

### 1.1. Correlación

(a) Dibujar el diagrama de dispersión e interpretarlo.

(b) Calcular el coeficiente de correlación muestral y explíquelos.

(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.

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

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

(c) Hallar los valores de la dispersión para cada conjunto de datos.

(d) Hallar el coeficiente muestral de correlación lineal en cada caso.

(e) Observar, comentar y concluir.

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) 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?

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

(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?

(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?

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

# 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 510867 27.3   1137497 60.8   644245 34.5
## Vcells 906966  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")

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

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

[Hide](#)

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

## 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

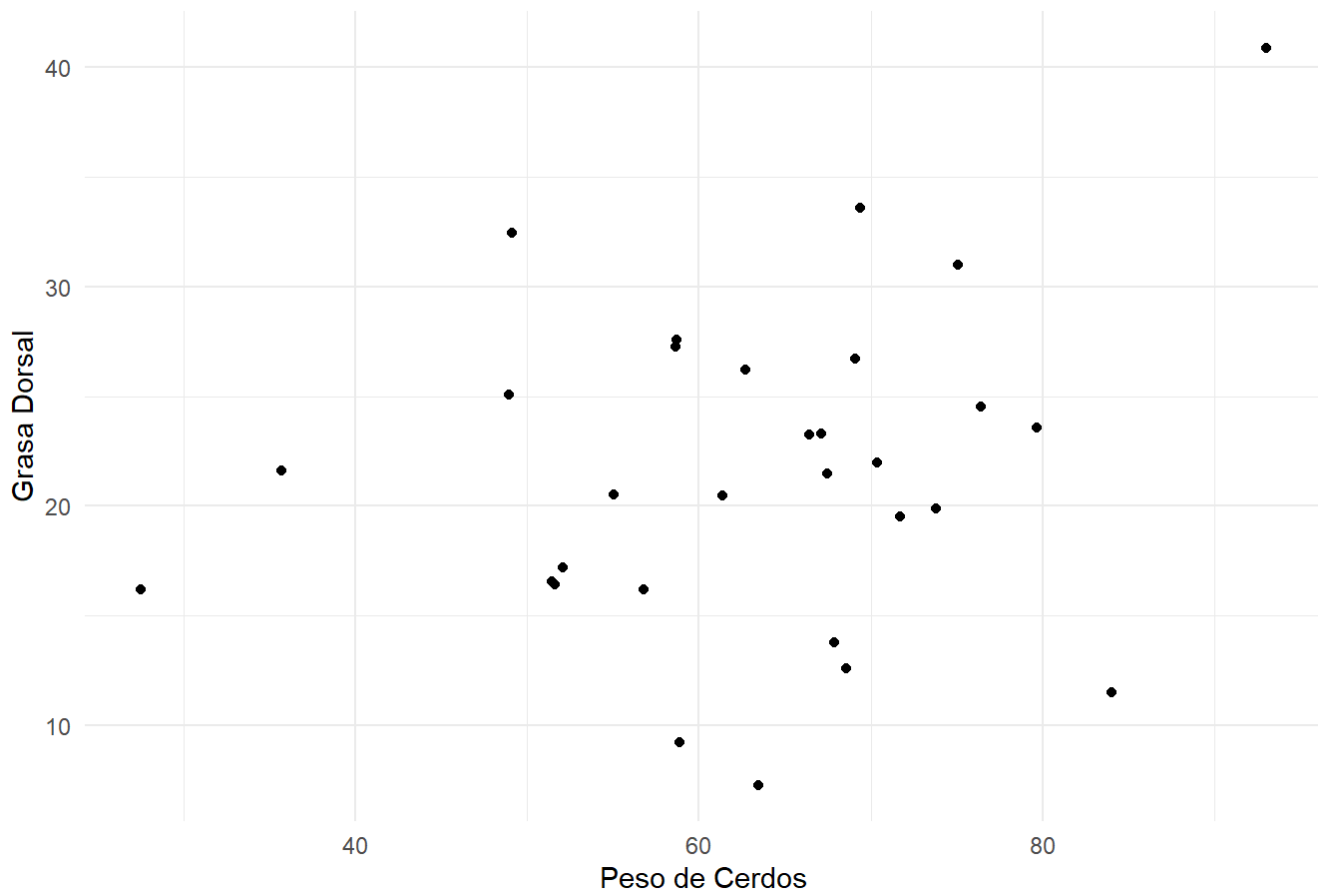
Hide

```
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 la
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 explíquelolo.

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
```

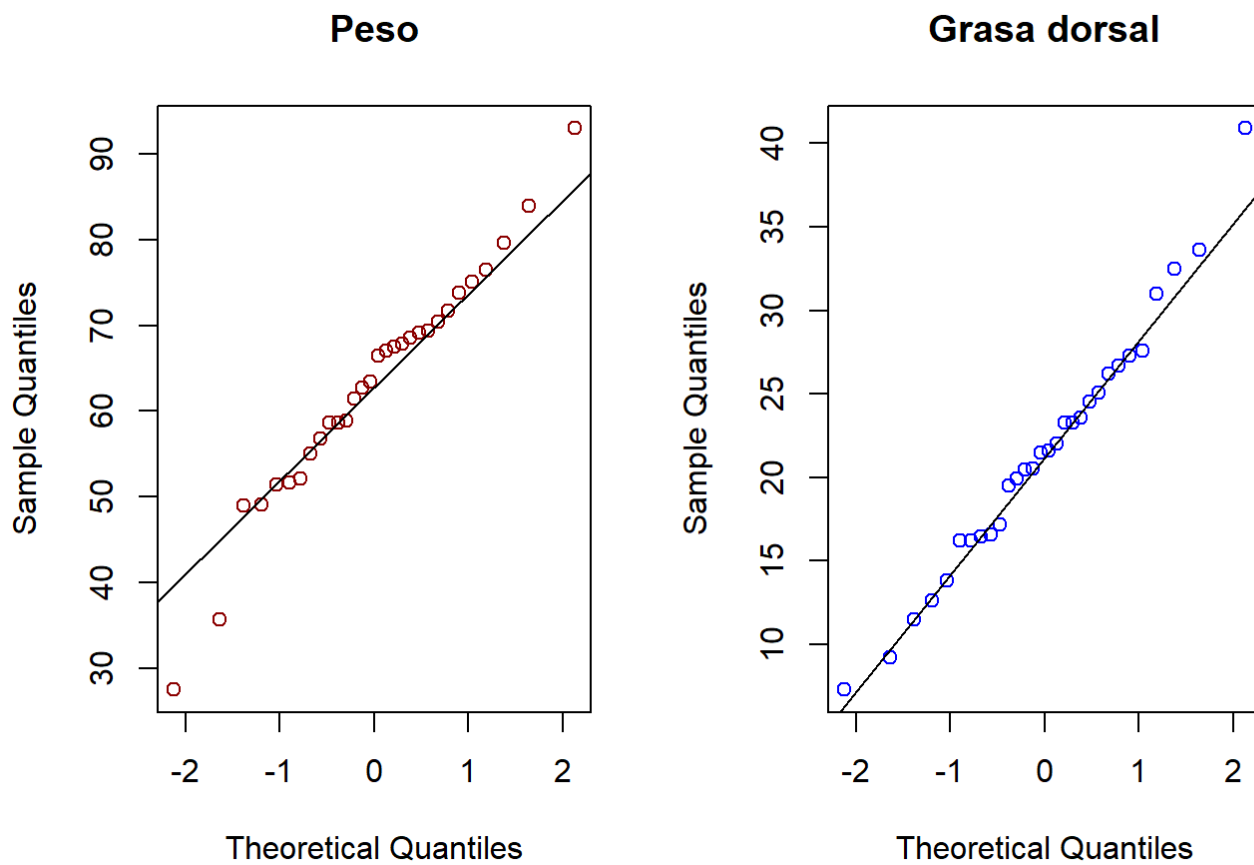
[Hide](#)

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

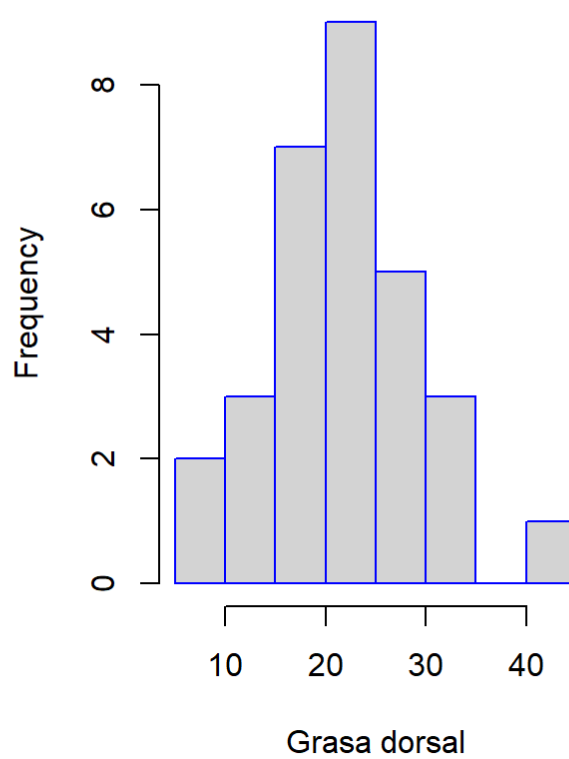
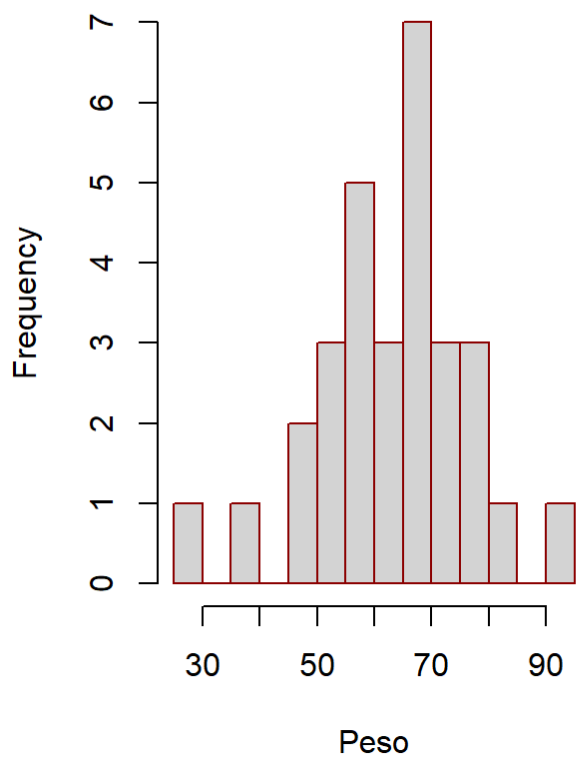
```
##  
##  Shapiro-Wilk normality test  
##  
## data:  grasacerdos$EGD  
## W = 0.98514, p-value = 0.9395
```

[Hide](#)

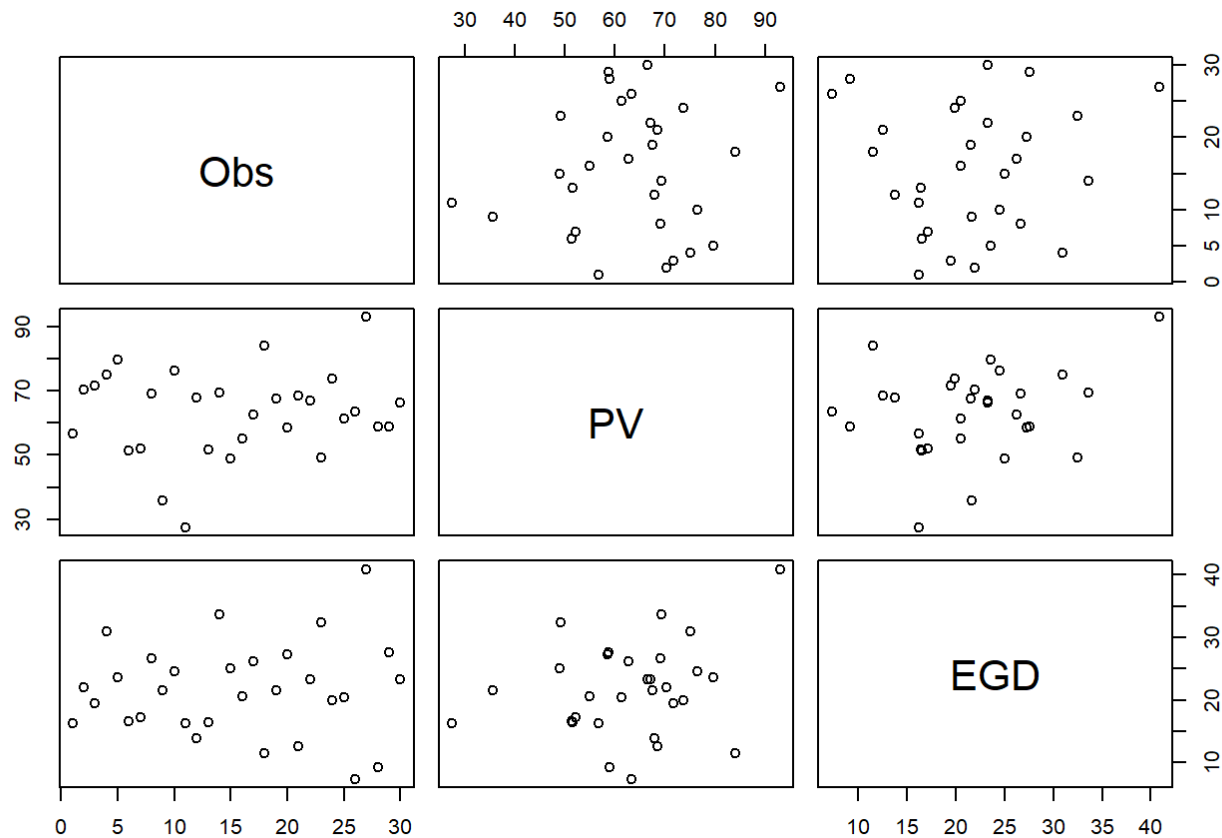
```
par(mfrow = c(1, 2))  
qqnorm(grasacerdos$PV, main = "Peso", col = "darkred")  
qqline(grasacerdos$PV)  
qqnorm(grasacerdos$EGD, main = "Grasa dorsal", col = "blue")  
qqline(grasacerdos$EGD)
```

[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 i
nternet y se trabaja con Anscombe's Quartet.xlsx (dimensiones 12x8)
anscombe<-read_excel("C:/Users/Josvaldes/Documents/Maestria/Austral/1ano/regresionAvanzad
a/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

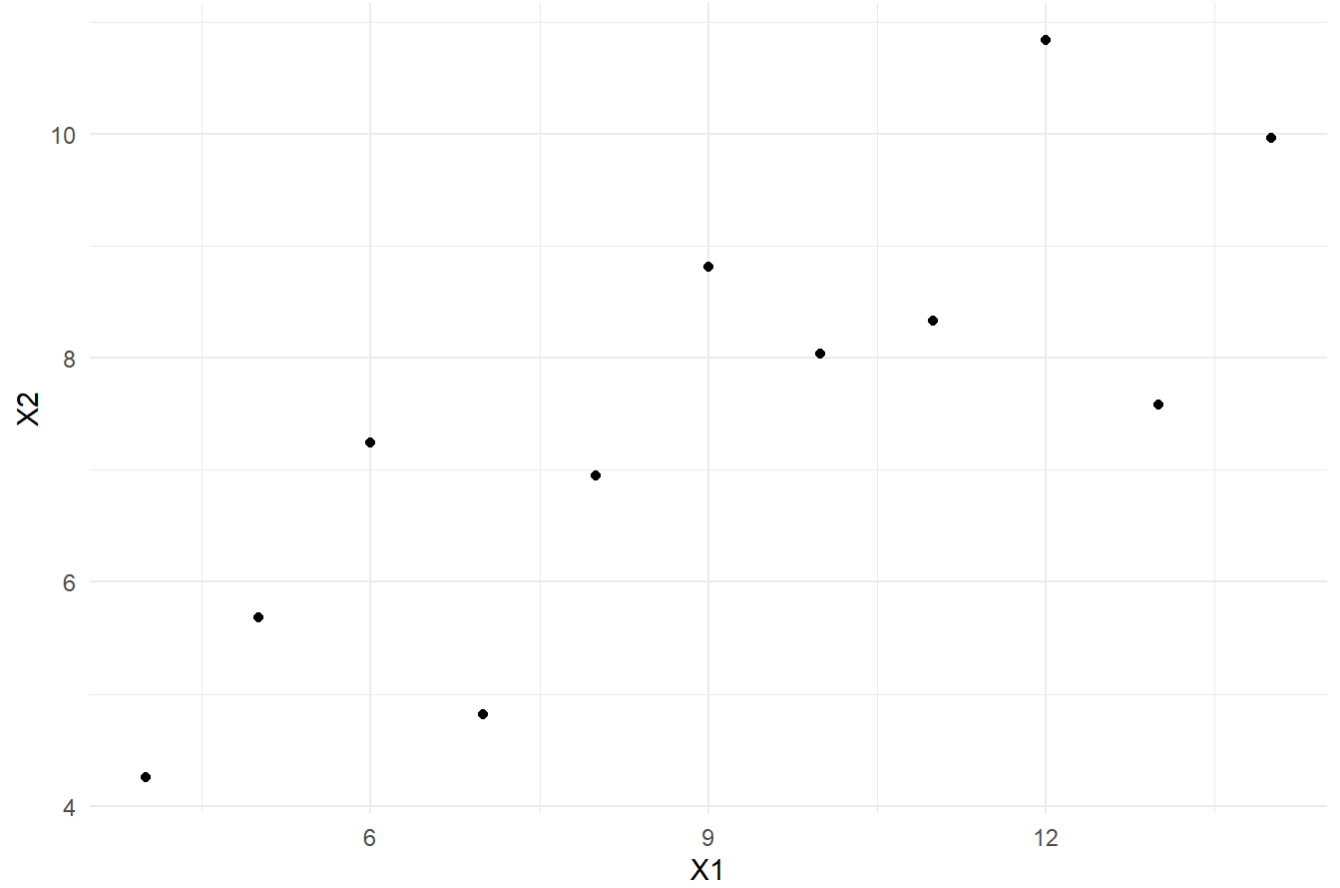
X1 <dbl>	X2 <dbl>	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <dbl>
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04

6 rows

Hide

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

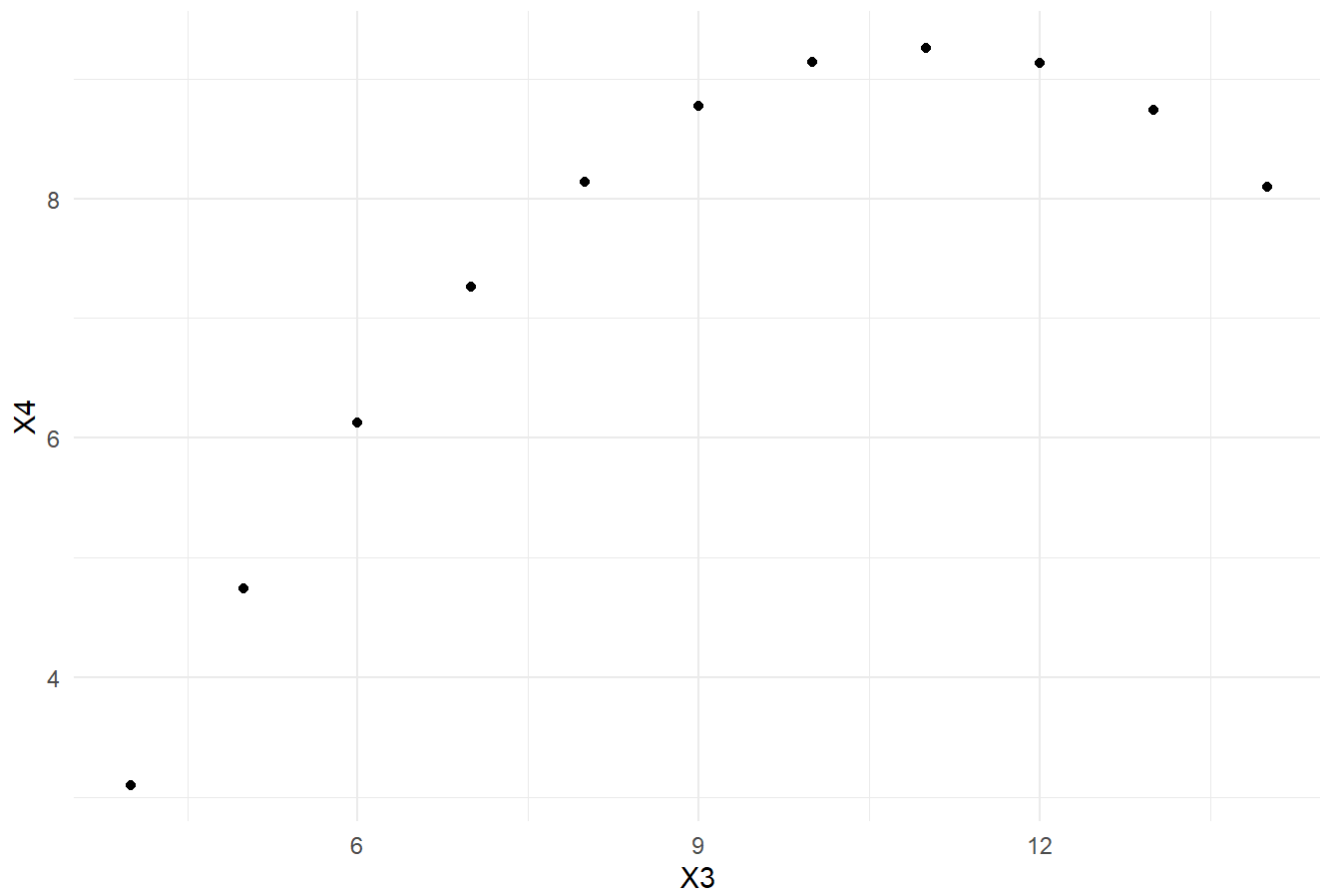
Diagrama de Dispersión X1 vs X2



Hide

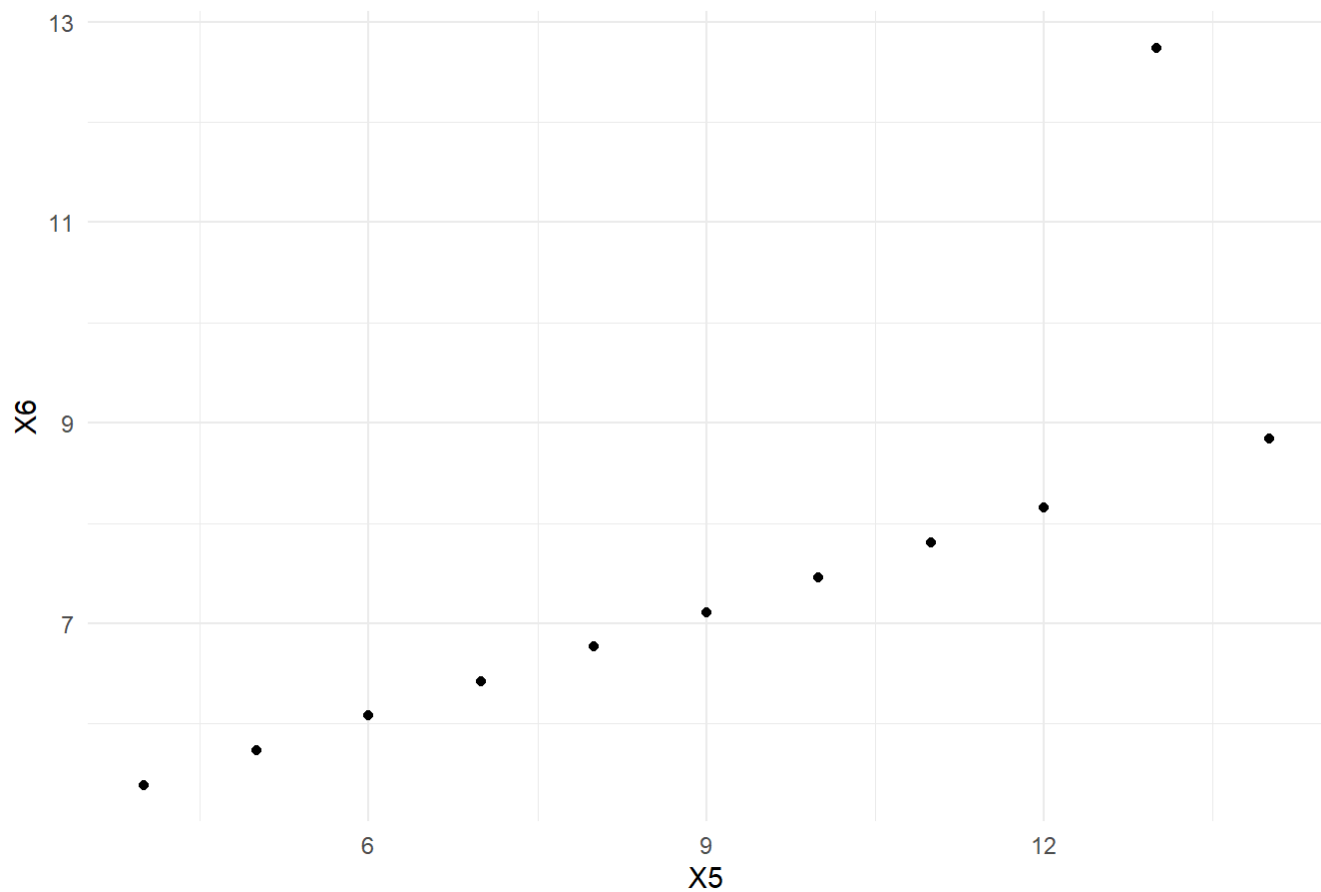
```
dd2=ggplot(anscombe, aes(X3, X4)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n X3 vs X4")  
dd2
```

## Diagrama de Dispersión X3 vs X4

[Hide](#)

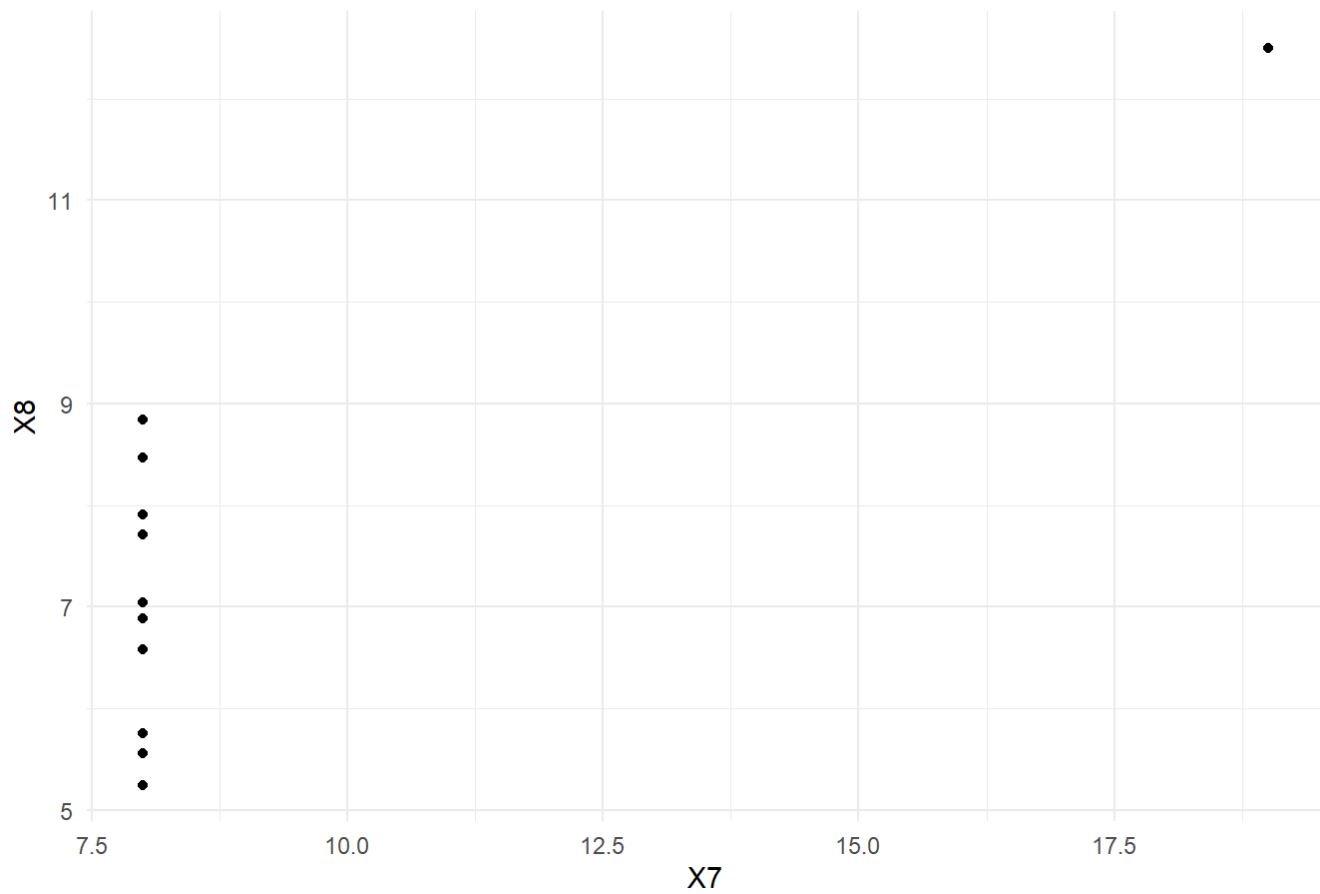
```
dd3=ggplot(anscombe, aes(X5, X6)) +  
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n X5 vs X6")  
dd3
```

## Diagrama de Dispersión X5 vs X6

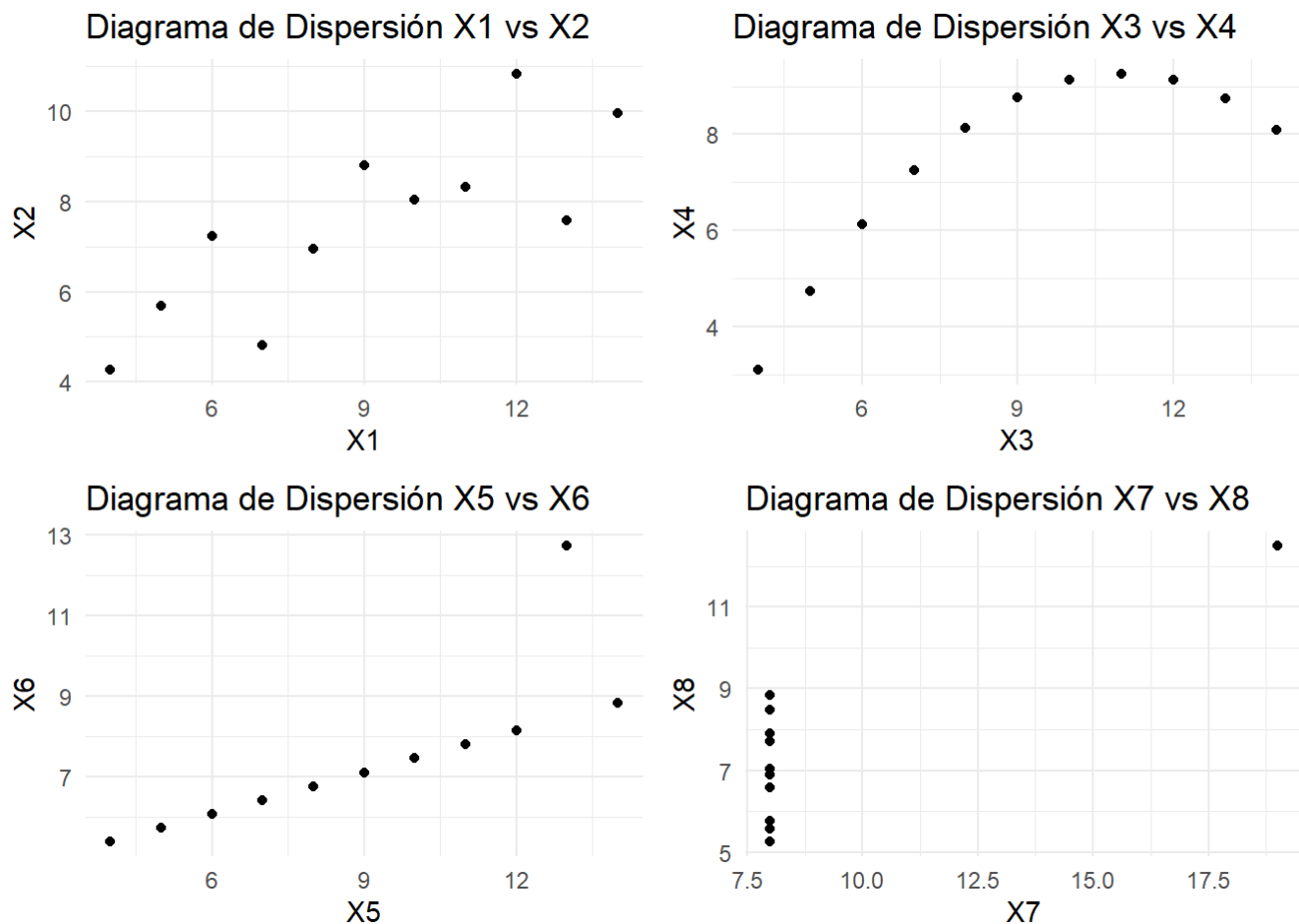
[Hide](#)

```
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)
```



(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

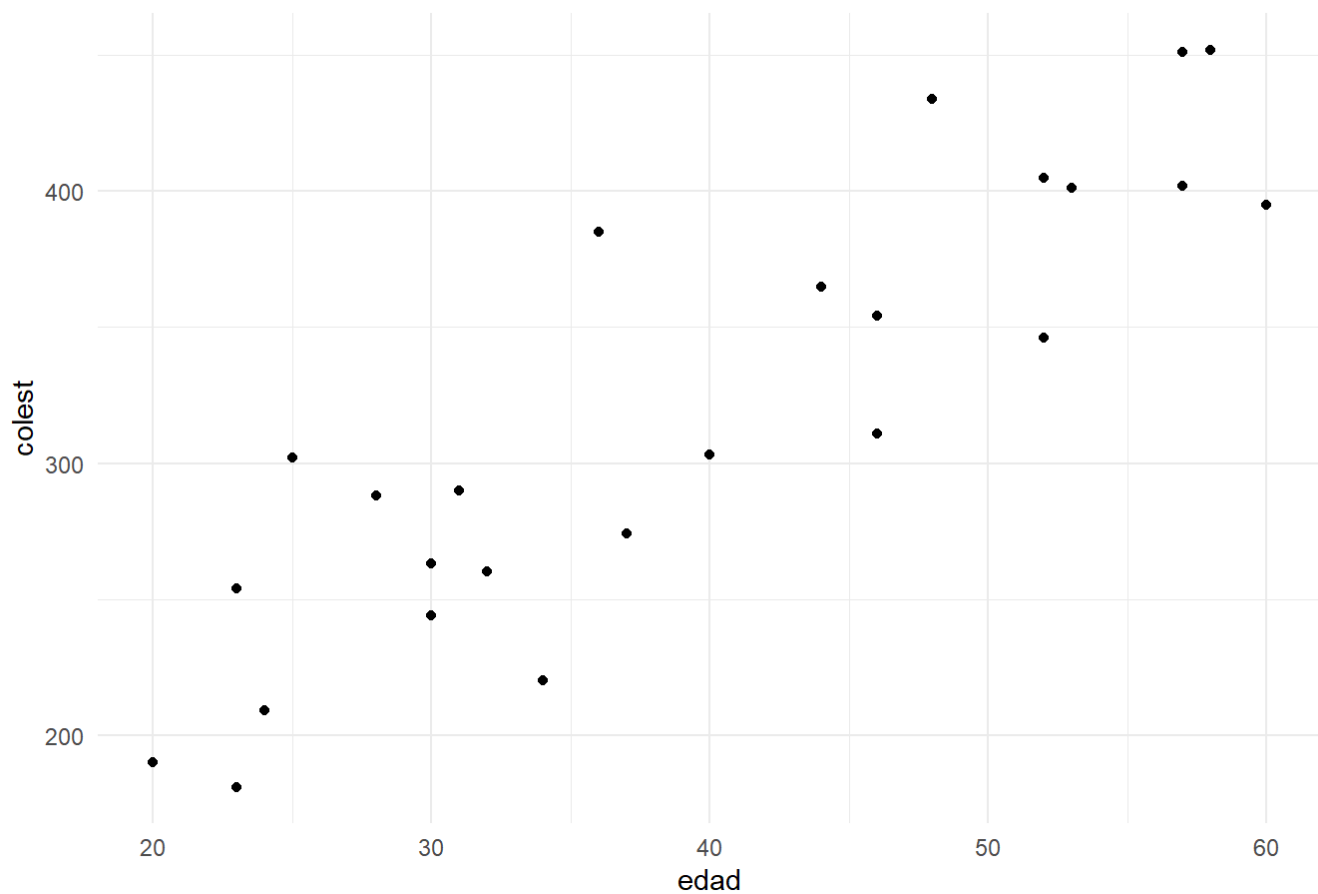
6 rows

Se realizan los diagramas de dispersión solicitados

[Hide](#)

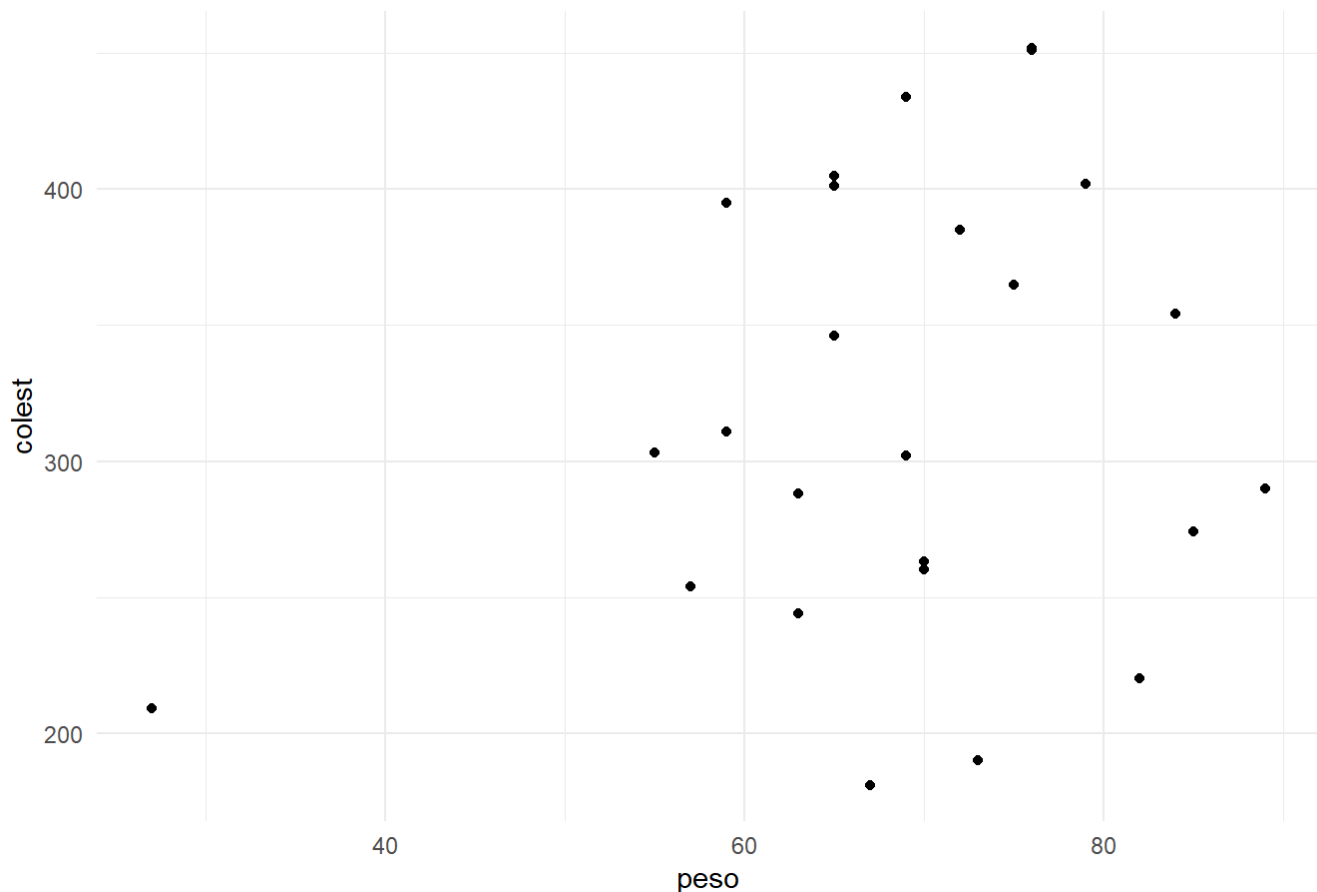
```
#Diagrama de dispersión colesterol en función de la edad
dd112=ggplot(colesterol, aes(edad, colest)) +
  geom_point() + theme_minimal() + labs(title = "Diagrama de Dispersi\u00F3n edad vs coles
terol")
dd112
```

## 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
```

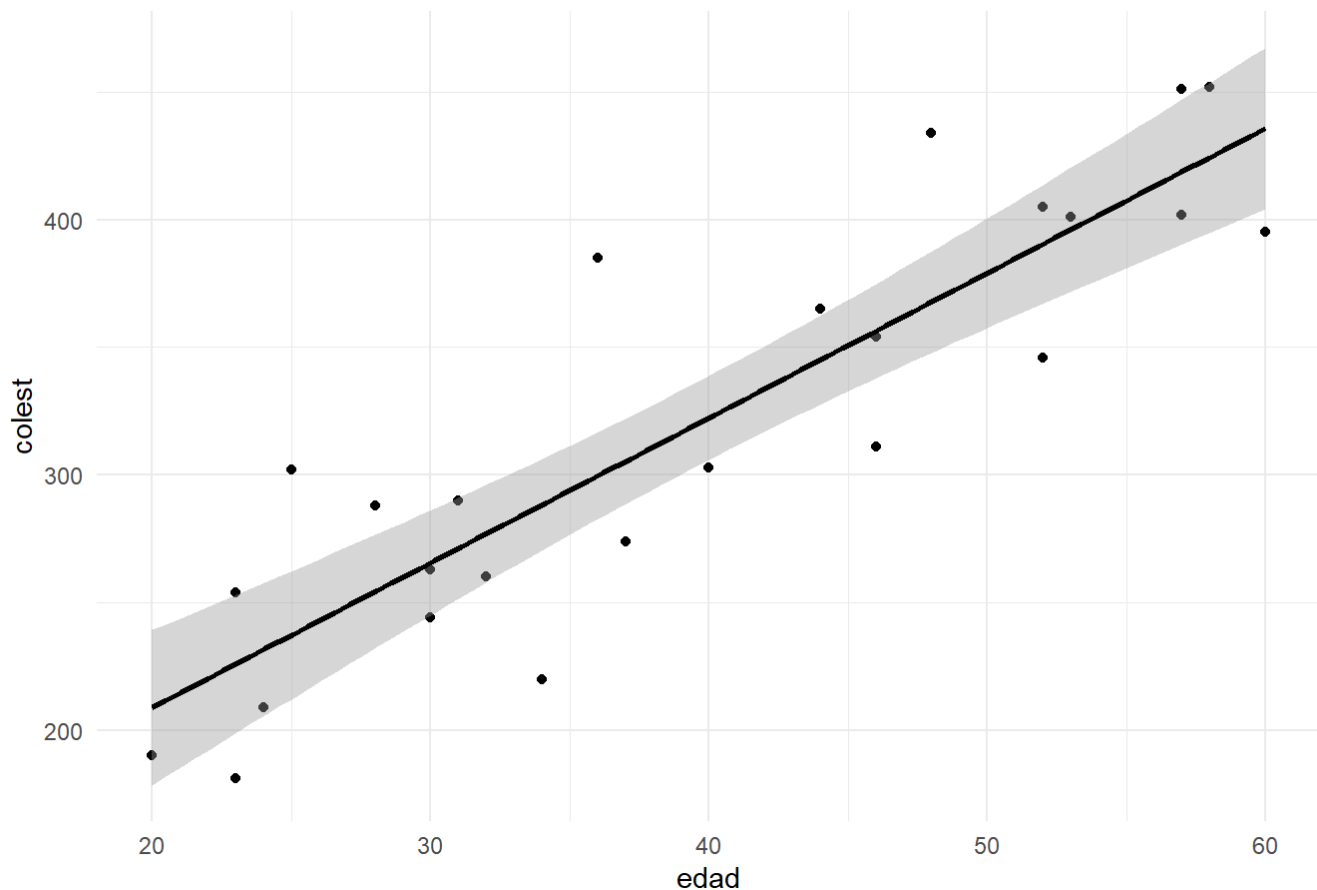
Grafica del modelo y las bandas de error estándar alrededor de la línea de regresión

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?

(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?

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