

1.1. Correlación

(a)

(b)

(c)

(a)

(b)

(c)

(d)

(e)

Ejercicio 1.3. El archivo peso_edad_colest.xlsx disponible contiene registros correspondientes a 25 individuos respecto de su peso, su edad y el nivel de colesterol total en sangre.

(a)

(b)

(c)

(d)

(e)

Ejercicio 1.4.

(a)

(b)

(c)

(d)

Trabajo Practico Regresión Avanzada

Code ▼

Jose Valdes

2023-06-05

Hide

```
#limpio la memoria  
rm( list= ls(all.names= TRUE) ) #remove all objects  
gc( full= TRUE )                #garbage collection
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 511599 27.4   1139588 60.9   644245 34.5
## Vcells 913126  7.0    8388608 64.0  1635137 12.5
```

Se realiza validación de la instalación de los paquetes necesarios para ejecutar el script

[Hide](#)

```
# Bibliotecas a cargar

check_packages <- function(packages) {
  if (all(packages %in% rownames(installed.packages()))) {
    TRUE
  } else{
    cat(
      "Instalar los siguientes packages antes de ejecutar el presente script\n",
      packages[!(packages %in% rownames(installed.packages()))],
      "\n"
    )
  }
}

packages_needed <- c("readxl","ggplot2","MVN","gridExtra","aod","MASS","carData","car")

# Se llama a la funcion check_packages
check_packages(packages_needed)
```

```
## [1] TRUE
```

[Hide](#)

```
library(readxl)
library(ggplot2)
library(MVN)
library(gridExtra)
library(aod)
library(MASS)
library(carData)
library(car)
```

1.1. Correlación

Ejercicio 1.1.

En el archivo grasacerdos.xlsx se encuentran los datos del peso vivo (PV, en Kg) y al espesor de grasa dorsal (EGD, en mm) de 30 lechones elegidos al azar de una población de porcinos Duroc Jersey del Oeste de la provincia de Buenos Aires. Se pide

(a)

Dibujar el diagrama de dispersión e interpretarlo.

Hide

```
library(readxl)
library(ggplot2)
library(MVN)
library(gridExtra)

grasacerdos<-read_excel("C:/Users/Josvaldes/Documents/Maestria/Austral/1ano/regresionAvanzada/TPRegresion/TPRegresion/grasacerdos.xlsx")
dim(grasacerdos)
```

```
## [1] 30  3
```

Hide

```
head(grasacerdos)
```

	Obs	PV	EGD
	<dbl>	<chr>	<chr>
	1	56,81	16,19
	2	70,40	22,00
	3	71,73	19,52
	4	75,10	31,00
	5	79,65	23,58
	6	51,43	16,58

6 rows

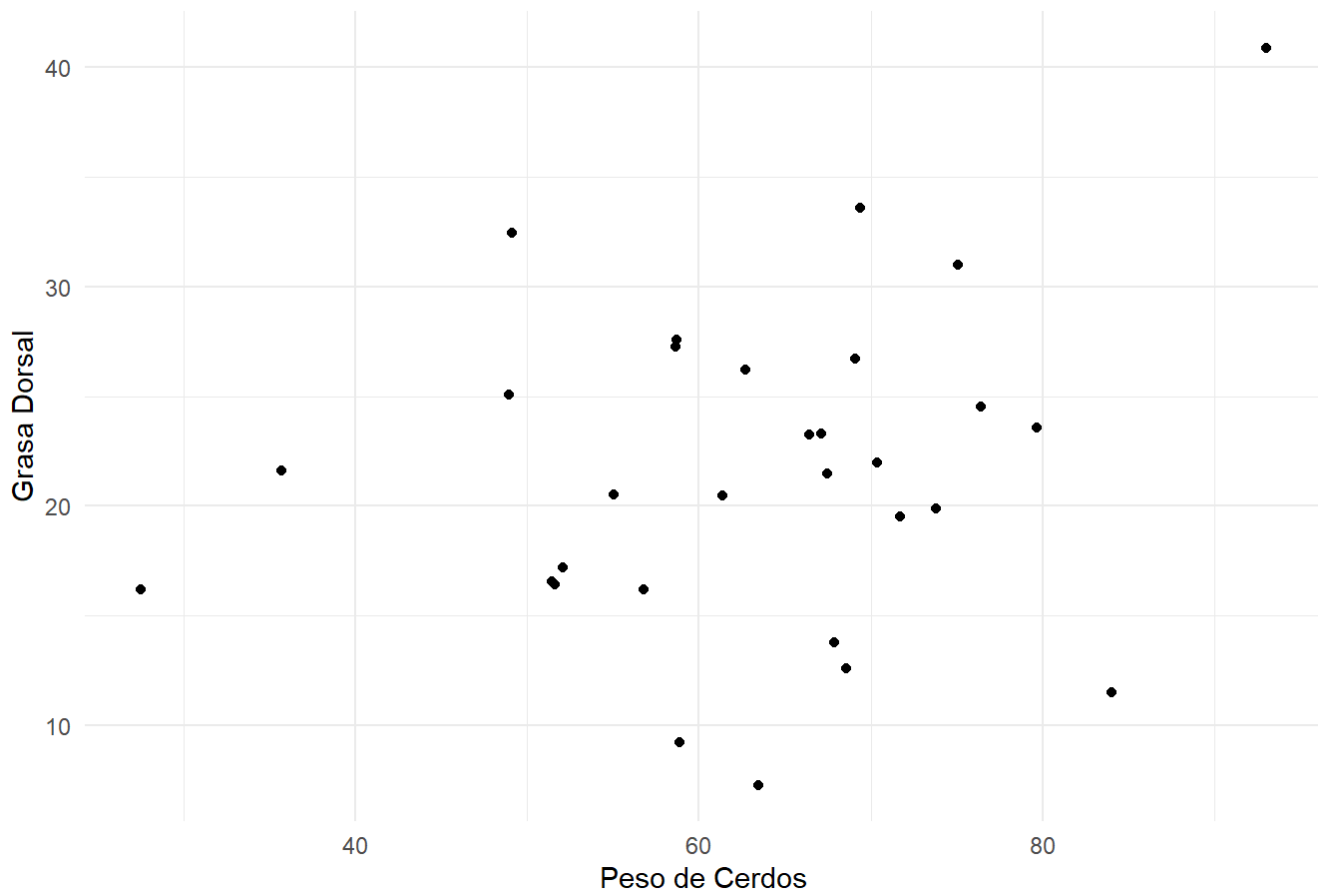
Hide

```
grasacerdos$PV <- as.numeric(gsub(",", ".", grasacerdos$PV))
grasacerdos$EGD <- as.numeric(gsub(",", ".", grasacerdos$EGD))
```

Hide

```
ggplot(grasacerdos, aes(PV, EGD)) +
  geom_point() +
  theme_minimal() +
  labs(x = "Peso de Cerdos", y = "Grasa Dorsal",
       title = ("Diagrama de Dispersi\u00F3n Peso de Cerdos vs Grasa Dorsal")) # se deja l
a Letra "ó" con \u00F3, que
es La representación Unicode de esa Letra
```

Diagrama de Dispersión Peso de Cerdos vs Grasa Dorsal



No se observa correlación entre las variables

(b)

Calcular el coeficiente de correlación muestral y explíquelo.

Hide

```
biNormTest <- mvn(grasacerdos, mvnTest = "hz")
print(biNormTest$multivariateNormality)
```

```
##          Test      HZ   p value MVN
## 1 Henze-Zirkler 0.6379234 0.3891766 YES
```

Por el resultado se puede sostener el supuesto de una distribución normal bivariada para estas variables. En tal sentido, se procede a realizar el test de Pearson para determinar la relación de las variables:

Hide

```
corCoeff <- cor(grasacerdos$PV, grasacerdos$EGD, method = "pearson")
corCoeff
```

```
## [1] 0.2543434
```

La prueba de correlación de Pearson muestra que existe una correlación positiva débil entre las variables. Esto significa que hay una tendencia a que los valores de las variables aumenten juntos, pero la relación no es muy fuerte.

(c)

¿Hay suficiente evidencia para admitir asociación entre el peso y el espesor de grasa? ($\alpha = 0,05$). Verifique los supuestos para decidir el indicador que va a utilizar.

Para determinar si hay suficiente evidencia para admitir una asociación entre el peso y el espesor de grasa, es necesario verificar los supuestos y luego utilizar un indicador apropiado para evaluar la correlación entre las variables.

A continuación, se describen los supuestos que se deben verificar antes de seleccionar el indicador:

1 - Supuesto de normalidad: Se debe verificar si las variables peso y espesor de grasa siguen una distribución normal. Esto se puede hacer mediante métodos gráficos, como histogramas o gráficos de Q-Q, y pruebas estadísticas, como el test de normalidad (por ejemplo, el test de Shapiro-Wilk).

Hide

```
shapiro.test(grasacerdos$PV)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  grasacerdos$PV  
## W = 0.97533, p-value = 0.6925
```

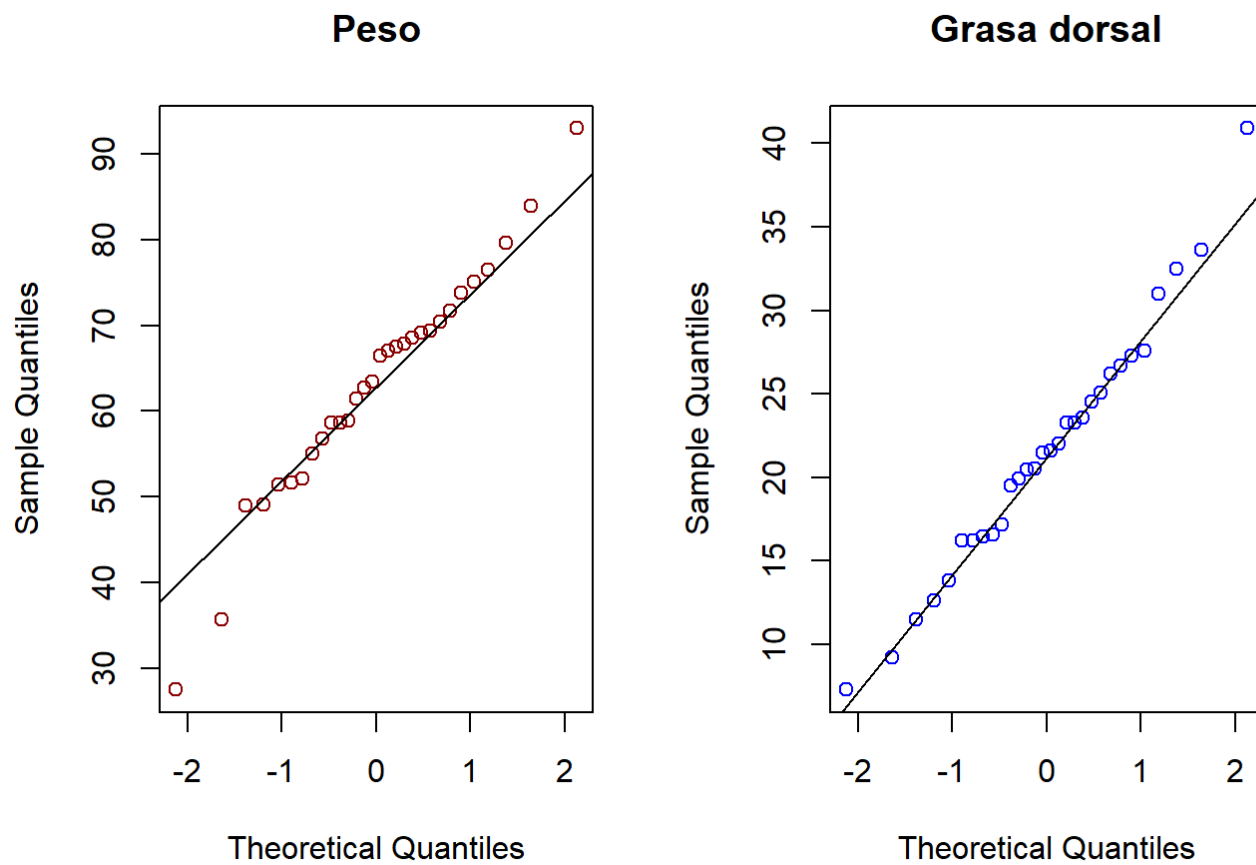
Hide

```
shapiro.test(grasacerdos$EGD)
```

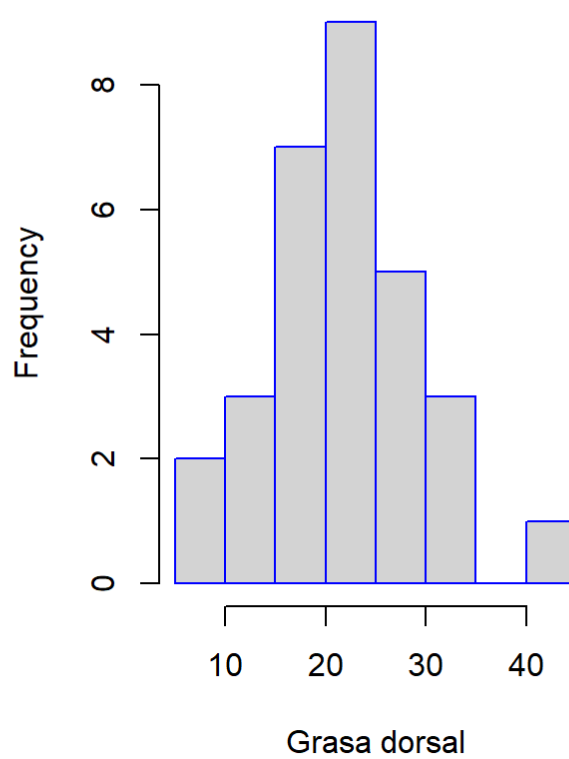
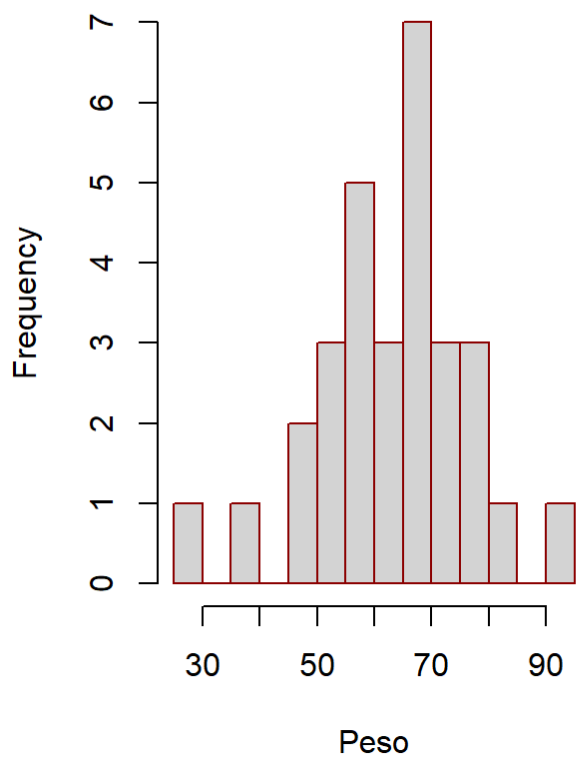
```
##  
##  Shapiro-Wilk normality test  
##  
## data:  grasacerdos$EGD  
## W = 0.98514, p-value = 0.9395
```

Hide

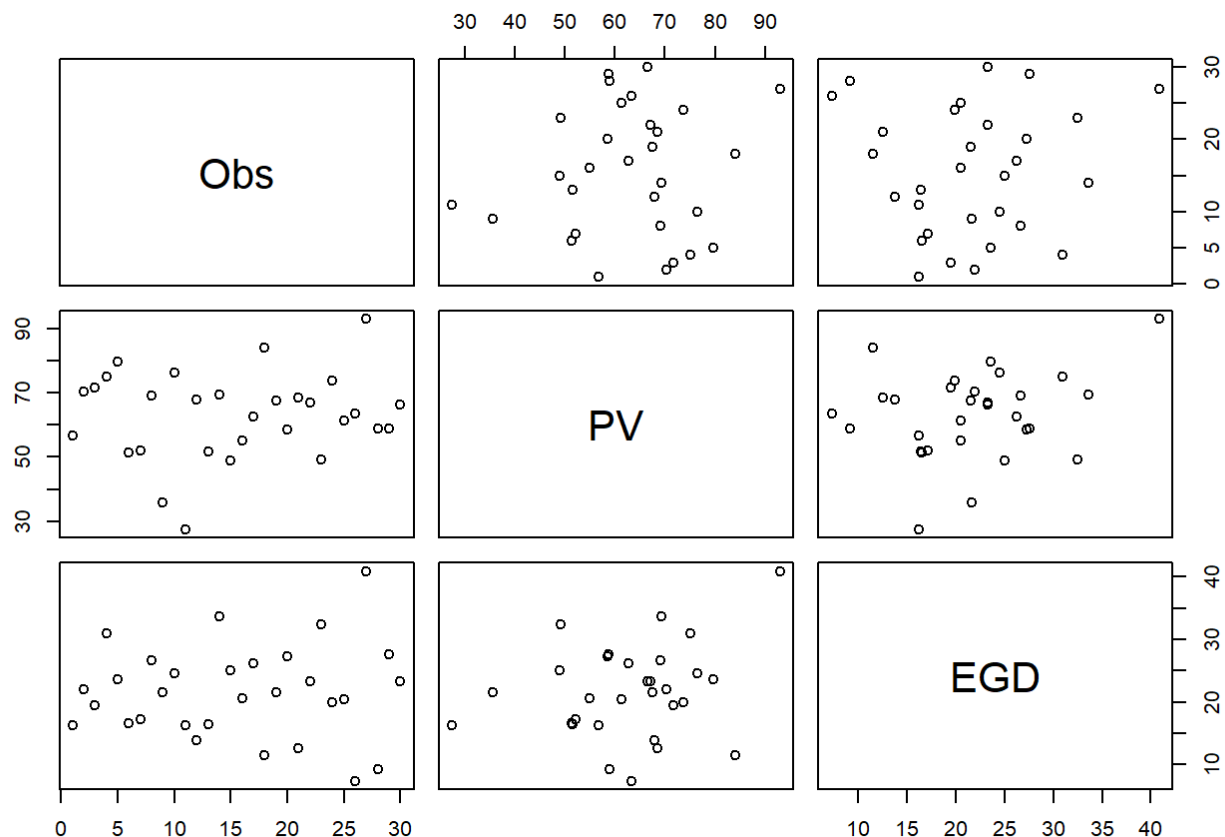
```
par(mfrow = c(1, 2))  
qqnorm(grasacerdos$PV, main = "Peso", col = "darkred")  
qqline(grasacerdos$PV)  
qqnorm(grasacerdos$EGD, main = "Grasa dorsal", col = "blue")  
qqline(grasacerdos$EGD)
```

[Hide](#)

```
par(mfrow = c(1, 2))  
hist(grasacerdos$PV, breaks = 10, main = "", xlab = "Peso", border = "darkred")  
hist(grasacerdos$EGD, breaks = 10, main = "", xlab = "Grasa dorsal", border = "blue")
```

[Hide](#)

```
par(bg="white")  
pairs(grasacerdos) # representa todos los diagramas de dispersión de a pares
```



2 - Supuesto de linealidad: Se debe verificar si la relación entre el peso y el espesor de grasa es lineal. Esto se puede explorar mediante un diagrama de dispersión o mediante técnicas de análisis exploratorio de datos.

3 - Supuesto de homogeneidad de varianzas: Se debe verificar si la varianza del espesor de grasa es constante en diferentes niveles de peso. Esto se puede evaluar mediante gráficos de dispersión y pruebas estadísticas, como el test de Levene.

Una vez que se han verificado los supuestos, puedes seleccionar un indicador apropiado para evaluar la asociación entre el peso y el espesor de grasa. Dado que estamos analizando una relación entre dos variables continuas, el coeficiente de correlación de Pearson sería un indicador adecuado.

Para determinar si hay suficiente evidencia para admitir la asociación entre el peso y el espesor de grasa, se puede realizar una prueba de hipótesis utilizando el coeficiente de correlación de Pearson. El enunciado de las hipótesis sería:

Hipótesis nula (H_0): No hay asociación entre el peso y el espesor de grasa ($\rho = 0$). Hipótesis alternativa (H_A): Hay asociación entre el peso y el espesor de grasa ($\rho \neq 0$).

Hide

```
corTest <- cor.test(grasacerdos$PV, grasacerdos$EGD, method = "pearson")
corTest
```



```
##
## Pearson's product-moment correlation
##
## data: grasacerdos$PV and grasacerdos$EGD
## t = 1.3916, df = 28, p-value = 0.175
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.1166112 0.5630217
## sample estimates:
## cor
## 0.2543434
```

El resultado del test de correlación de Pearson como se mostró en el punto b corresponde a una correlación positiva baja entre las variables y un P-valor de 0.1749942 que sería mayor que el nivel de significancia $\alpha = 0,05$ de la prueba, por tal razón, no se puede afirmar la presencia de una asociación significativa entre las variables.

Ejercicio 1.2. Los datos del cuarteto de Anscombe se encuentran en el archivo anscombe.xlsx

Se pide explorar los datos de la siguiente manera:

(a)

Graficar los cuatro pares de datos en un diagrama de dispersión cada uno.

Hide

```
# se observa que el archivo esta incompleto anscombe.xlsx (dimensiones 6x8), se busca en internet y se trabaja con Anscombe's Quartet.xlsx (dimensiones 12x8)
anscombe<-read_excel("C:/Users/Josvaldes/Documents/Maestria/Austral/1ano/regresionAvanzada/TPRegresion/TPRegresion/Anscombe's Quartet.xlsx")
dim(anscombe)
```

```
## [1] 11 8
```

Hide

```
head(anscombe)
```

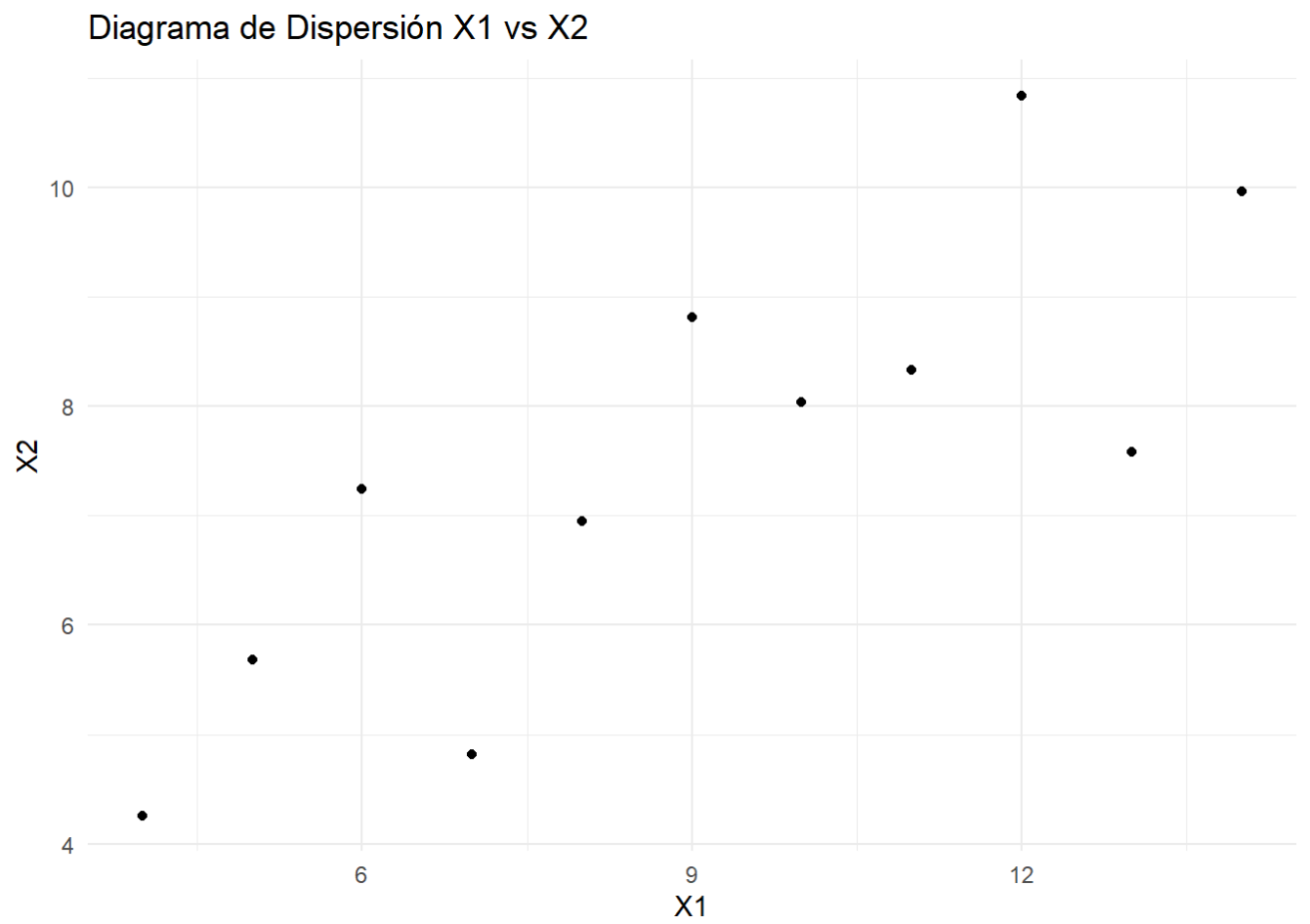
X1 <dbl>	X2 <dbl>	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <dbl>
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84

X1 <dbl>	X2 <dbl>	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <dbl>
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04

6 rows

Hide

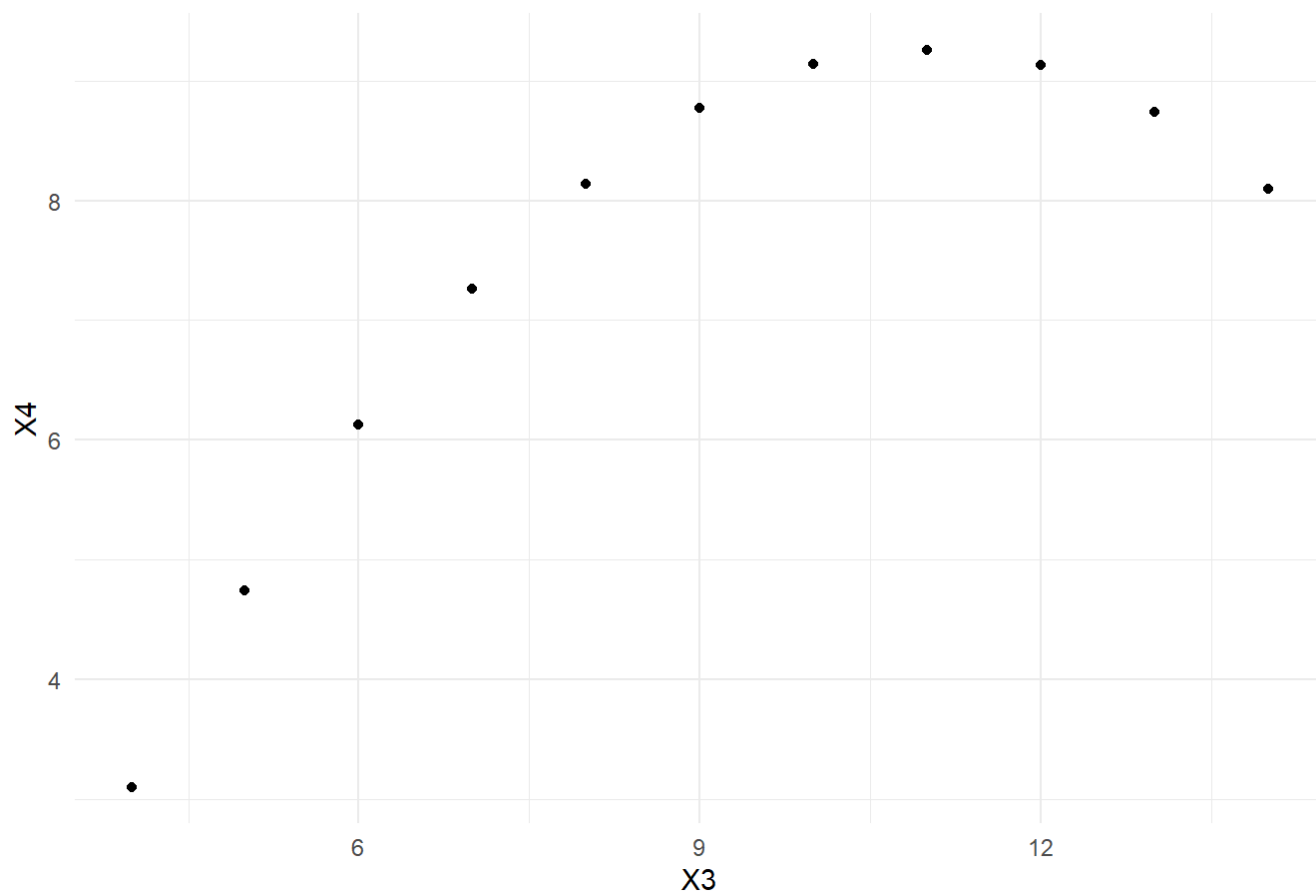
```
dd1=ggplot(anscombe, aes(X1, X2)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n X1 vs X2")  
dd1
```



Hide

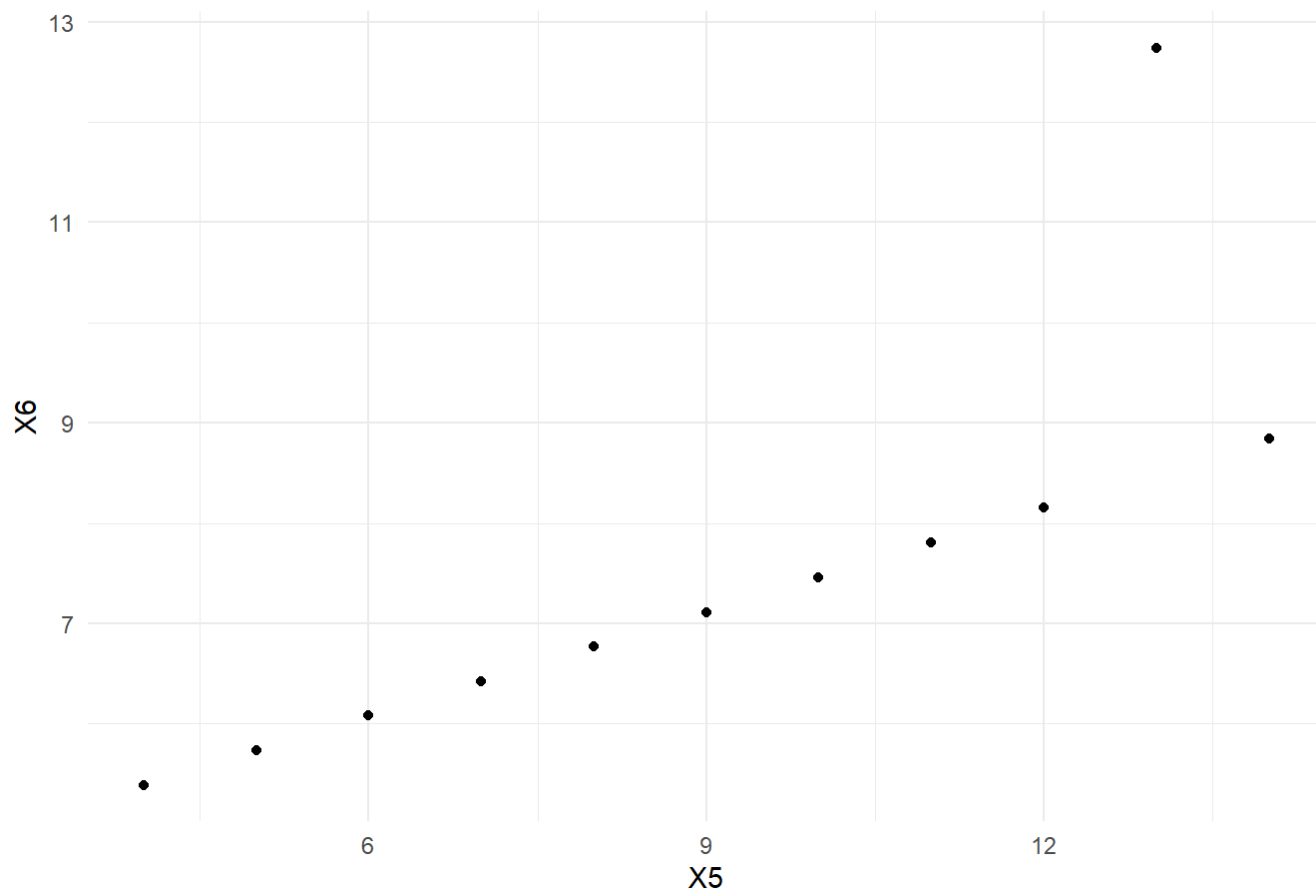
```
dd2=ggplot(anscombe, aes(X3, X4)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n X3 vs X4")  
dd2
```

Diagrama de Dispersión X3 vs X4

[Hide](#)

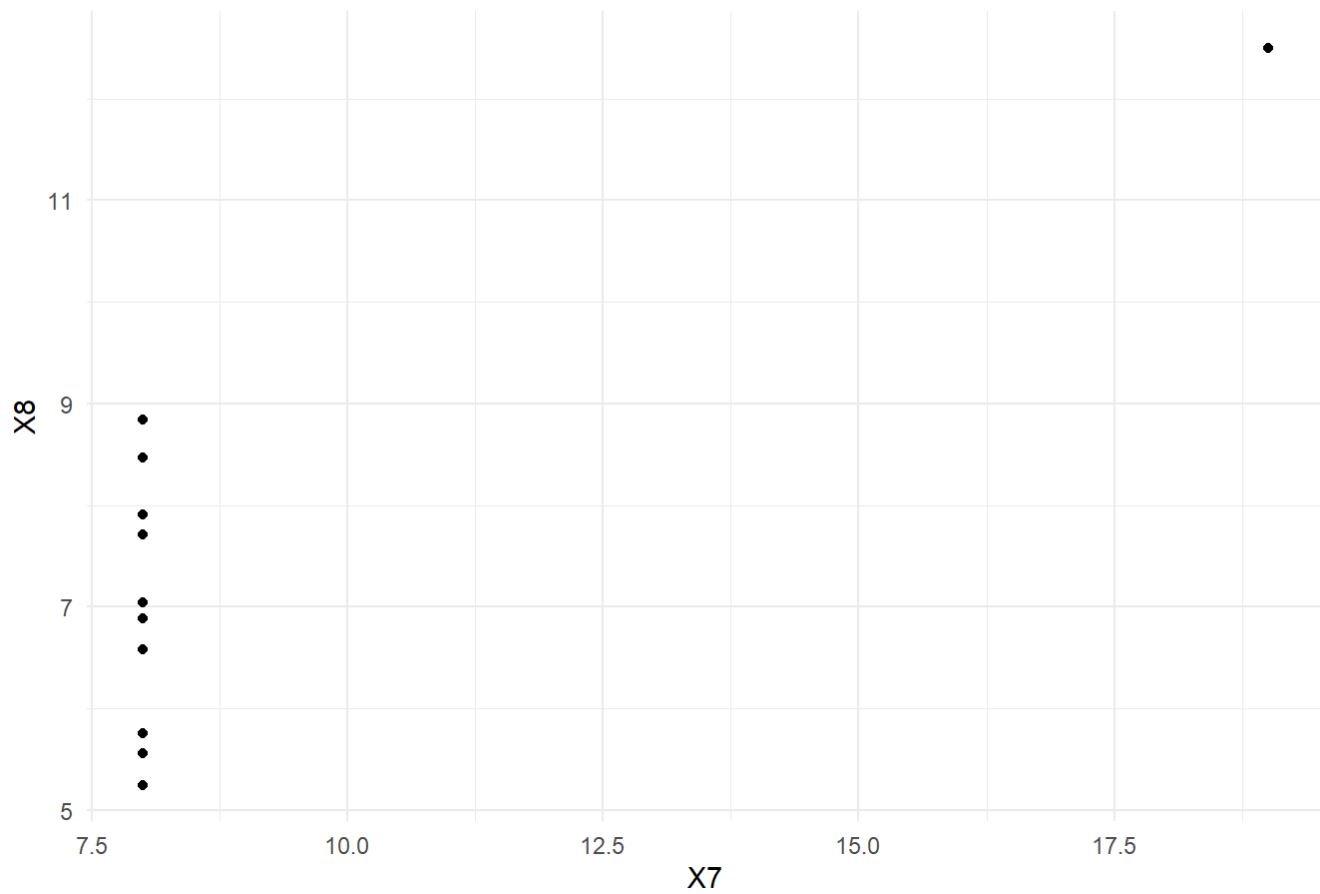
```
dd3=ggplot(anscombe, aes(X5, X6)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n X5 vs X6")  
dd3
```

Diagrama de Dispersión X5 vs X6

[Hide](#)

```
dd4=ggplot(anscombe, aes(X7, X8)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n X7 vs X8")  
dd4
```

Diagrama de Dispersión X7 vs X8

[Hide](#)

```
#resumen  
grid.arrange(dd1,dd2,dd3,dd4, ncol = 2, nrow = 2)
```

Diagrama de Dispersión X1 vs X2

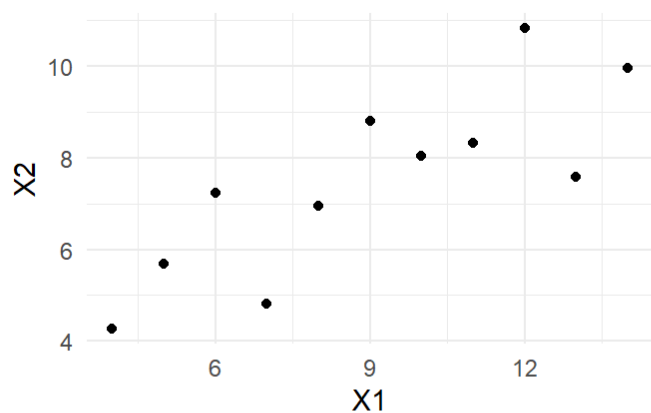


Diagrama de Dispersión X3 vs X4

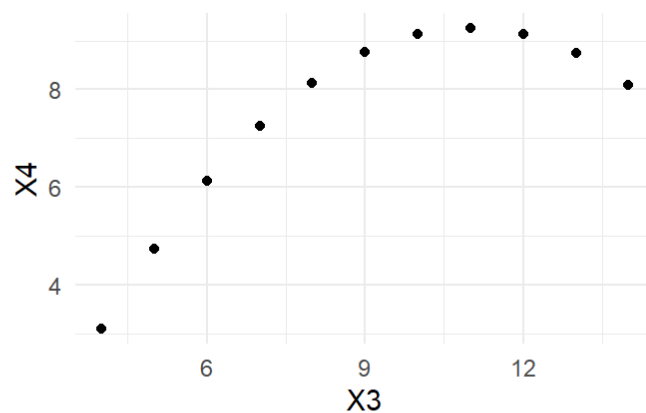


Diagrama de Dispersión X5 vs X6

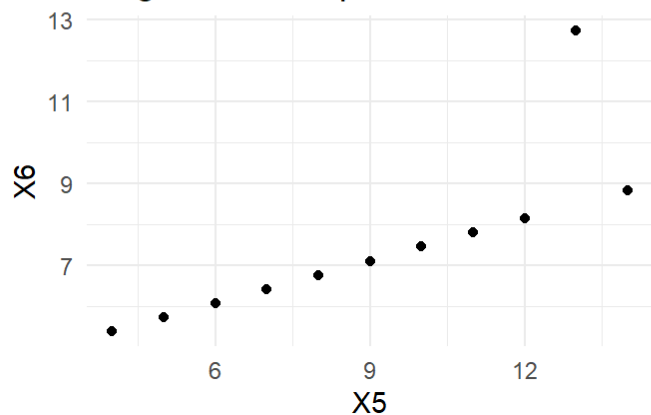
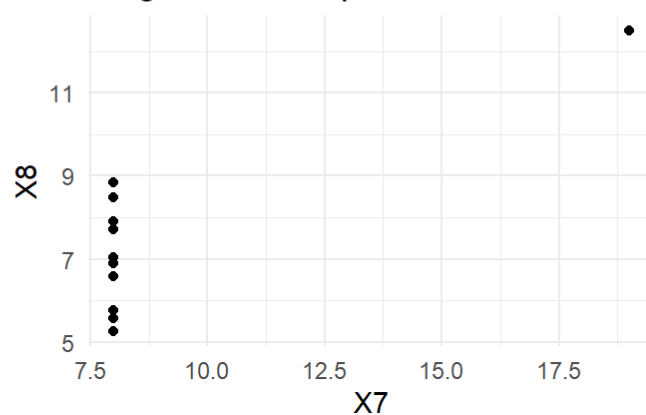


Diagrama de Dispersión X7 vs X8



(b)

Hallar los valores medios de las variables para cada par de datos.

Hide

colMeans(anscombe)

```
##      X1      X2      X3      X4      X5      X6      X7      X8
## 9.000000 7.500909 9.000000 7.500909 9.000000 7.500000 9.000000 7.500909
```

(c)

Hallar los valores de la dispersión para cada conjunto de datos.

Hide

sapply(anscombe, sd)

```
##      X1      X2      X3      X4      X5      X6      X7      X8
## 3.316625 2.031568 3.316625 2.031657 3.316625 2.030424 3.316625 2.030579
```

(d)

Hallar el coeficiente muestral de correlación lineal en cada caso.

Hide

```
mvn(data = anscombe[c(1,2)], mvnTest = "hz")$multivariateNormality$MVN
```

```
## [1] "YES"
```

Hide

```
mvn(data = anscombe[c(3,4)], mvnTest = "hz")$multivariateNormality$MVN
```

```
## [1] "NO"
```

Hide

```
mvn(data = anscombe[c(5,6)], mvnTest = "hz")$multivariateNormality$MVN
```

```
## [1] "NO"
```

Hide

```
mvn(data = anscombe[c(7,8)], mvnTest = "hz")$multivariateNormality$MVN
```

```
## [1] "NO"
```

Hide

```
cor.test(anscombe$X1, anscombe$X2, method="pearson")$p.value
```

```
## [1] 0.002169629
```

Hide

```
cor.test(anscombe$X3, anscombe$X4, method="spearman")$p.value
```

```
## [1] 0.02305887
```

Hide

```
cor.test(anscombe$X5, anscombe$X6, method="spearman")$p.value
```

```
## [1] 0
```

Hide

```
cor.test(anscombe$X7,anscombe$X8,method="spearman")$p.value
```

```
## Warning in cor.test.default(anscombe$X7, anscombe$X8, method = "spearman"):  
## Cannot compute exact p-value with ties
```

```
## [1] 0.1173068
```

Debido al Warning obtenido (Cannot compute exact p-value with ties[1] 0.1173068), se calcula el coeficiente de correlación con el método de Spearman, aun así que el test de Henze-Zirkler dice como resultado NO.

[Hide](#)

```
cor.test(anscombe$X7,anscombe$X8,method="pearson")$p.value
```

```
## [1] 0.002164602
```

(e)

Observar, comentar y concluir.

Por los resultados obtenidos en el primer par de variables se utiliza el coeficiente de correlación de Pearson y para los tres pares restantes el de Spearman. Aunque para la relación de variables X7 y X8 aunque se obtuvo con test de Henze-Zirkler como resultado NO, se recibe una warning por el cual se hace la prueba con el test de Pearson.

1.2. Modelo Lineal Simple

Ejercicio 1.3. El archivo peso_edad_colest.xlsx disponible contiene registros correspondientes a 25 individuos respecto de su peso, su edad y el nivel de colesterol total en sangre.

Se pide:

(a)

Realizar el diagrama de dispersión de colesterol en función de la edad y de colesterol en función de peso. Le parece adecuado ajustar un modelo lineal para alguno de estos dos pares de variables?

[Hide](#)


```
#Se cargan los datos
colesterol <- read_excel('peso_edad_colest.xlsx')

#Se visualizan la estructura
head(colesterol)
```

peso <dbl>	edad <dbl>	colest <dbl>
84	46	354
73	20	190
65	52	405
70	30	263
76	57	451
69	25	302

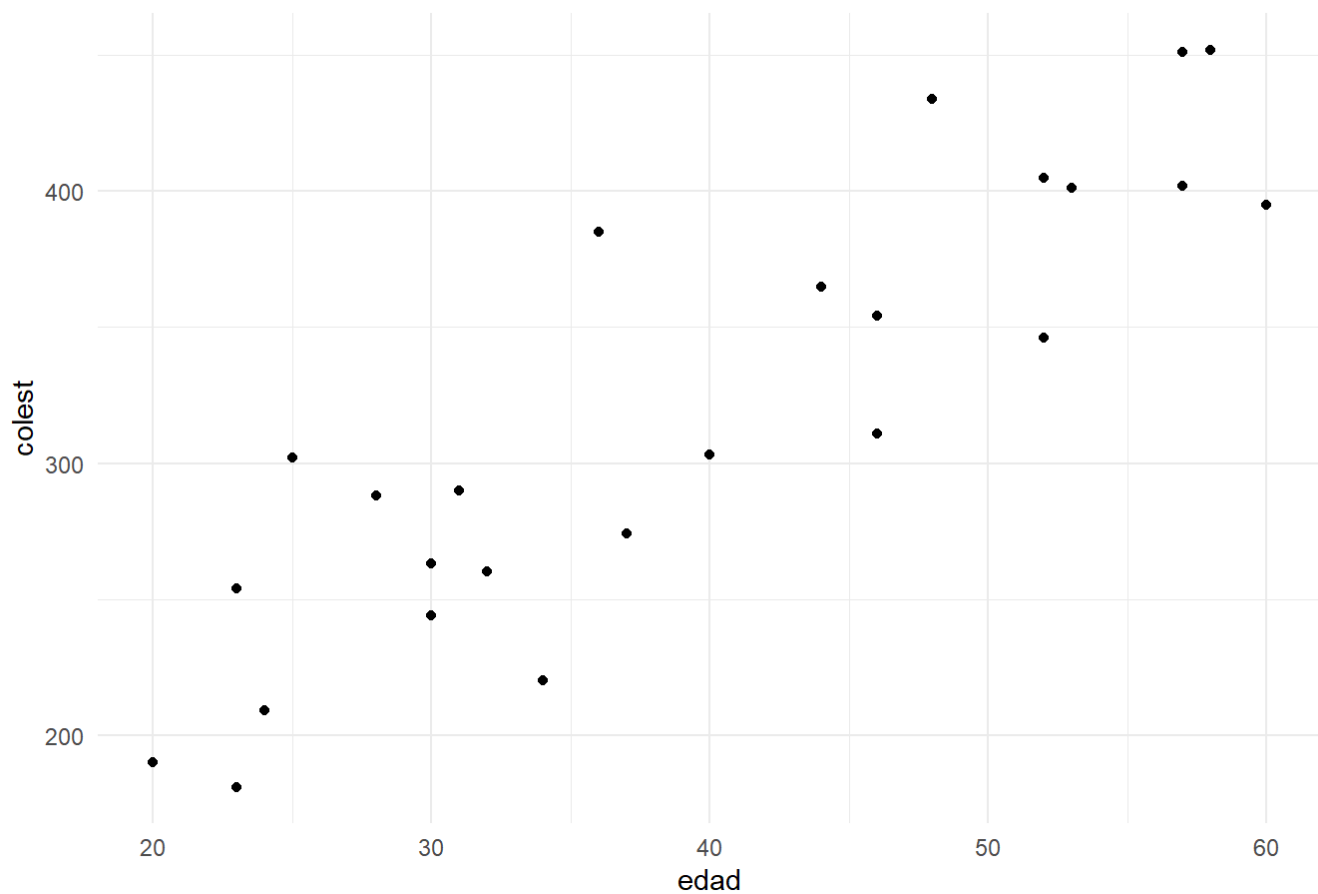
6 rows

Se realizan los diagramas de dispersión solicitados

[Hide](#)

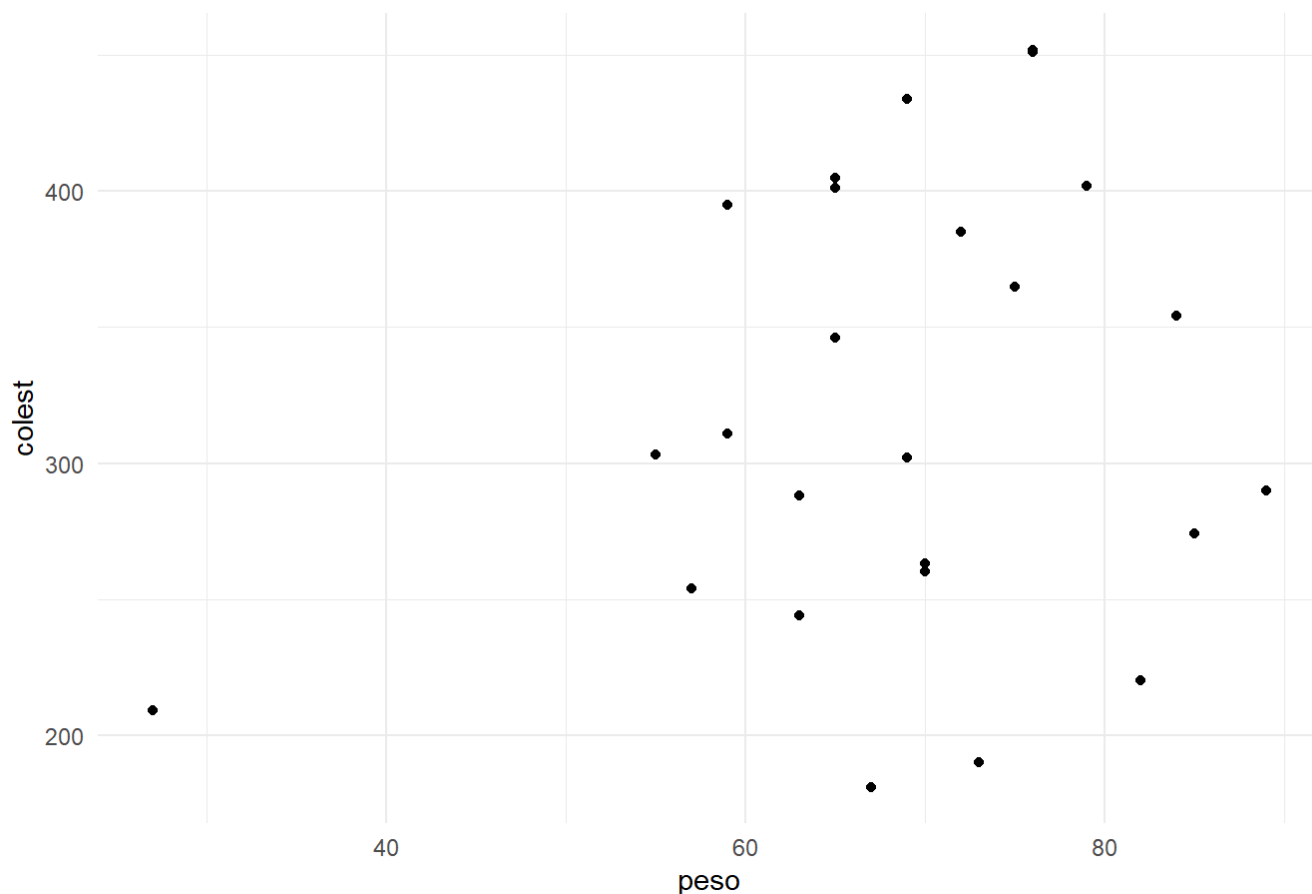
```
#Diagrama de dispersión colesterol en función de la edad
dd112=ggplot(colesterol, aes(edad, colest)) +
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n edad vs coles
terol")
dd112
```

Diagrama de Dispersión edad vs colesterol

[Hide](#)

```
#Diagrama de dispersión colesterol en función del peso
dd212=ggplot(colesterol, aes(peso, colest)) +
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n peso vs coles
terol")
dd212
```

Diagrama de Dispersión peso vs colesterol



Por las gráficas se podría pensar que se ajuste un modelo lineal entre las variables edad y colesterol.

(b)

Estime los coeficientes del modelo lineal para el colesterol en función de la edad.

Coeficientes

Hide

```
#Modelo lineal para el colesterol en función de la edad.
model <- lm(colest ~ edad, data = colesterol)
model$coefficients
```

```
## (Intercept)      edad
##  95.502004    5.670842
```

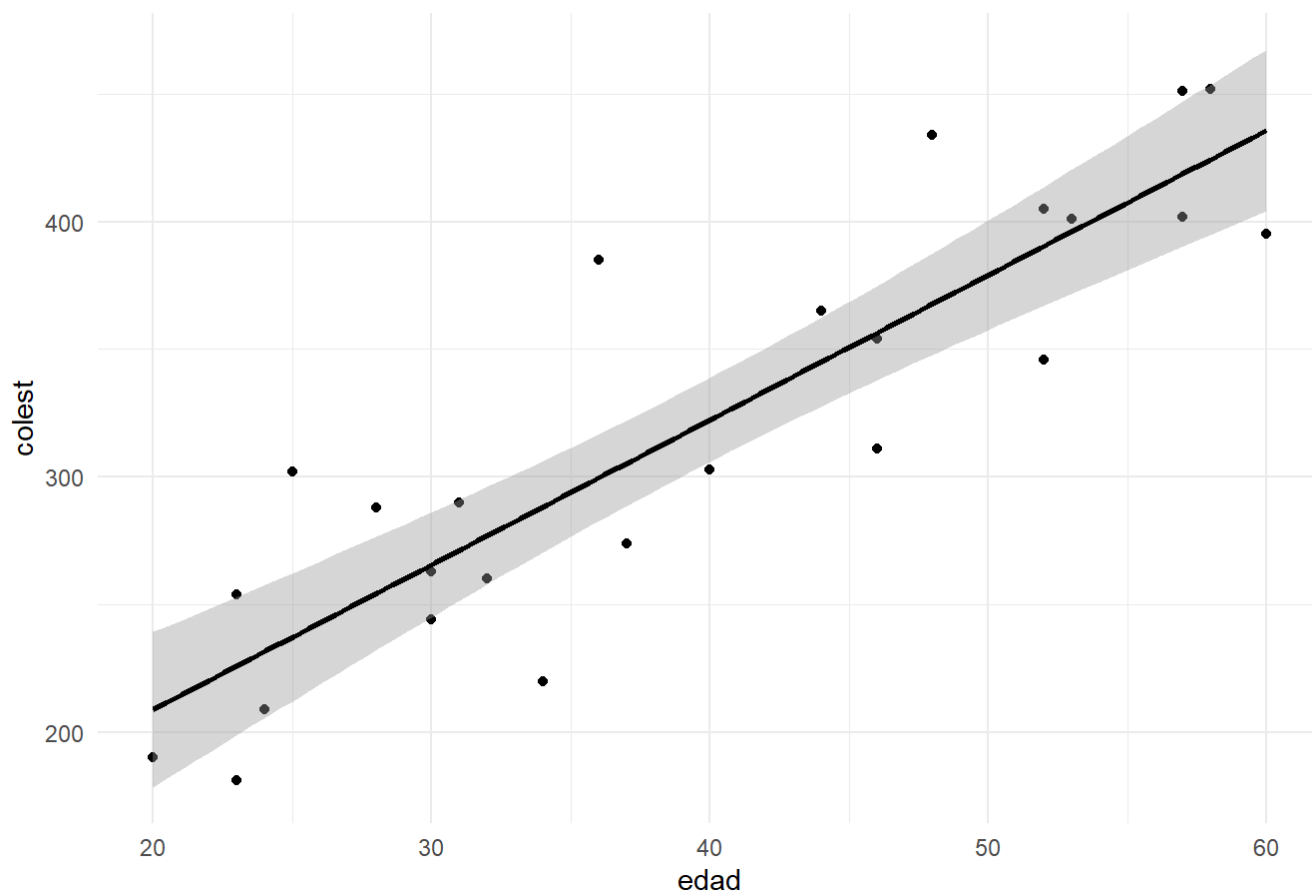
Grafica del modelo y las bandas de error estándar alrededor de la línea de regresión

Hide

```
(dd112+ geom_smooth(method = "lm", se = TRUE, color = "black") )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Diagrama de Dispersión edad vs colesterol



(c)

Estime intervalos de confianza del 95% para los coeficientes del modelo y compare estos resultados con el test de Wald para los coeficientes. Le parece que hay asociación entre estos test y el test de la regresión?

Hide

```
ic <- confint(model, level = 0.95)
ic
```

```
##           2.5 %    97.5 %
## (Intercept) 41.190390 149.813618
## edad        4.358216  6.983467
```

Test de Wald

Hide

```
library(aod)
coef(model)
```

```
## (Intercept)      edad
##   95.502004    5.670842
```

Hide

```
testWald=wald.test(Sigma = vcov(model), b = coef(model), Terms = 1)
testWald
```

```
## Wald test:
## -----
##
## Chi-squared test:
## X2 = 13.2, df = 1, P(> X2) = 0.00028
```

Las anteriores salidas muestra los coeficientes estimados del modelo de regresión lineal y los resultados del test de Wald para evaluar la significancia de los coeficientes.

Los coeficientes del modelo indican lo siguiente:

- El coeficiente de intercepto (Intercept) es de aproximadamente 95.502004.
- El coeficiente para la variable "edad" es de aproximadamente 5.670842.

El test de Wald se utiliza para evaluar la significancia estadística de los coeficientes del modelo. En este caso, se realiza el test de Wald para el coeficiente del intercepto (intercept). El resultado del test muestra que el estadístico de prueba chi-cuadrado (X^2) es de 13.2, con 1 grado de libertad y un valor p ($P(>X^2)$) de 0.00028.

Se puede concluir lo siguiente:

El coeficiente de intercepto es significativamente diferente de cero, debido a que el valor p es muy pequeño (0.00028). Esto indica que hay evidencia de una asociación entre la variable de respuesta y la variable de intercepto.

En cuanto al coeficiente de la variable "edad", se realizan los siguientes cálculos para obtener el test de Wald:

Hide

```
# Se obtiene la matriz de varianza-covarianza de los coeficientes del modelo
vcov_model <- vcov(model)

# Se obtienen los coeficientes estimados del modelo
coef_model <- coef(model)

# Calculo del estadístico de prueba utilizando la fórmula del test de Wald:
wald_stat <- (coef_model["edad"] - 0) / sqrt(vcov_model["edad", "edad"])

# Calculo del valor p correspondiente al estadístico de prueba
p_value <- 1 - pchisq(wald_stat^2, df = 1)

# Imprimir resultado
cat("Test de Wald para la variable 'edad':\n")
```

```
## Test de Wald para la variable 'edad':
```

Hide

```
cat("-----\n")
```

```
## -----
```

Hide

```
cat("Estadístico de prueba:", wald_stat, "\n")
```

```
## Estadístico de prueba: 8.937073
```

Hide

```
cat("Valor p:", p_value, "\n")
```

```
## Valor p: 0
```

En resumen, hay evidencia de asociación entre el coeficiente de intercepto y la variable de respuesta según el test de Wald. Para la variable “edad” se tiene un estadístico de prueba de 8.937073 y un valor p de 0. Esto indica que hay evidencia significativa para rechazar la hipótesis nula de que el coeficiente de “edad” sea igual a cero.

(d)

A partir de esta recta estime los valores de $E(Y)$ para $x = 25$ años y $x = 48$ años. Podría estimarse el valor de $E(Y)$ para $x = 80$ años?

Para estimar los valores de $E(Y)$ para diferentes valores de x utilizando la recta ajustada en el modelo de regresión, se pueden utilizar los coeficientes del modelo.

En este caso, los coeficientes del modelo son:

Intercepto: 95.502004 Coeficiente para la variable “edad”: 5.670842

$E(Y) = \text{Intercepto} + \text{Coeficiente} * x$

Hide

```
predict(model, newdata = data.frame(edad = c(25,80)))
```

```
##          1          2
## 237.2730 549.1693
```

Sin embargo, para valores de x más allá del rango de los datos observados, como $x = 80$ años, la extrapolación puede no ser confiable. La recta ajustada se basa en los datos observados y su validez puede estar limitada a ese rango. Por lo tanto, no se recomienda estimar el valor de $E(Y)$ para $x = 80$ años utilizando este modelo de regresión.

(e)

Testee la normalidad de los residuos y haga un gráfico para ver si son homocedásticos.

Hide

```
# Prueba de normalidad de Shapiro-Wilk  
residuos <- residuals(model)  
shapiro.test(residuos)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  residuos  
## W = 0.96478, p-value = 0.5175
```

El resultado de esta prueba proporciona un valor p que indica que no hay suficiente evidencia para rechazar la hipótesis nula de normalidad de los residuos. Como el valor p es mayor que un umbral de significancia (por ejemplo, 0.05), se puede concluir que los residuos siguen una distribución normal.

Grafico de los residuos del modelo

Hide

```
plot(residuos ~ fitted.values(model), ylab = "Residuos", xlab = "Valores ajustados")  
abline(h = 0, col = "red")
```

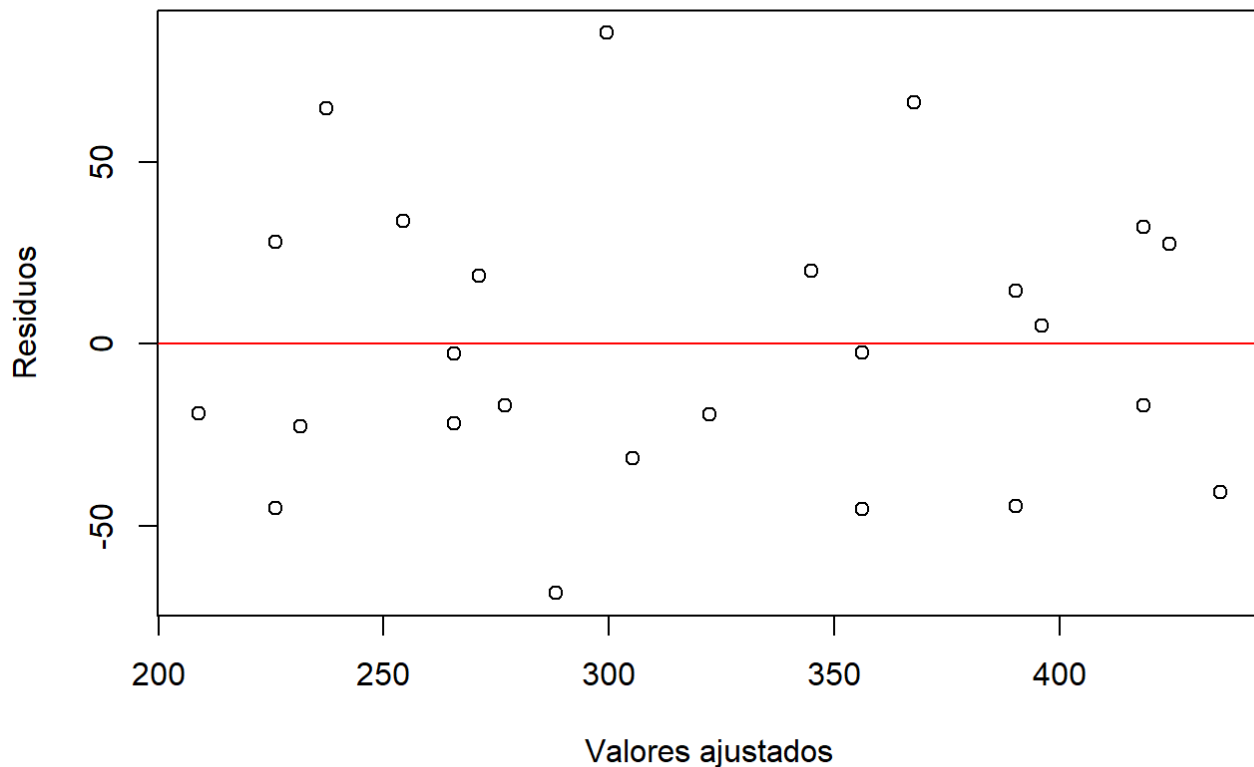


Grafico con lineas:

Hide

```

colest2<-colesterol
colest2$prediccion <- model$fitted.values
colest2$residuos <- model$residuals

ggplot(data = colest2, aes(x = prediccion, y = residuos)) +
  geom_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_hline(yintercept = 0) + geom_segment(aes(xend = prediccion, yend = 0), alpha = 0.2)
+
  labs(title = "Distribución de los residuos", x = "predicción modelo", y = "residuo") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

```

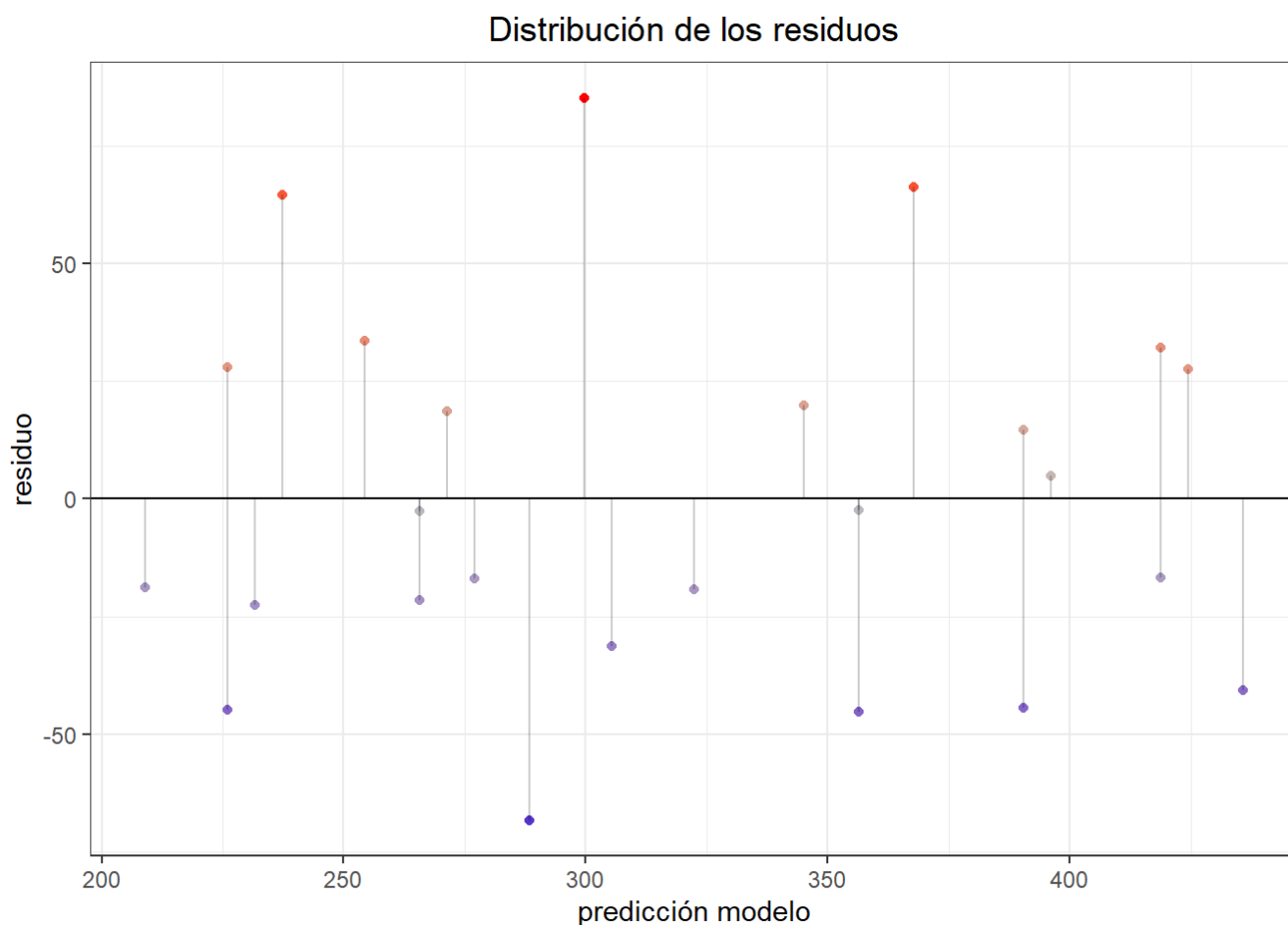


Grafico con histograma:

Hide

```

ggplot(data = colest2, aes(x = residuos)) + geom_histogram(aes(y = after_stat(density))) +
  labs(title = "Histograma de los residuos") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Histograma de los residuos

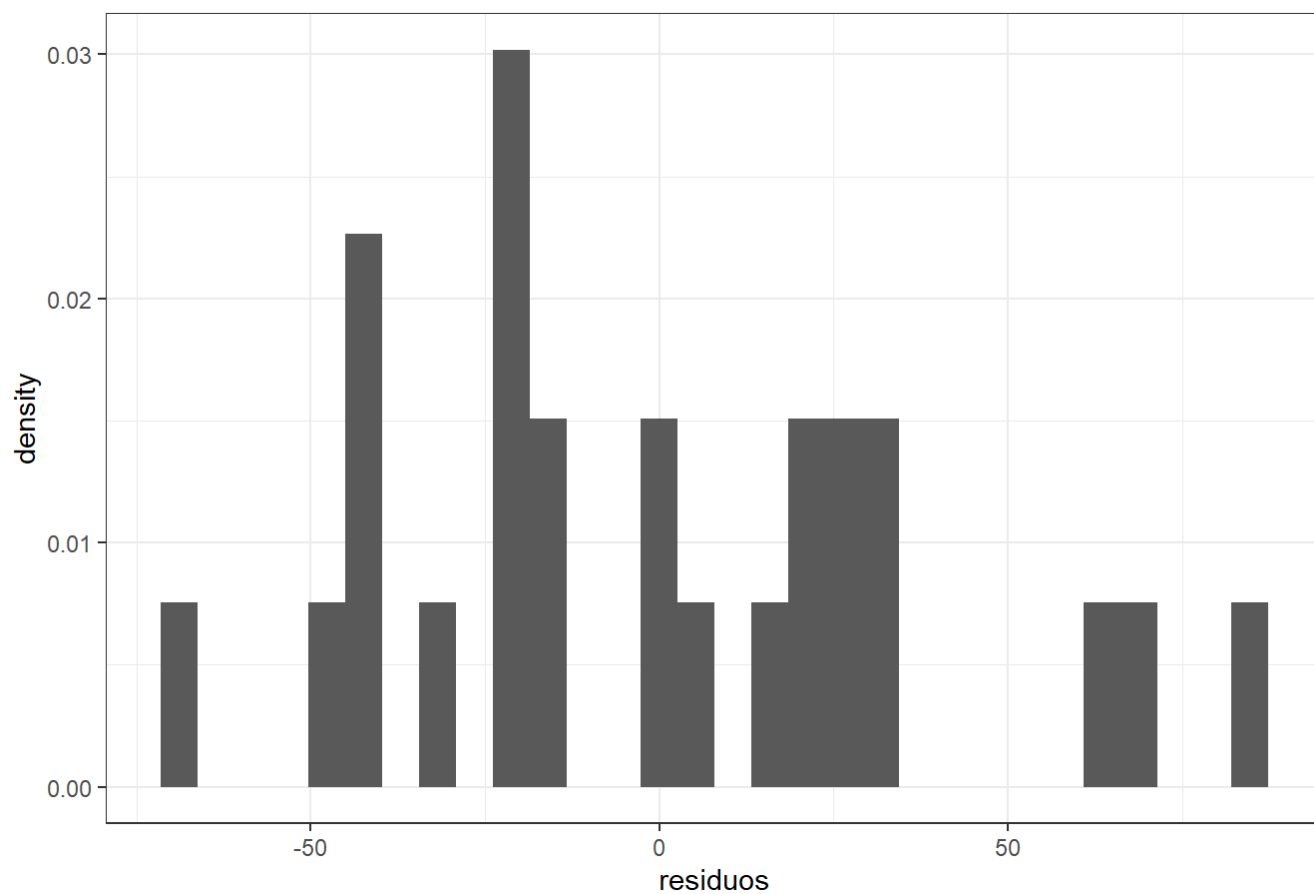
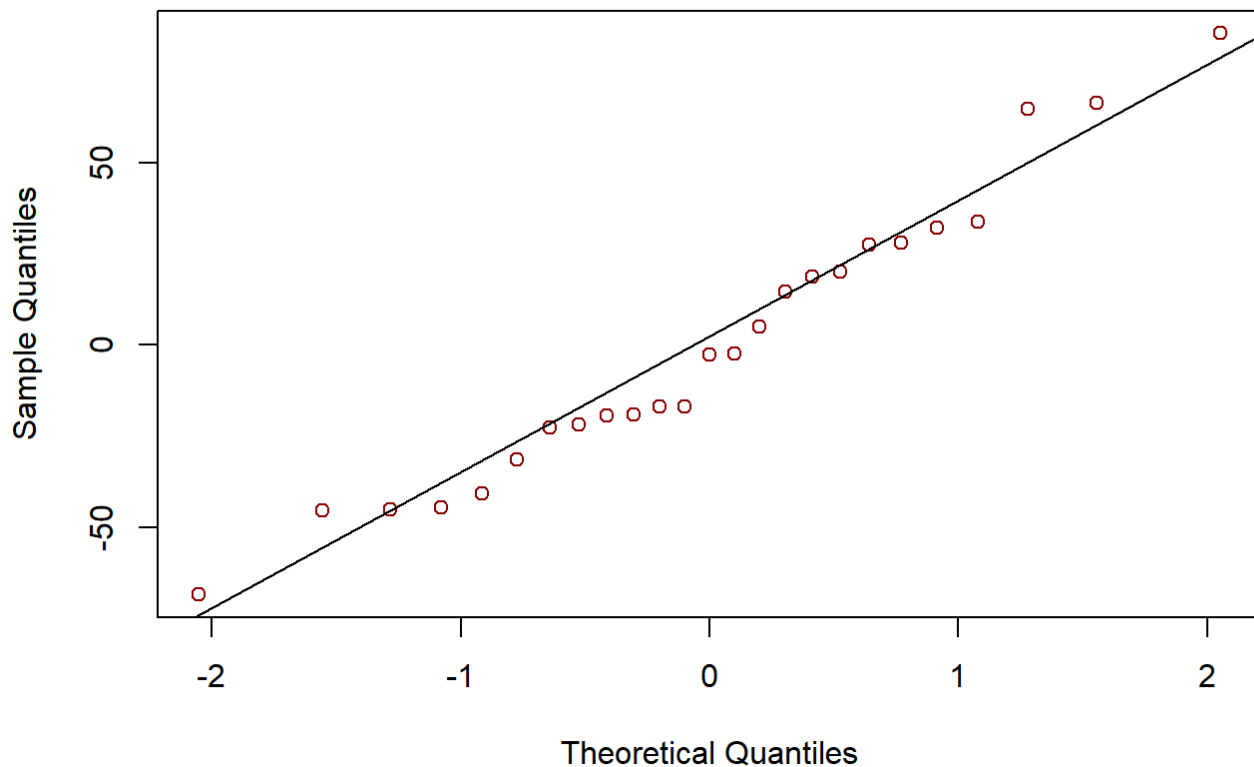


Grafico QQ

[Hide](#)

```
qqnorm(model$residuals, main = "Residuos del modelo", col = "darkred")  
qqline(model$residuals)
```

Residuos del modelo



De los resultados anteriores se puede suponer que los residuos del modelo siguen una distribución normal y no son homocedasticos.

1.3. Transformación de Variables

Ejercicio 1.4.

Una empresa desarrolló un sistema de energía solar para calentar el agua para una caldera que es parte del sistema de energía del proceso productivo. Existe el interés de controlar la estabilidad del sistema, para ello se monitorea el mismo y se registran los datos cada hora. Los datos se encuentran disponibles en el archivo energia.xlsx

(a)

Realizar el diagrama de dispersión y evaluar si un modelo de regresión lineal es adecuado.

Hide

```
# Se cargan los datos
energia <- read_excel('energia.xlsx')

#Se visualizan la estructura
head(energia)
```

Hora <dbl>	Energía <dbl>
1	598
2	527
3	530
4	528
5	452
6	497
6 rows	

[Hide](#)

```
#Dimensiones  
dim(energia)
```

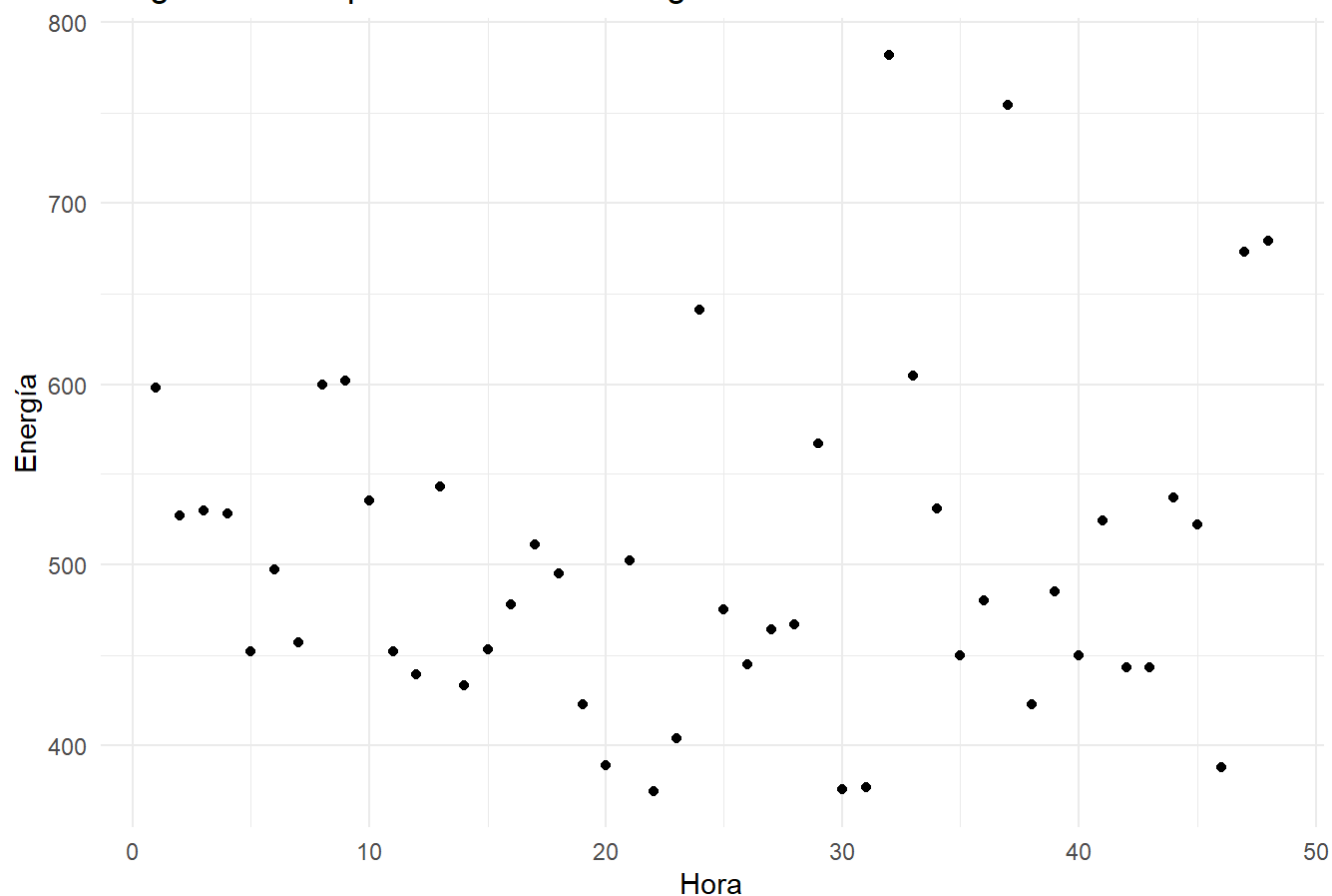
```
## [1] 48  2
```

Diagrama de dispersión

[Hide](#)

```
#Diagrama de dispersión colesterol en función del peso  
dd14=ggplot(energia, aes(Hora, Energía)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de dispersi\u00F3n Hora vs Energ  
ía")  
dd14
```

Diagrama de dispersión Hora vs Energía



Hide

```
# Validación de una distribución normal bivariada para estas variables
biNormTest14 <- mvn(energia, mvnTest = "hz")
biNormTest14
```

```
## $multivariateNormality
##           Test      HZ      p value MVN
## 1 Henze-Zirkler 1.355059 0.002347283 NO
##
## $univariateNormality
##           Test Variable Statistic  p value Normality
## 1 Anderson-Darling Hora      0.5128    0.1849    YES
## 2 Anderson-Darling Energía    1.1299    0.0053    NO
##
## $Descriptives
##           n Mean Std.Dev Median Min Max 25th 75th Skew Kurtosis
## Hora      48 24.50 14.00000 24.5 1 48 12.75 36.25 0.000000 -1.2752179
## Energía   48 504.25 93.07615 482.5 375 782 444.50 535.50 1.032494 0.8324672
```

Por arrojar un resultado de MVN NO se realiza el test de Spearman

Hide

```
cor.test(energia$Hora, energia$Energía, method="spearman")$p.value
```

```
## Warning in cor.test.default(energia$Hora, energia$Energía, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
## [1] 0.806419
```

Hide

```
# métodos robustos para manejar empates
cor.test(energia$Hora, energia$Energía, method = "spearman", exact = FALSE)
```

```
##
## Spearman's rank correlation rho
##
## data: energia$Hora and energia$Energía
## S = 19093, p-value = 0.8064
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## -0.03631528
```

La salida corresponde a la prueba de correlación de rangos de Spearman y se puede interpretar de la siguiente manera:

- La primera línea indica que se realizó la prueba de correlación de rangos de Spearman en los datos de las variables "Hora" y "Energía" del dataframe "energia".
- El valor de S es 19093, que es la suma de los cuadrados de las diferencias entre los rangos de las dos variables.
- El valor p es 0.8064, que es el valor p obtenido de la prueba de hipótesis. En este caso, como el valor p es mayor que 0.05 (nivel de significancia comúnmente utilizado), no hay suficiente evidencia para rechazar la hipótesis nula de que no hay correlación entre las dos variables.
- La hipótesis alternativa indica que el verdadero coeficiente de correlación rho no es igual a cero.
- La estimación de rho basada en la muestra es -0.03631528, lo que indica una correlación negativa muy débil entre las dos variables.

En resumen, la salida sugiere que no hay evidencia suficiente para concluir que hay una correlación significativa entre las variables "Hora" y "Energía" en el conjunto de datos analizado.

(b)

Estimar un modelo lineal y verificar la normalidad de los residuos del mismo.

Hide

```
model14 = lm(Hora ~ Energía, data=energia)
summary(model14)
```

```
##
## Call:
## lm(formula = Hora ~ Energía, data = energia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.605 -11.787   0.263  11.301  22.870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.55802   11.33302   1.638   0.108
## Energía      0.01178    0.02211   0.533   0.597
##
## Residual standard error: 14.11 on 46 degrees of freedom
## Multiple R-squared:  0.006138,    Adjusted R-squared:  -0.01547
## F-statistic: 0.2841 on 1 and 46 DF,  p-value: 0.5966
```

El modelo de regresión lineal ajustado es el siguiente:

$$\text{Hora} = 18.55802 + 0.01178 * \text{Energía}$$

Se interpreta:

- **Residuals:** Muestra los valores mínimos, primer cuartil, mediana, tercer cuartil y máximo de los residuos del modelo. Los residuos representan la diferencia entre los valores observados de la variable "Hora" y los valores predichos por el modelo.
- **Coefficients:** Muestra los coeficientes estimados del modelo. El intercepto (Intercept) tiene un valor estimado de 18.55802 y un error estándar de 11.33302. La variable explicativa "Energía" tiene un coeficiente estimado de 0.01178 y un error estándar de 0.02211.
- **Residual standard error:** Indica la desviación estándar de los residuos, que es una medida de la variabilidad no explicada por el modelo. En este caso, el valor es de 14.11.
- **El R-cuadrado múltiple** indica la proporción de la variabilidad total de la variable "Hora" explicada por el modelo, y en este caso es de 0.006138. El R-cuadrado ajustado tiene en cuenta el número de variables explicativas y los grados de libertad, y en este caso es de -0.01547. Ambos valores son bajos, lo que sugiere que el modelo tiene una capacidad limitada para explicar la variabilidad en la variable respuesta.
- **F-statistic y p-value:** El F-statistic evalúa si el modelo en su conjunto es significativo. En este caso, el valor del estadístico F es de 0.2841 con 1 y 46 grados de libertad. El p-value asociado al F-statistic es de 0.5966, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula de que los coeficientes del modelo sean igual a cero.

El modelo de regresión lineal ajustado no parece ser estadísticamente significativo, ya que el p-value asociado al F-statistic es alto y los coeficientes de las variables no son significativamente diferentes de cero. Además, el R-cuadrado es muy bajo, lo que sugiere que el modelo no explica de manera efectiva la variabilidad en la variable "Hora" basándose únicamente en la variable "Energía".

(c)

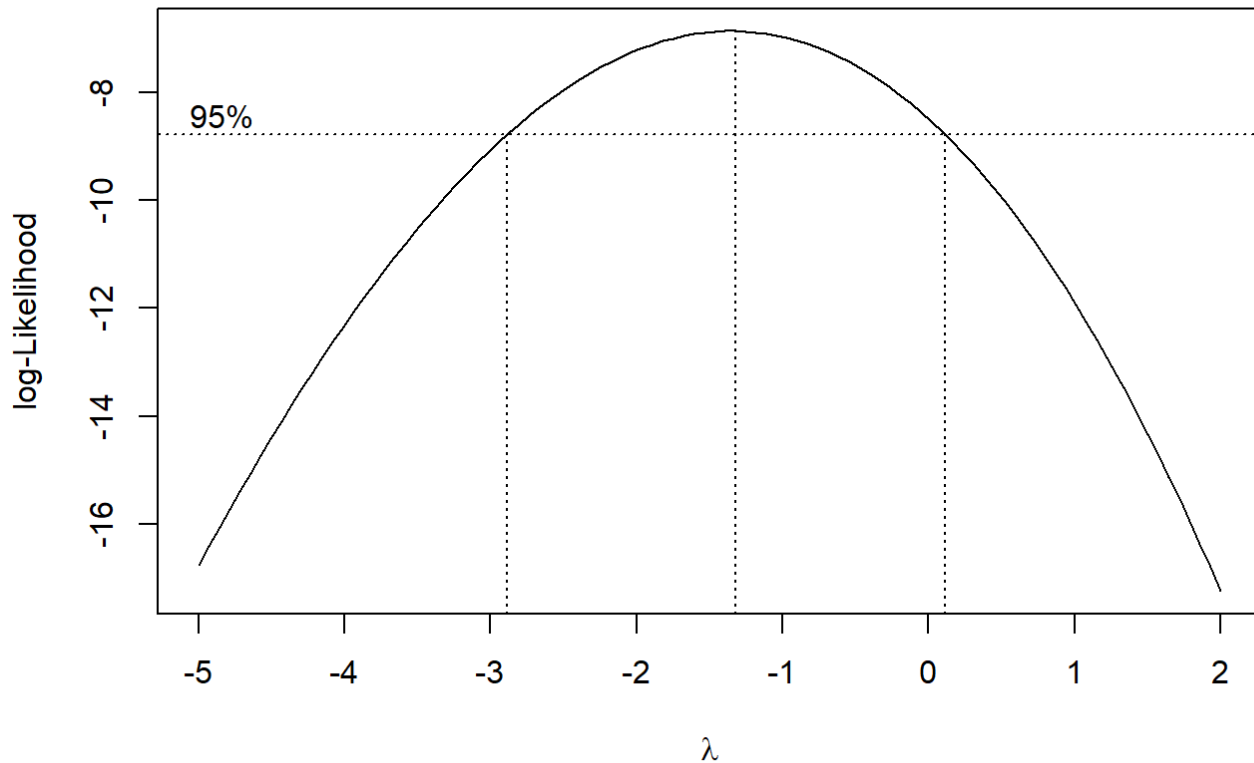
En caso de rechazar este supuesto buscar una transformación lineal para este modelo y aplicarla.

Hide

```
library(MASS)
```

```
# Aplica la transformación de Box-Cox a La variable dependiente "Energía" en función de La variable independiente "Hora"
```

```
box_cox_result <- boxcox(Energía ~ Hora, lambda = -5:2, data = energia)
```



Hide

```
# Se encuentra el valor óptimo de lambda que maximiza el logaritmo de verosimilitud
```

```
best_box_cox <- box_cox_result$x[which.max(box_cox_result$y)]
```

```
# Se ajusta un modelo de regresión lineal utilizando La variable dependiente "Energía" elevada a la potencia óptima de lambda (best_box_cox) como la variable de respuesta y la variable independiente "Hora".
```

```
modelE2 <- lm((Energía)^(best_box_cox) ~ Hora, data = energia)
```

```
summary(modelE2)
```

```
##
## Call:
## lm(formula = (Energía)^(best_box_cox) ~ Hora, data = energia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.290e-04 -3.263e-05  3.849e-06  3.599e-05  1.150e-04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.779e-04  1.787e-05   15.55  <2e-16 ***
## Hora        -1.251e-08  6.350e-07   -0.02    0.984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.094e-05 on 46 degrees of freedom
## Multiple R-squared:  8.444e-06, Adjusted R-squared:  -0.02173
## F-statistic: 0.0003884 on 1 and 46 DF,  p-value: 0.9844
```

Hide

```
shapiro.test(modelE2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelE2$residuals
## W = 0.98002, p-value = 0.5796
```

Según el gráfico, el lambda óptimo se encuentra cerca de -1. Entonces consideraremos la transformación logarítmica sobre la variable respuesta.

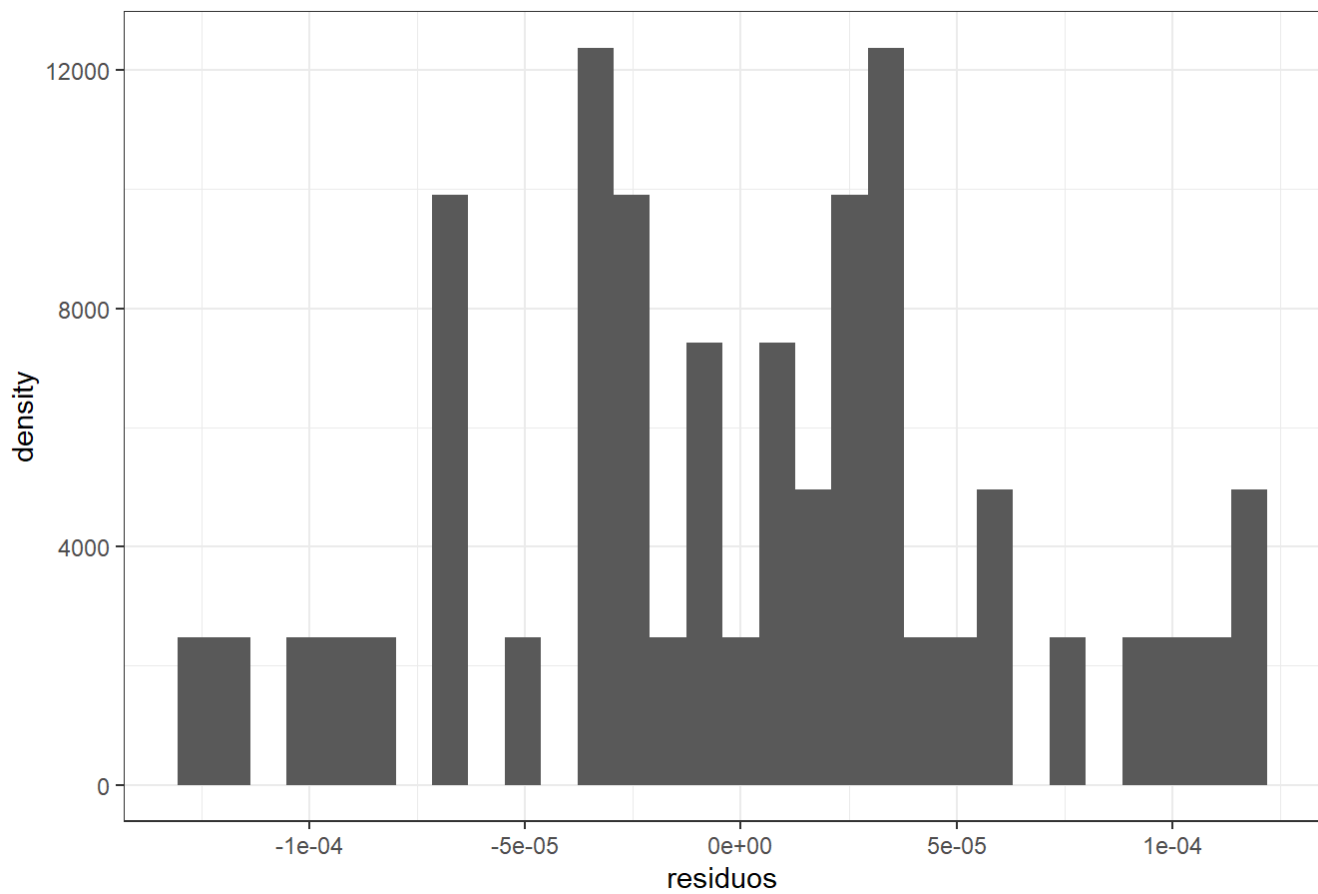
Hide

```
energia3<-energia
energia3$Energía <- log(energia$Energía)
energia3$prediccion <- modelE2$fitted.values
energia3$residuos <- modelE2$residuals
ggplot(data = energia3, aes(x = residuos)) + geom_histogram(aes(y = ..density..)) +
  labs(title = "Histograma de los residuos") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

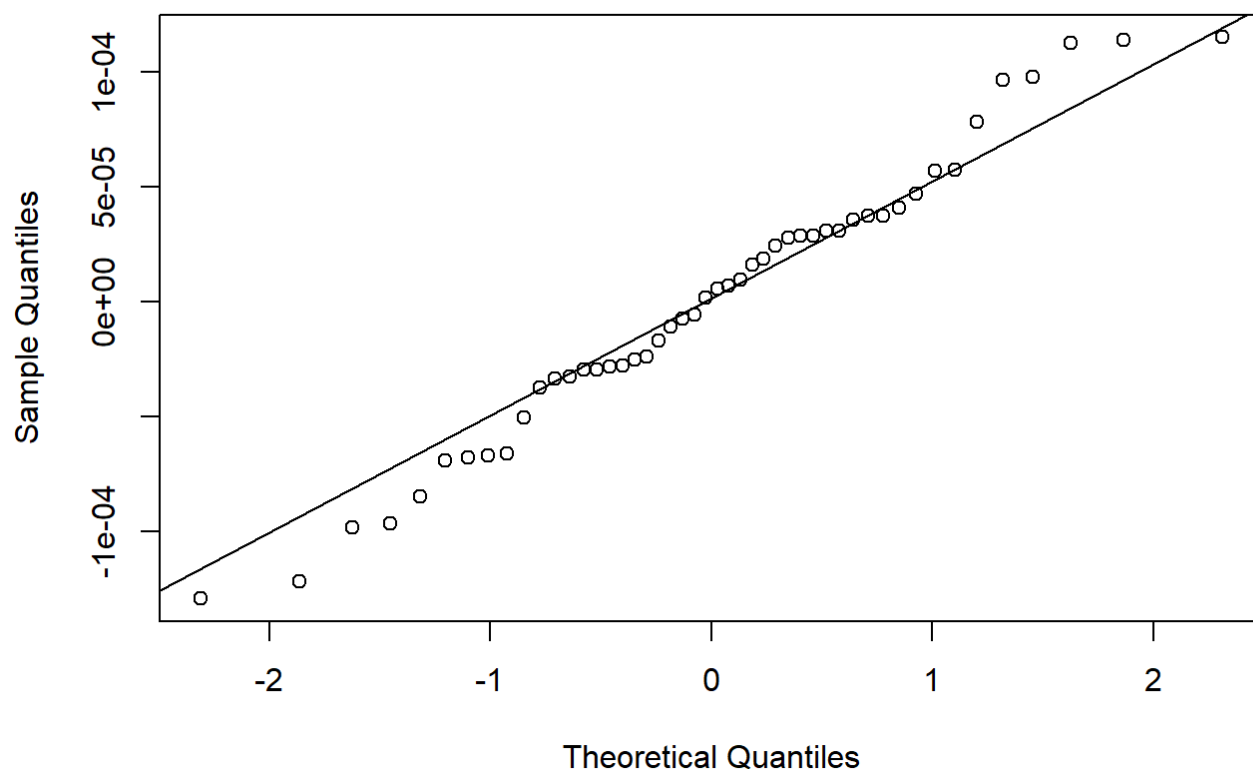
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Histograma de los residuos

[Hide](#)

```
qqnorm(modelE2$residuals)  
qqline(modelE2$residuals)
```

Normal Q-Q Plot


[Hide](#)

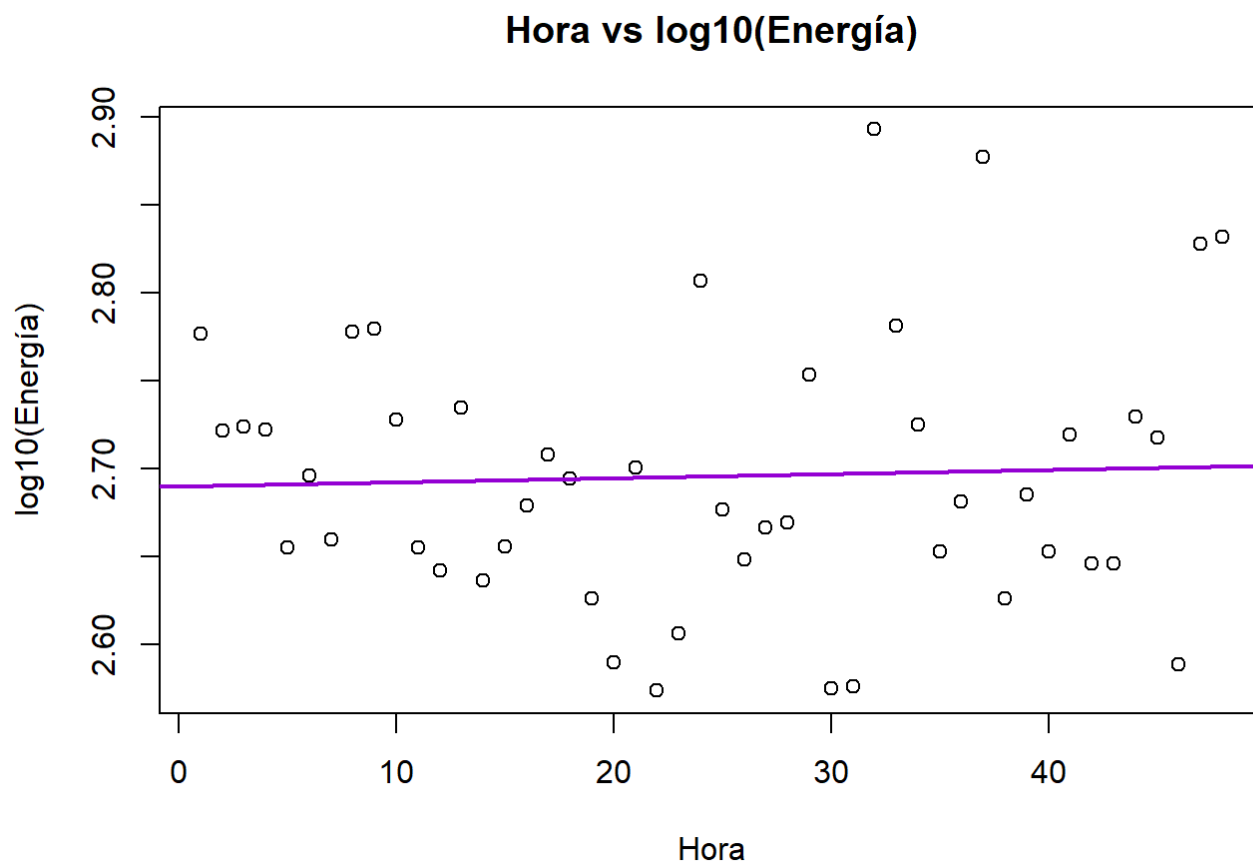
```
linMod2 <- lm(log10(Energía) ~ Hora, data = energia)
summary(linMod2)
```

```
##
## Call:
## lm(formula = log10(Energía) ~ Hora, data = energia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12212 -0.04859 -0.01411  0.03415  0.19541
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6899875   0.0224064  120.055   <2e-16 ***
## Hora         0.0002440   0.0007961   0.306     0.761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07641 on 46 degrees of freedom
## Multiple R-squared:  0.002038,    Adjusted R-squared:  -0.01966
## F-statistic: 0.09393 on 1 and 46 DF,  p-value: 0.7606
```

[Hide](#)

```
plot(energia$Hora, log10(energia$Energía), xlab="Hora", ylab="log10(Energía)",
     main="Hora vs log10(Energía)")

abline(linMod2, col="darkviolet", lwd=2)
```



análisis diagnóstico

[Hide](#)

```
shapiro.test(linMod2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  linMod2$residuals
## W = 0.96393, p-value = 0.1454
```

[Hide](#)

```
library(car)

# Prueba de heterocedasticidad
ncvTest(modelE2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.758408, Df = 1, p = 0.096744
```

[Hide](#)

```
dwt(linMod2)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.0159792 1.877106 0.502
## Alternative hypothesis: rho != 0
```

(d)

Realizar el análisis diagnóstico del nuevo modelo y estimar un intervalo de confianza y un intervalo de predicción para 27.5 hs con ambos modelos. Comparar los intervalos.