 x_{04}$$

```
y_pred <- as.data.frame(predict(model, x_test, interval="prediction", level=0.95, type="response"))
colnames(y_pred) <- c('ajuste', 'cota inferiror', 'cota superior')
y_pred <- select(y_pred, 2, 1, 3)
display_markdown('#### **Intervalo de predicción:**')
as.htmlwidget(formattable(as.data.frame(y_pred), align="c"), width="50%")
display_markdown(sprintf("Intervalo de predicción del 95%% para la media de la potencia mensual consumid
a: $\\left(%.4f, %.4f\\right)$", y_pred[1], y_pred[3]))
```

Intervalo de predicción:

cota inferiror	ajuste	cota superior
243.7175	287.5618	331.4062

Intervalo de predicción del 95% para la media de la potencia mensual consumida: $(243.7175, 331.4062)$

i) Para este modelo calcule R^2. Interprete esta cantidad.

```
r_squared <- summary(model)$r.squared
display_markdown(sprintf('$R^2 = %f$', r_squared))
```

$R^2 = 0.744750$

El coeficiente de determinación R^2 indica qué proporción de la dispersión observada en la variable dependiente (y) es predecible a partir de las variables independientes $\left(x_j\right)$. Un valor $R^2 = 0.744750$ es un resultado aceptable.

Respondo en conjunto:

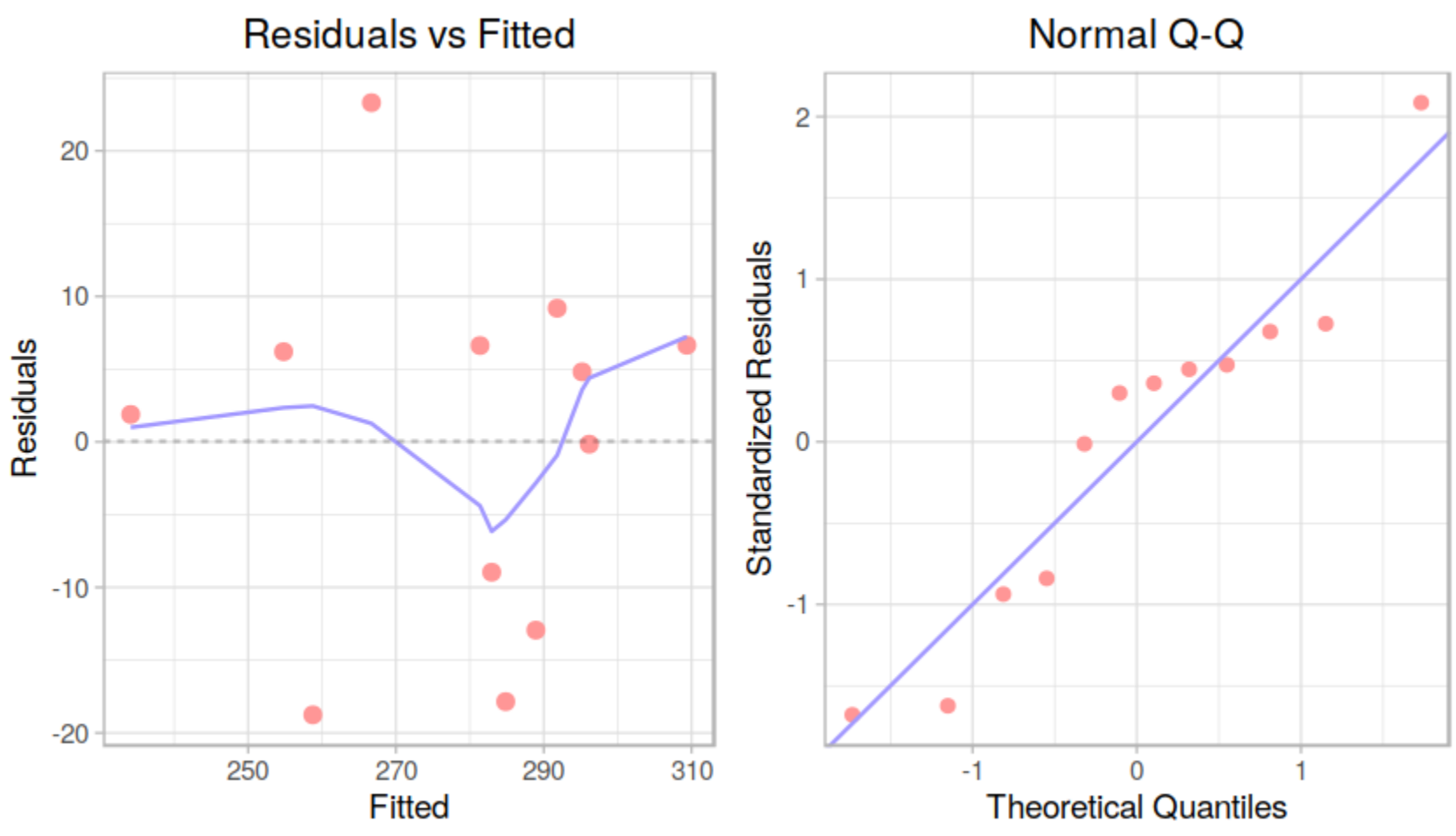
- j) Haga una gráfica de residuos contra $\hat{y}$. Interprete esta gráfica.
- k) Construya una gráfica de probabilidad normal de los residuos y haga un comentario sobre la suposición de normalidad.

```
# LOWESS line (la misma linea que ajusta los valores usando plot(model)):
smoothed <- data.frame(with(df, lowess(x = model$fitted, y = model$residuals)))
```

```r
# Gráficos
options(repr.plot.width=7, repr.plot.height=4)
res_vs_fit <- ggplot(model) +
    geom_point(aes(x=model$fitted, y=model$residuals),color= '#ff9696', size=2) +
    geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff") +
    geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
    ggtitle("Residuals vs Fitted") +
    xlab("Fitted") +
    ylab("Residuals") +
    theme_light() +
    theme(plot.title = element_text(hjust = 0.5))

qq_plot <- ggplot(model, aes(qqnorm(.stdresid)[[1]], .stdresid)) +
    geom_point(na.rm = TRUE,color= '#ff9696') +
    geom_abline(col="#a399ff") +
    xlab("Theoretical Quantiles") +
    ylab("Standardized Residuals") +
    ggtitle("Normal Q-Q") +
    theme_light() +
    theme(plot.title = element_text(hjust = 0.5))

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



La gráfica de residuos vs valores ajustados sugiere que la varianza de los errores podría no ser constante, dado que se observa mayor dispersión en los valores intermedios que en los extremos. Sin embargo, esto no implica una violación grave del principio de varianza constante.

En cuanto a la gráfica de probabilidad normal, los valores no se ajustan a una recta, lo cual pone en duda la normalidad de los residuos.

Se debe tener en cuenta que la cantidad de muestras es chica y esto dificulta un análisis visual concluyente.

Ejercicio 3

Se efectúa un estudio sobre el desgaste de un cojinete (y) y su relación con $x_1 = \text{viscosidad del aceite}$, y $x_2 = \text{carga}$. Se obtienen los siguientes datos:

y	x_1	x_2
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

```
df <- data.frame(read.table("./TP3_tables/data3.txt", header = TRUE))
names(df) <- c("y", "x1", "x2")
```

```
display_markdown("#### **Resumen de los datos**")
df_summary <- as.data.frame(apply(df, 2, summary))
df_summary$values <- rownames(df_summary)
df_summary <- df_summary[,c(4,1,2,3)]
colnames(df_summary)[1] <- " "
rownames(df_summary) <- c()
table <- formattable(df_summary, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL)
```

Resumen de los datos

	y	x1	x2
Min.	91.00	1.600	816.000
1st Qu.	116.00	17.125	902.750
Median	148.50	27.500	1086.500
Mean	154.00	25.850	1066.333
3rd Qu.	187.75	38.250	1179.500
Max.	230.00	43.000	1357.000

a) Ajuste un modelo de regresión lineal múltiple a los datos contenidos en la tabla anterior.

```
model_1 <- lm(y ~ x1 + x2, data=df)
```

```
display_markdown('#### **Primer modelo** $\\left(y = \\beta_0 + \\beta_1 x_1 + \\beta_2 x_2 \\right)$')
coef <- as.data.frame(x=model_1$coefficients)
colnames(coef) <- c('')
rownames(coef) <- c('β0', 'β1', 'β2')
coef <- as.data.frame(t(coef))
table <- formattable(coef, align = 'c')
as.htmlwidget(table, width="50%", height=NULL)
```

Primer modelo $(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2)$

β0	β1	β2
350.9943	-1.271994	-0.1539042

b) Utilice el modelo para predecir el desgaste cuando $x_1 = 25$ y $x_2 = 1000$.

```
x_test <- data.frame(x1=25, x2=1000)
y_test_1 <- predict(model_1, x_test, interval = "none", type = "response")
display_markdown(paste("$\\hat{y} = \\hat{\\beta}_0 + \\hat{\\beta}_1 \\times 25 + \\hat{\\beta}_2 \\time
s 1000 = $", round(y_test_1, 4)))
```

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 25 + \hat{\beta}_2 \times 1000 = 165.2902$

c) Ajuste un modelo de regresión lineal múltiple con un término de interacción entre los datos

```
model_2 <- lm(y ~ x1 * x2, data=df)
```

```
display_markdown('#### **Segundo modelo** $\\left(y = \\beta_0 + \\beta_1 x_1 + \\beta_2 x_2 + \\beta_{1:2} x_1 x_2\\right)$')
coef <- as.data.frame(x=model_2$coefficients)
colnames(coef) <- c('')
rownames(coef) <- c('β0', 'β1', 'β2', 'β12')
coef <- as.data.frame(t(coef))
table <- formattable(coef, align = 'c')
as.htmlwidget(table, width="50%", height=NULL)
```

Segundo modelo $\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1:2} x_1 x_2\right)$

	β0	β1	β2	β12
	125.8655	7.758641	0.09430397	-0.009185794

d) Utilice el modelo del inciso ***c)*** para hacer una predicción cuando $x_! = 25$ y $x_2 = 1000$. Compare esta predicción con el valor calculado en el inciso ***b)***.

```
y_test_2 <- predict(model_2, x_test, interval = "none", type = "response")
display_markdown(paste("$\\hat{y} = \\hat{\\beta}_0 + \\hat{\\beta}_1 \\times 25 + \\hat{\\beta}_2 \\time
s 1000 + \\hat{\\beta}_{1:2} \\times 25 \\times 1000 = $", round(y_test_2, 4)))
```

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 25 + \hat{\beta}_2 \times 1000 + \hat{\beta}_{1:2} \times 25 \times 1000 = 184.4907$

El primer modelo predijo un valor de 165.2902, mientras que el segundo modelo predijo 184.4907. Algunas observaciones:

- En ambos casos el test de significancia de la regresión arrojó p-valores mayores a 0.05 $(\text{p-valor}_1 = 0.05138 \text{ y p-valor}_2 = 0.1169)$. Por lo tanto se aceptan las hipótesis nulas de que la variable dependiente (y) no depende de las variables regresoras $(x_1 \text{ y } x_2)$.
- En el primer modelo las estimaciones de los coeficientes β_1 y β_2 arrojaron valores negativos, lo cual indica una relación inversa entre la variable dependiente y las regresoras. En el segundo modelo estas estimaciones resultaron positivas, mientras que la estimación del coeficiente del término de interacción fue negativa.

e) Pruebe la significancia de la regresión utilizando $\alpha = 0.05$. ¿Cuál es el p-valor de la prueba? ¿Qué conclusiones se obtienen?

Para determinar la significancia de la regresión se prueba la hipótesis nula:

$$H_{0_1} : \beta_1 = \beta_2 = 0$$

Para ello se utiliza el estadístico F.

```
display_markdown('#### **Primer modelo** $\\left(y = \\beta_0 + \\beta_1 x_1 + \\beta_2 x_2 \\right)$')
f_stat<- as.data.frame(t(summary(model_1)$fstatistic))
f_stat$p_value <- c(pf(f_stat[,1], f_stat[,2], f_stat[,3], lower.tail=FALSE))
f_stat <- as.data.frame(cbind('F-statistic', f_stat))
colnames(f_stat) <- c(' ', 'value', 'df1', 'df2', 'p-value')
table <- formattable(f_stat, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL, )
```

Primer modelo $\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2\right)$

	value	df1	df2	p-value
F-statistic	9.353035	2	3	0.05138189

Para el primer modelo, el valor del estadístico es $F = 9.353$ y el $\text{p-valor} = 0.05138 > 0.05$. Como se había anticipado en el inciso anterior, no es posible rechazar la hipótesis nula.

De forma análoga para el segundo modelo:

$$H_{0_2} : \beta_1 = \beta_2 = \beta_{1:2} = 0$$

```
display_markdown('#### **Segundo modelo** $\\left(y = \\beta_0 + \\beta_1 x_1 + \\beta_2 x_2 + \\beta_{1:
2} x_1 x_2\\right)$')
f_stat<- as.data.frame(t(summary(model_2)$fstatistic))
f_stat$p_value <- c(pf(f_stat[,1], f_stat[,2], f_stat[,3], lower.tail=FALSE))
f_stat <- as.data.frame(cbind('F-statistic', f_stat))
colnames(f_stat) <- c(' ', 'value', 'df1', 'df2', 'p-value')
table <- formattable(f_stat, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL, )
```

Segundo modelo $\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1:2} x_1 x_2\right)$

	value	df1	df2	p-value
F-statistic	7.714138	3	2	0.116915

En este caso el valor del estadístico es $F = 7.714$ y el $\text{p-valor} = 0.1169 > 0.05$. Nuevamente, no es posible rechazar la hipótesis nula.

f) Utilice la prueba t para evaluar la contribución al modelo de cada variable de regresión. Si se emplea $\alpha = 0.05$, ¿qué conclusiones pueden obtenerse?

Se utiliza el estadístico t de Student para probar la hipótesis nula de que el valor verdadero de cada coeficiente β_j es igual a 0.

```
display_markdown('#### **Primer modelo** $\\left(y = \\beta_0 + \\beta_1 x_1 + \\beta_2 x_2 \\right)$')
model_coef <- format(round(summary(model_1)$coefficients, 5), nsmall=5)
model_coef <- cbind(coef=c('β0', 'β1', 'β2'), model_coef)
colnames(model_coef)[1] <- "Coefficient"
rownames(model_coef) <- c()
table <- formattable(as.data.frame(model_coef, 5), align='c', list('Coefficient' = formatter("span",style
= ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL)
```

Primer modelo $\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2\right)$

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β0	350.99427	74.75307	4.69538	0.01827
β1	-1.27199	1.16914	-1.08797	0.35620
β2	-0.15390	0.08953	-1.71903	0.18410

Solo el p-valor del coeficiente β_0 es menor que 0.05, lo cual sugiere que las variables independientes no son útiles para predecir la variable respuesta.

Para el segundo modelo:

```
display_markdown('#### **Segundo modelo** $\\left(y = \\beta_0 + \\beta_1 x_1 + \\beta_2 x_2 + \\beta_{1:
2} x_1 x_2\\right)$')
model_coef <- format(round(summary(model_2)$coefficients, 5), nsmall=5)
model_coef <- cbind(coef=c('β0', 'β1', 'β2', 'β12'), model_coef)
colnames(model_coef)[1] <- "Coefficient"
rownames(model_coef) <- c()
table <- formattable(as.data.frame(model_coef, 5), align='c', list('Coefficient' = formatter("span",style
= ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL)
```

Segundo modelo $\left(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1:2} x_1 x_2\right)$

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
β0	125.86555	197.95717	0.63582	0.58994
β1	7.75864	7.51479	1.03245	0.41036
β2	0.09430	0.22066	0.42738	0.71072
β12	-0.00919	0.00756	-1.21447	0.34850

En este caso, todos los p-valores son mayores que 0.05. Esto indicaría que ninguno de los regresores contribuye al modelo.

g) Utilice el método de la suma adicional de cuadrados para investigar la utilidad que tiene la adición de la variable $x_2 = \text{carga}$, a un modelo que ya contiene a la variable $x_1 = \text{viscosidad del aceite}$. Utilice $\alpha = 0.05$.

Se quiere investigar la contribución de la variable x_2 a un modelo que continene la variable x_1. Para ello se plantea un modelo completo que contiene ambas variables y un modelo reducido que solo contiene a x_1. El modelo completo es:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Y el modelo reducido:

$$Y = \beta_0 + \beta_1 x_1$$

Se puede utilizar un estadístico F para probar la hipótesis $H_0 : \beta_2 = 0$. La prueba se realiza ajustando ambos modelos y comparando sus sumas residuales de cuadrados. Sean SSE_{MC} la suma de residuos cuadrados del modelo completo y SSE_{MR} la del modelo reducido, el estadístico resulta:

$$F_0 = \frac{\left(SSE_{MR} - SSE_{MC}\right) / \left(k - r\right)}{SSE_{MC} / \left(n - p\right)}$$

donde:

- $k =$ número de regresores en el modelo completo $= 2$
- $r =$ número de regresores en el modelo reducido $= 1$
- $n =$ cantidad de residuos $= 6$
- $p =$ número de parámetros en el modelo completo $= 3$

Modelo completo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

```
full_model <- lm(y ~ x1 + x2, data=df)
sse_full <- sum((predict(full_model) - df$y)^2)
display_markdown(paste('$SSE_{MC} =$', round(sse_full, 4)))
```

$SSE_{MC} =$ 1950.4222

Modelo reducido: $y = \beta_0 + \beta_1 x_1$

```
reduced_model <- lm(y ~ x1, data=df)
sse_reduced <- sum((predict(reduced_model) - df$y)^2)
display_markdown(paste('$SSE_{MR} =$', round(sse_reduced, 4)))
```

$SSE_{MR} =$ 3871.6309

```
# Cálculo de F0
k <- 2; r <- 1; n <- nrow(df); p <- 3
f0 <- ((sse_reduced - sse_full) / (k - r)) / (sse_full / (n - p))
display_markdown(paste('$F_0 =$', round(f0, 4)))
```

$F_0 =$ 2.9551

```
# Cálculo del p-valor
alpha <- 0.05
p_value <- 1 - pf(f0, df1=k-r, df2=n-p)
display_markdown(paste('$\\text{p-valor} =$', round(p_value, 4)))
```

$\text{p-valor} =$ 0.1841

El p-valor de la prueba es $0.1841 > 0.05$. Por lo tanto no es posible rechazar la hipótesis nula y se concluye que la adición de la variable x_2 no contribuye al modelo de forma significativa.

h) Encuentre intervalos de confianza del 99% para β_1 y β_2.

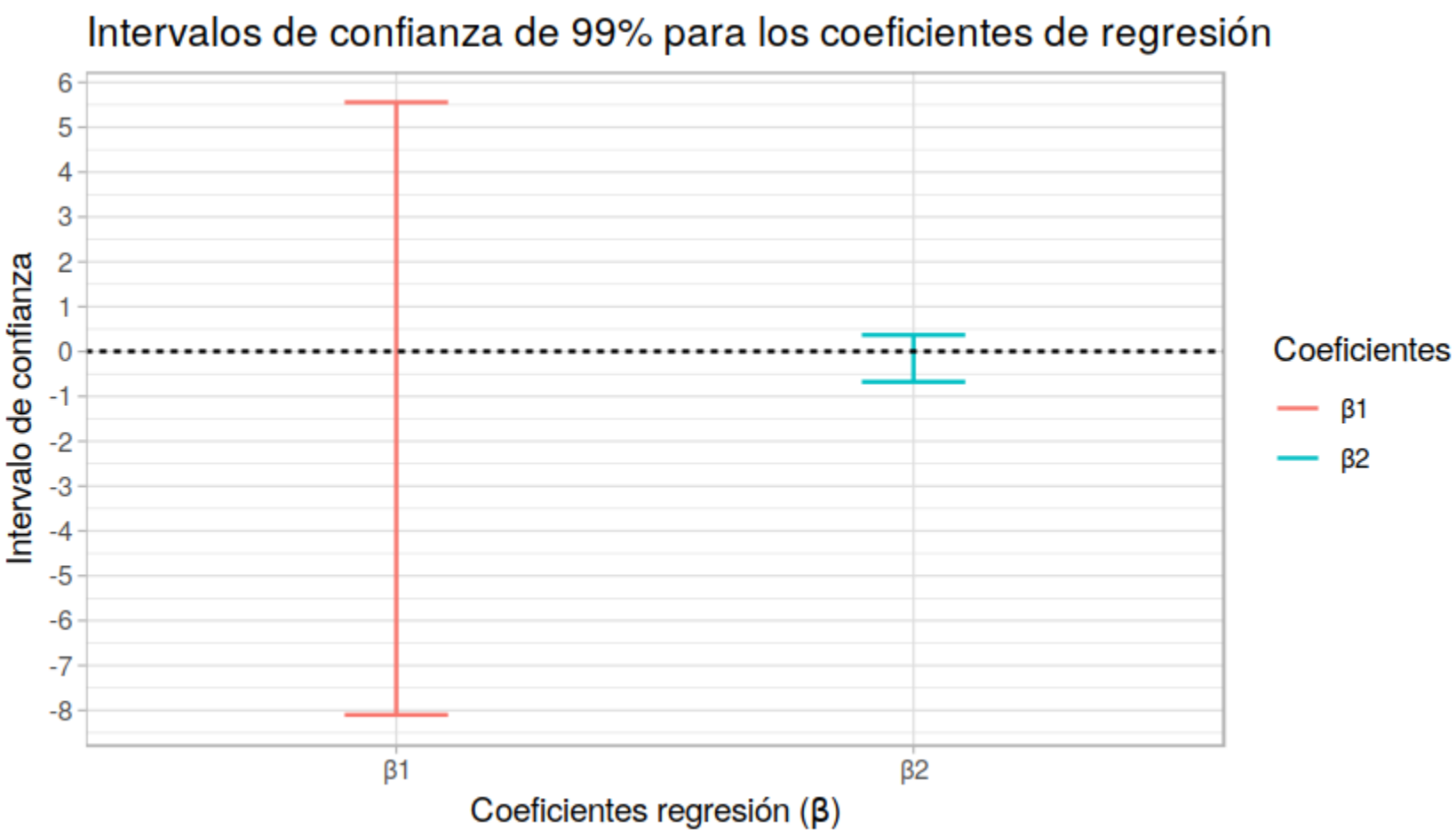
```
display_markdown('Intervalos de confianza del $99\\%$ para $\\beta_1$ y $\\beta_2$:')
conf_int <- as.data.frame(confint(full_model, c('x1', 'x2'), level = 0.99))
conf_int <- cbind(c('β1', 'β2'), conf_int)
colnames(conf_int)[1] <- ' '
rownames(conf_int) <- c()
as.htmlwidget(formattable(conf_int, align='c', list(' ' = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left')))), width="50%")
```

Intervalos de confianza del 99% para β_1 y β_2:

	0.5 %	99.5 %
β1	-8.1008357	5.5568468
β2	-0.6768389	0.3690305

```
colnames(conf_int) <- c('lwr', 'upr')
conf_int$avg <- apply(conf_int, 1, mean)
coef <- c('β1', 'β2')
```

```
# Gráfico de los intervalos de confianza
options(repr.plot.width=7, repr.plot.height=4)
ggplot(conf_int, aes(coef, avg, colour=coef)) +
    ggtitle("Intervalos de confianza de 99% para los coeficientes de regresión") +
    geom_errorbar(aes(ymin=lwr, ymax=upr), width = 0.2) +
    labs(x=TeX('Coeficientes regresión ($\\beta$ )'), y='Intervalo de confianza', color='Coeficientes') +
    geom_hline(yintercept=0, linetype="dashed", col="black") +
    scale_y_continuous(breaks=round(seq(min(conf_int) - 1, max(conf_int) + 1), 0)) +
    theme_light()
```



Ambos intervalos incluyen al cero, y por lo tanto no es posible rechazar las hipótesis $H_{0_1} : \beta_1 = 0$ y $H_{0_2} : \beta_2 = 0$.

i) Vuelva a calcular los intervalos de confianza del inciso h) después de añadir al modelo el término de interacción $x_1 x_2$. Compare la longitud de estos intervalos con las de los calculados en el inciso h). ¿La longitud de estos intervalos proporciona alguna información sobre la contribución al modelo del término de interacción?

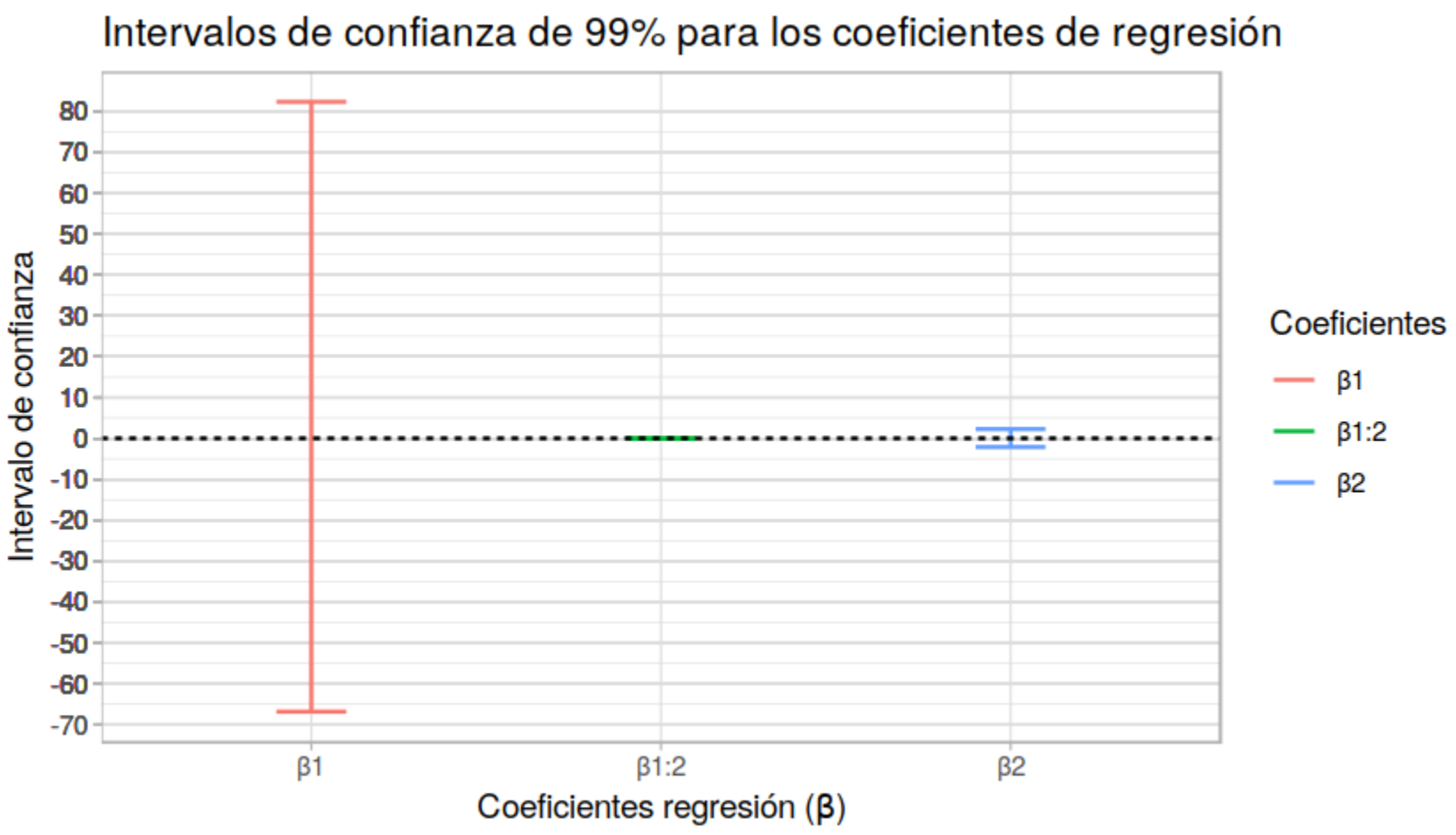
```
inter_model <- lm(y ~ x1*x2, data=df)
display_markdown('Intervalos de confianza del $99\\%$ para $\\beta_1$, $\\beta_2$ y $\\beta_{1:2}$:')
conf_int <- as.data.frame(confint(inter_model, c('x1', 'x2', 'x1:x2'), level = 0.99))
conf_int <- cbind(c('β1', 'β2', 'β1:2'), conf_int)
colnames(conf_int)[1] <- ' '
rownames(conf_int) <- c()
as.htmlwidget(formattable(conf_int, align='c', list(' ' = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left')))), width="50%")
```

Intervalos de confianza del 99% para β_1, β_2 y $\beta_{1:2}$:

	0.5 %	99.5 %
β1	-66.82452015	82.34180241
β2	-2.09568096	2.28428889
β1:2	-0.08425356	0.06588197

```
colnames(conf_int) <- c('lwr', 'upr')
conf_int$avg <- apply(conf_int, 1, mean)
coef <- rownames(conf_int)
```

```
# Gráfico de los intervalos de confianza
options(repr.plot.width=7, repr.plot.height=4)
ggplot(conf_int, aes(coef, avg, colour=coef)) +
    ggtitle("Intervalos de confianza de 99% para los coeficientes de regresión") +
    geom_errorbar(aes(ymin=lwr, ymax=upr), width = 0.2) +
    labs(x=TeX('Coeficientes regresión ($\\beta$ )'), y='Intervalo de confianza', color='Coeficientes') +
    geom_hline(yintercept=0, linetype="dashed", col="black") +
    scale_y_continuous(breaks=round(seq(min(conf_int) - 1, max(conf_int) + 1), -1)) +
    theme_light()
```



Obervando la longitud de los intervalos de confianza, se puede asegurar con un alto grado de seguridad que el término de interacción no contribuye al modelo.

j) Para el modelo que utiliza las variables de regresión x_1 y x_2 calcule R^2.

```
r_squared <- summary(full_model)$r.squared
display_markdown(paste('$R^2 =$', round(r_squared, 4)))
```

$R^2 = 0.8618$

k) ¿Qué sucede con el valor de R^2 cuando se añade al modelo un término de interacción x_1x_2?

```
r_squared <- summary(inter_model)$r.squared
display_markdown(paste('$R^2 =$', round(r_squared, 4)))
```

$R^2 = 0.9205$

La adición del término de interacción x_1x_2 logró un incremento del coeficiente de determinación.

Ejercicio 4

Se ajusta el modelo de regresión $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ a una muestra de $n = 25$ observaciones. Los cocientes t calculados

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{V\left(\hat{\beta}_j\right)}}$$

son los siguientes:

	β_1	β_2	β_3
t_0	4.82	8.21	0.98

a) Encuentre el p-valor para cada uno de los estadísticos t.

```
n <- 25 # cantidad de observaciones
t_values <- c(4.82, 8.21, 0.98) # cocientes calculados
p <- length(t_values) + 1 # número de coeficientes = 4
df <- n - p # grados de libertad = número de coeficientes = 4
p_values <- pt(q=t_values, df=df, lower.tail=FALSE)
```

```
display_html(sprintf('
<table style="width: 40%%;">
    <thead style="text-align: center">
        <th style="text-align: center; font-weight: normal">Coeficiente</th>
        <th style="text-align: center">$t_0$</th>
        <th style="text-align: center">$\\text{p-valor}$</th>
    </thead>
    <tbody>
        <tr>
            <td style="text-align: center">$\\beta_1$</td>
            <td style="text-align: center">%.4f</td>
            <td style="text-align: center">%.e</td>
        </tr>
        <tr>
            <td style="text-align: center">$\\beta_2$</td>
            <td style="text-align: center">%.2f</td>
            <td style="text-align: center">%.4f</td>
        </tr>
        <tr>
            <td style="text-align: center">$\\beta_3$</td>
            <td style="text-align: center">%.2f</td>
            <td style="text-align: center">%.4f</td>
        </tr>
    </tbody>
</table>
', t_values[1], p_values[1], t_values[2], p_values[2], t_values[3], p_values[3]))
```

Coeficiente	t_0	p-valor
β_1	4.8200	5e-05
β_2	8.21	0.0000
β_3	0.98	0.1691

b) Si se emplea $\alpha = 0.05$, ¿qué conclusiones pueden obtenerse sobre la variable de regresión x_3?. ¿Es posible que esta variable de regresión tenga una contribución significativa en el modelo?

El test $H_0 : \beta_3 = 0$ arrojó un $\text{p-valor} = 0.1913 > 0.05$, lo cual indica que la variable regresora x_3 no tiene un aporte significativo en el modelo propuesto. En general, este resultado sugiere que la variable x_3 debe ser removida del modelo. También existe la posibilidad de que la relación entre x_3 y la variable respuesta no sea lineal. En este caso, se podría aplicar transformaciones a x_3 para linealizar la relación.

Ejercicio 5

Los datos que aparecen a continuación se recopilaron durante un experimento para determinar el cambio en la eficiencia del impulso (y, en por ciento), a medida que cambia el ángulo de divergencia de la nariz de un cohete (x).

y	24.60	24.71	23.90	39.50	39.60	57.12	67.11	67.24	67.15	77.87	80.11	84.67
x	4.00	4.00	4.00	5.00	5.00	6.00	6.50	6.50	6.75	7.00	7.10	7.30

```r
x <- c(4.00, 4.00, 4.00, 5.00, 5.00, 6.00, 6.50, 6.50, 6.75, 7.00, 7.10, 7.30)
y <- c(24.60, 24.71, 23.90, 39.50, 39.60, 57.12, 67.11, 67.24, 67.15, 77.87, 80.11, 84.67)
data <- as.data.frame(cbind(x, y))
```

```r
display_markdown("#### **Resumen de los datos**")
df_summary <- as.data.frame(apply(data, 2, summary))
df_summary$values <- rownames(df_summary)
df_summary <- df_summary[,c(3,1,2)]
colnames(df_summary)[1] <- " "
rownames(df_summary) <- c()
table <- formattable(df_summary, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL)
```

Resumen de los datos

	x	y
Min.	4.0000	23.9000
1st Qu.	4.7500	35.8025
Median	6.2500	62.1150
Mean	5.7625	54.4650
3rd Qu.	6.8125	69.8975
Max.	7.3000	84.6700

a) Ajuste a los datos un modelo de segundo orden.

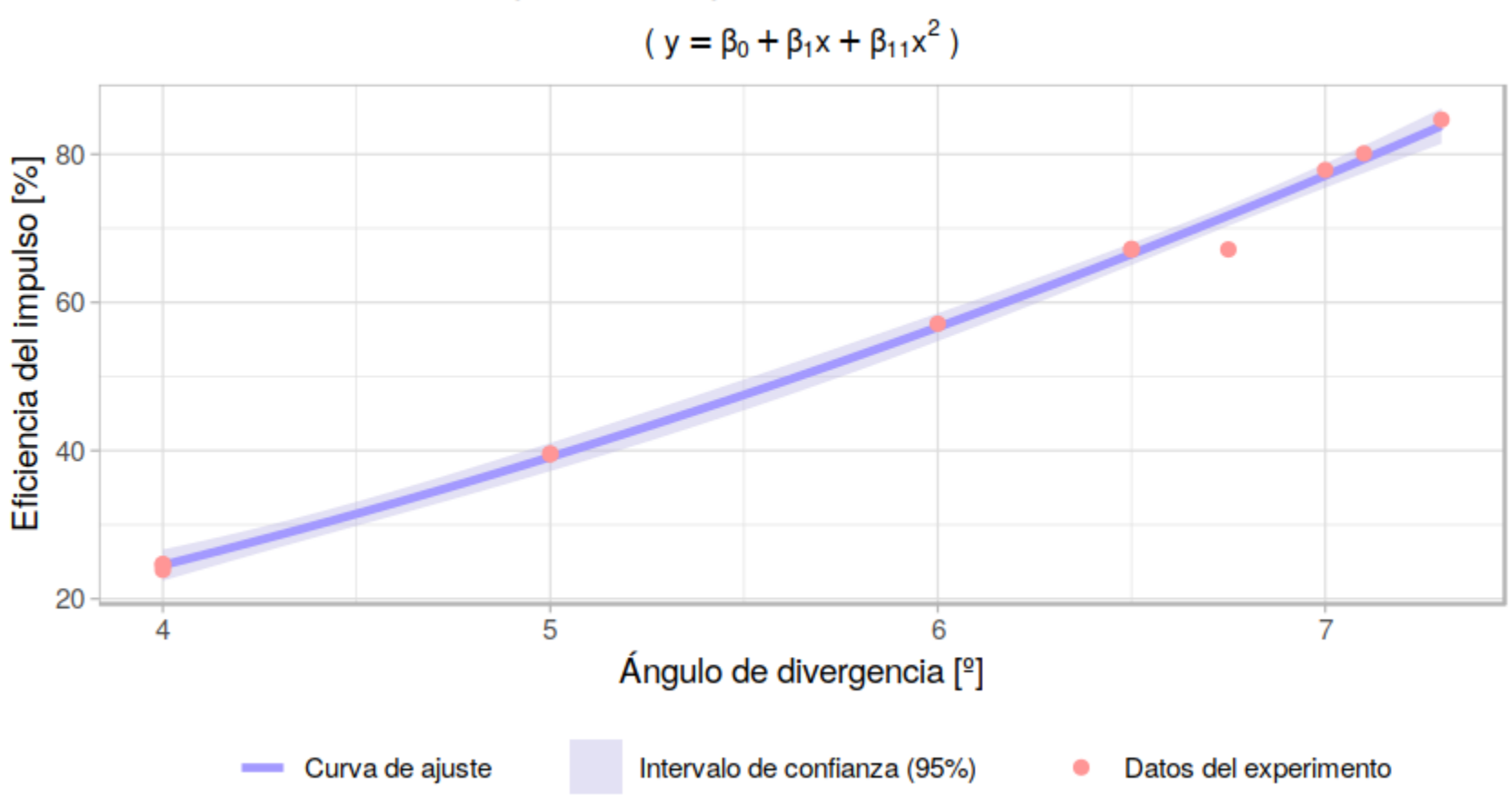
Modelo de segundo orden: $y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$

```r
model <- lm(y ~ x + I(x^2), data=data)
```

```r
coef <- as.data.frame(x=model$coefficients)
colnames(coef) <- c('')
rownames(coef) <- c('β0', 'β1', 'β11')
coef <- as.data.frame(t(coef))
table <- formattable(coef, align = 'c')
as.htmlwidget(table, width="50%", height=NULL)
```

β0	β1	β11
-4.459494	1.383712	1.467047

```
# Gráfico del modelo
options(repr.plot.width=7, repr.plot.height=4)
ggplot(data, aes(x=x, y=y)) +
    labs(
        title="Modelo de regresión de segundo orden para Eficiencia del impulso",
        subtitle=TeX("( $y = β_0 + β_1 x + β_{11} x^2$ )"),
        x="Ángulo de divergencia [º]",
        y="Eficiencia del impulso [%]") +
    stat_smooth(aes(x=x, y=y, col="Curva de ajuste", fill="Intervalo de confianza (95%)"),
                method="lm", formula=y ~ x + I(x^2), se=TRUE, size=1) +
    geom_point(aes(shape='Datos del experimento'), col='#ff9696') +
    theme_light() +
    theme(
        plot.title = element_text(size=10, hjust = 0.5),
        plot.subtitle = element_text(size=10, hjust = 0.5),
        legend.position = "bottom") +
    scale_fill_manual(NULL, values = '#BAB5E3') +
    scale_color_manual(NULL, values = '#a399ff') +
    scale_shape_manual(NULL, values = 19) +
    guides(
        color=guide_legend(override.aes = list(fill=NA), order=1),
        fill=guide_legend(override.aes = list(color=NA), order=2),
        shape=guide_legend(order=3))
```



Se observa un valor atípico entre los datos del experimento que no puede ser explicado por el modelo propuesto.

b) Pruebe la significancia de la regresión y la adecuación del ajuste, utilizando $\alpha = 0.05$.

```
f_stat<- as.data.frame(t(summary(model)$fstatistic))
f_stat$p_value <- c(pf(f_stat[,1], f_stat[,2], f_stat[,3], lower.tail=FALSE))
f_stat <- as.data.frame(cbind('F-statistic', f_stat))
colnames(f_stat) <- c(' ', 'value', 'df1', 'df2', 'p-value')
table <- formattable(f_stat, align='c', list(` ` = formatter("span",style = ~ style(
    'font-weight'='bold', 'text-align'='left'))))
as.htmlwidget(table, width="50%", height=NULL, )
```

	value	df1	df2	p-value
F-statistic	1044.995	2	9	2.213323e-11

El p-valor de la prueba de significancia de la regresión es $2.213e-11 < 0.05$, por lo que se rechaza la hipótesis nula $H_0 : \beta_1 = \beta_{1:1} = 0$.

Para medir la bondad del ajuste se utiliza el coeficiente de determinación:

$$R^2 = 1 - \frac{SSE}{SST}$$

```
r_squared <- summary(model)$r.squared
display_markdown(sprintf('$R^2 = %.4f$', r_squared))
```

$R^2 = 0.9957$

Este coeficiente indica qué proporción de la dispersión de la variable de salida es explicada por el modelo. Para el modelo propuesto $R^2 = 0.9957$, muy cercano al valor máximo 1, lo que significa que el nivel de ajuste es bueno.

c) Pruebe la hipótesis $\beta_{11} = 0$ con $\alpha = 0.05$.

Para probar la hipótesis $H_0 : \beta_{11} = 0$ se contrasta el modelo completo (MC) con un modelo reducido que no contiene el término cuadrático$(MR)(MR)$. Se utiliza un estadístico F:

$$F_0 = \frac{(SSE_{MR} - SSE_{MC}) \,/\, (k - r)}{SSE_{MC} / \,(n - p)}$$

Modelo completo: $y = \beta_0 + \beta_1 x + \beta_{11} x^2$

```
sse_full <- sum((predict(model) - data$y)^2)
display_markdown(paste('$SSE_{MC} =$', round(sse_full, 4)))
```

$SSE_{MC} =$ 24.7202

Modelo reducido: $y = \beta_0 + \beta_1 x$

```
reduced_model <- lm(y ~ x, data=data)
sse_reduced <- sum((predict(reduced_model) - data$y)^2)
display_markdown(paste('$SSE_{MR} =$', round(sse_reduced, 4)))
```

$SSE_{MR} =$ 48.9825

```
# Cálculo de F0
k <- 2; r <- 1; n <- nrow(data); p <- length(coef)
f0 <- ((sse_reduced - sse_full) / (k - r)) / (sse_full / (n - p))
display_markdown(paste('$F_0 =$', round(f0, 4)))
```

$F_0 =$ 8.8333

```
# Cálculo del p-valor
alpha <- 0.05
p_value <- 1 - pf(f0, df1=k-r, df2=n-p)
display_markdown(paste('$\\text{p-valor} =$', round(p_value, 4)))
```

$\text{p-valor} =$ 0.0156

El p-valor de la prueba es $0.0156 < 0.05$. Por lo tanto se rechaza la hipótesis nula y se concluye que la adición del término cuadrático x^2 tiene un aporte significativo al modelo.

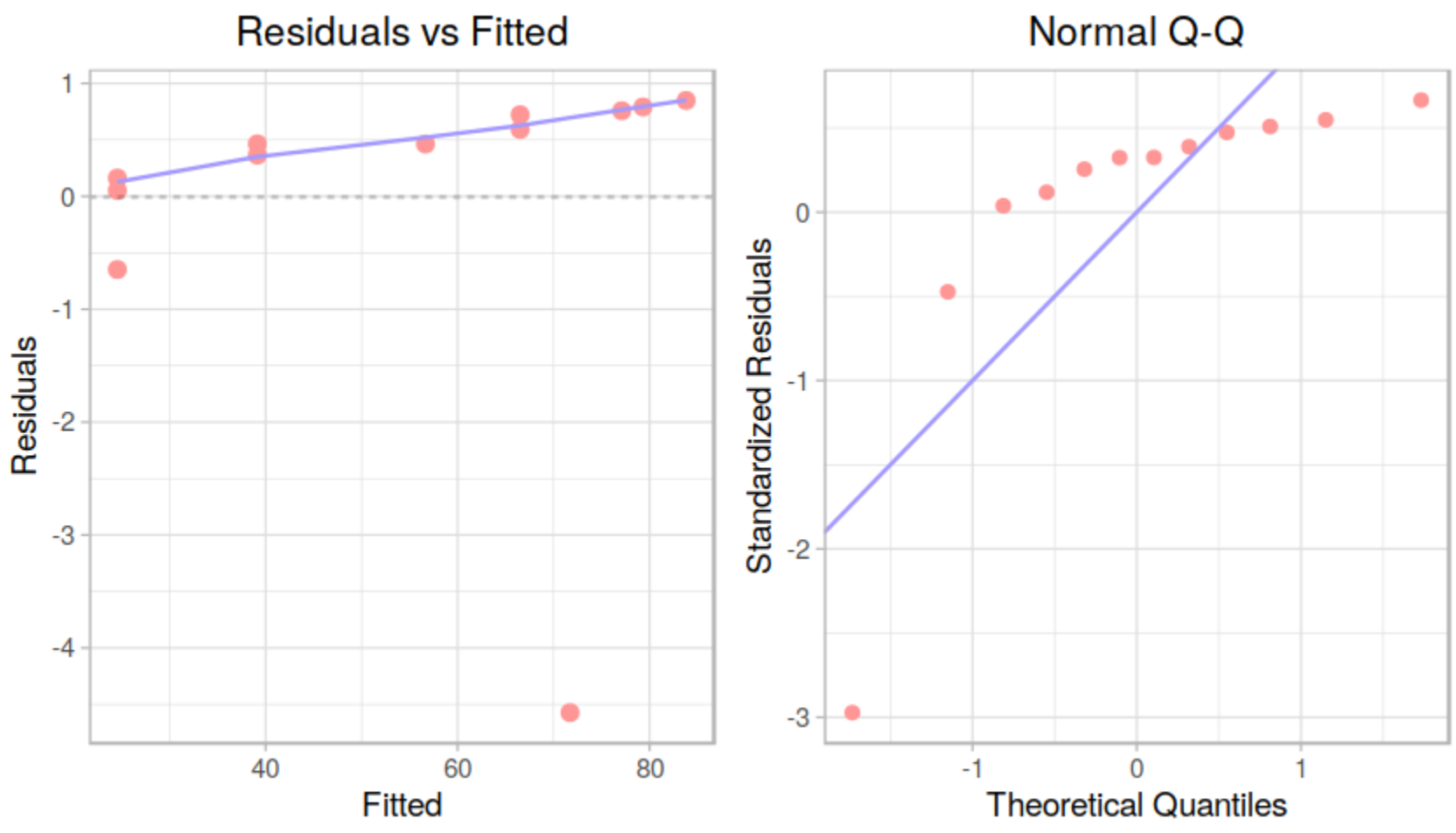
d) Haga una gráfica de los residuos y comente la adecuación del modelo.

```
# LOWESS line (la misma linea que ajusta los valores usando plot(model)):
smoothed <- data.frame(with(data, lowess(x = model$fitted, y = model$residuals)))
```

```r
# Gráficos
options(repr.plot.width=7, repr.plot.height=4)
res_vs_fit <- ggplot(model) +
    geom_point(aes(x=model$fitted, y=model$residuals), color= '#ff9696', size=2) +
    geom_path(data = smoothed, aes(x = x, y = y), col="#a399ff") +
    geom_hline(linetype = 2, yintercept=0, alpha=0.2) +
    ggtitle("Residuals vs Fitted") +
    xlab("Fitted") +
    ylab("Residuals") +
    theme_light() +
    theme(plot.title = element_text(hjust = 0.5))

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

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



En la gráfica de residuos vs valores ajustados se observa que la mayoría de los residuos tienen valores positivos, lo cual sugiere que la media de los residuos no es 0. Además, existe un valor significativamente atípico. Por otra parte, el gráfico Q-Q muestra claramente que los residuos no provienen de una distribución normal. Estos aspectos indican un incumplimiento de los supuestos necesarios para la validez del modelo.

e) Ajuste un modelo cúbico y pruebe la significancia del término cúbico, utilizando $\alpha = 0.05$

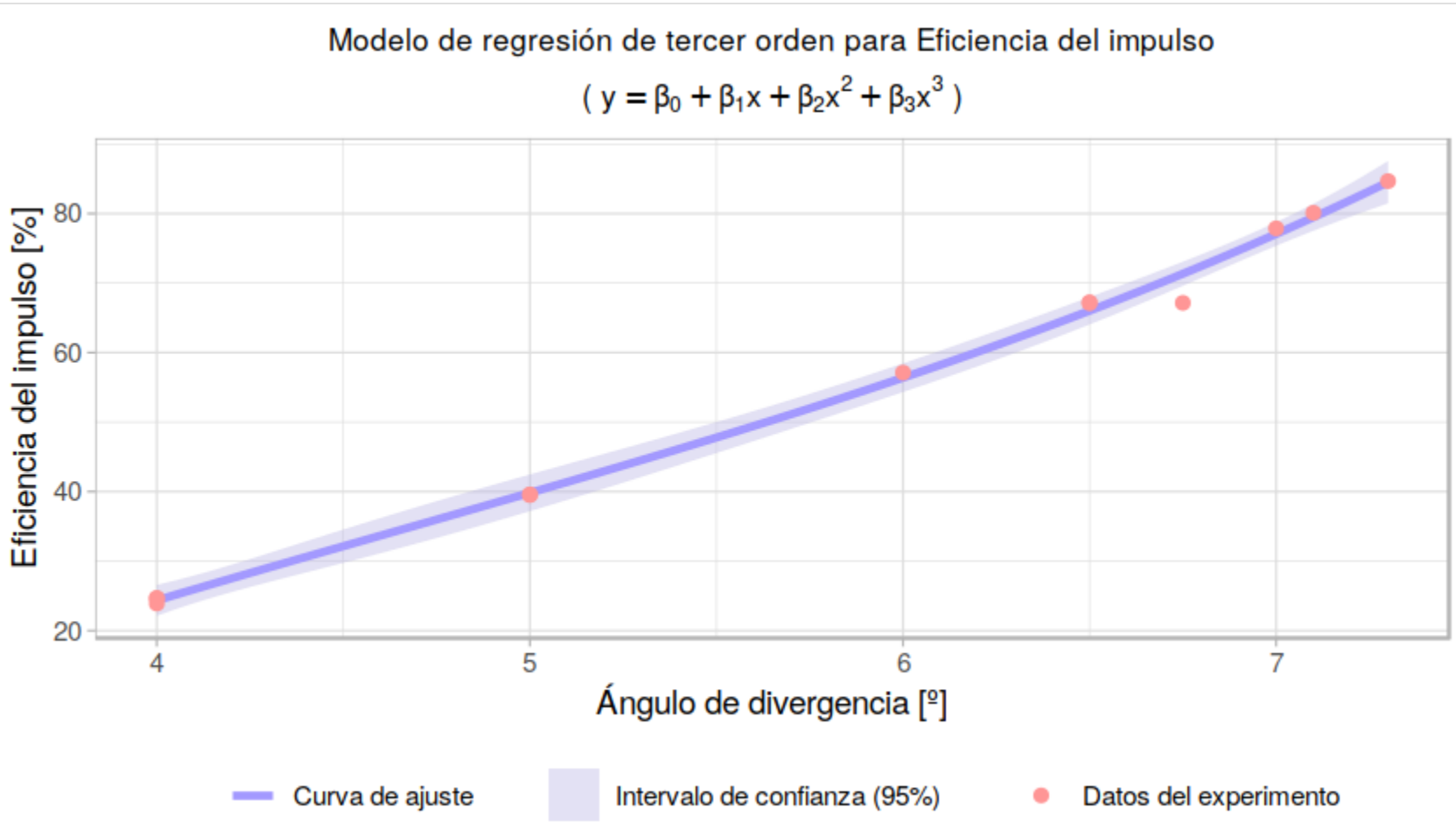
Modelo de tercer orden: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

```r
cubic_model <- lm(y ~ x + I(x^2) + I(x^3), data=data)
```

```r
coef <- as.data.frame(x=cubic_model$coefficients)
colnames(coef) <- c('')
rownames(coef) <- c('β0', 'β1', 'β2', 'β3')
coef <- as.data.frame(t(coef))
table <- formattable(coef, align = 'c')
as.htmlwidget(table, width="50%", height=NULL)
```

β0	β1	β2	β3
-87.35551	48.0087	-7.042795	0.5057031

```
# Gráfico del modelo
options(repr.plot.width=7, repr.plot.height=4)
ggplot(data, aes(x=x, y=y)) +
    labs(
        title="Modelo de regresión de tercer orden para Eficiencia del impulso",
        subtitle=TeX("( $y = β_0 + β_1 x + β_2 x^2 + β_3 x^3$ )"),
        x="Ángulo de divergencia [º]",
        y="Eficiencia del impulso [%]") +
    stat_smooth(aes(x=x, y=y, col="Curva de ajuste", fill="Intervalo de confianza (95%)"),
                method="lm", formula=y ~ x + I(x^2) + I(x^3), se=TRUE, size=1) +
    geom_point(aes(shape='Datos del experimento'), col='#ff9696') +
    theme_light() +
    theme(
        plot.title = element_text(size=10, hjust = 0.5),
        plot.subtitle = element_text(size=10, hjust = 0.5),
        legend.position = "bottom") +
    scale_fill_manual(NULL, values = '#BAB5E3') +
    scale_color_manual(NULL, values = '#a399ff') +
    scale_shape_manual(NULL, values = 19) +
    guides(
        color=guide_legend(override.aes = list(fill=NA), order=1),
        fill=guide_legend(override.aes = list(color=NA), order=2),
        shape=guide_legend(order=3))
```



De forma similar al caso anterior, se prueba la hipótesis $H_0 : \beta_3 = 0$ comparando el modelo completo con un modelo reducido, y utilizando un estadístico F.

Modelo completo: $y = \beta_0 + \beta x_1 + \beta_2 x^2 + \beta_3 x^3$

```
sse_full <- sum((predict(cubic_model) - data$y)^2)
display_markdown(paste('$SSE_{MC} =$', round(sse_full, 4)))
```

$SSE_{MC} = 22.4152$

Modelo reducido: $y = \beta_0 + \beta_1 x + \beta_2 x^2$

```
reduced_model <- lm(y ~ x + I(x^2), data=data)
sse_reduced <- sum((predict(reduced_model) - data$y)^2)
display_markdown(paste('$SSE_{MR} =$', round(sse_reduced, 4)))
```

$SSE_{MR} = 24.7202$

```
# Cálculo de F0
k <- 3; r <- 2; n <- nrow(data); p <- length(coef)
f0 <- ((sse_reduced - sse_full) / (k - r)) / (sse_full / (n - p))
display_markdown(paste('$F_0 =$', round(f0, 4)))
```

$F_0 = 0.8226$

```r
# Cálculo del p-valor
alpha <- 0.05
p_value <- 1 - pf(f0, df1=k-r, df2=n-p)
display_markdown(paste('$\\text{p-valor} =$', round(p_value, 4)))
```

$\text{p-valor} = 0.3909$

El p-valor de la prueba es $0.3909 > 0.05$. Por lo tanto no es posible rechazar la hipótesis nula y se concluye que la adición del término cúbico x^3 no contribuye al modelo de forma significativa.