1. Responda las siguientes preguntas.

a) Suponga que se realiza escalamiento de longitud unitaria en las predictoras pero no en la variable respuesta, ¿qué unidades tienen los coeficientes de la regresión una vez esta es ajustada?
b) ¿Por qué hay problemas de multicolinealidad cuando se tienen más covariables que observaciones en los datos?
c) Si la traza de la matriz $\mathbf{X}'\mathbf{X}$ es muy grande, ¿mayor es la distancia entre el vector de parámetros estimados y el verdadero vector de parámetros?
d) Si la correlación entre las variables X_j y X_k es pequeña, ¿se puede descartar la presencia de multicolinealidad?
e) ¿Hay problemas de multicolinealidad en un modelo de 7 predictoras en el cual para β_3, $R_j = \sqrt{\frac{4}{5}}$? Recuerde que R_j^2 es el coeficiente de determinación muestral obtenido de una regresión de X_j (como respuesta) en función de las otras variables predictoras consideradas en el modelo (actuando como predictoras de X_j).

a) $Y = \beta_0 + \beta_1 X_1^* + \cdots + \beta_2 X_k^* + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_1^* + \cdots \hat{\beta_2} X_k^*$

$[Y] = [Y] \Rightarrow \hat{\beta_j} = \text{pendiente} = \left[\frac{\text{unidades } Y}{\text{unidades } X_j}\right]$ → adimensional

Así, las unidades de $\hat{\beta_j}$ son las mismas de $\hat{Y}$

b) $0 \leq \text{Rango} \leq \min(n,p)$

como $n < p$

$0 \leq \text{Rango} \leq n$

Ahora: $p \geq \underbrace{p - \text{Rango}}_{\text{1 2 ... p}} \geq p - n > 0$

$$\begin{matrix} & 1 & 2 & \cdots & p \\ 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ n & 0 & 0 & 0 & 0 \end{matrix} \rightarrow p-2 \text{ variables libres}$$

$\therefore$ p-Rango = # var's libres y como p - Rango > 0

si se considera Rango$(X^TX) > 0$ $\therefore \det(X^TX) = 0$

Además, $X^TX\beta = X^Ty$

para estimar β, se usa $\hat{\beta} = (X^TX)^{-1}X^Ty$

pero $\det(X^TX) = 0$, es singular! No es ¡nvertible!

El sistema no tiene solución o tiene infinitas.

c) Falso. $Tr[X^TX] = \sum_{i=1}^{n} \lambda_j$ luego todos los $\lambda_j > 0$ y X^TX es invertible

Ahora bien, si algún λ_j fuese 0

$$Tr\left[(X^TX)^{-1}\right] = \sum_{j=1}^{n} \frac{1}{\lambda_j}$$

No existe, computacional/ si se logra calcular pero esta trata es muy grande!!! Por lo que en este caso la afirmación es cierta para $(X^TX)^{-1}$, no X^TX

d) Puede suceder que $corr(x_1, x_3) = 0$, $corr(x_1, x_2) = 0$ o muy pequeñas, pero que estén relacionadas como

$$x_1 = aX_2 + bX_3, \quad a, b \in \mathbb{R}$$

luego es mejor comprobar por otro método.

e) Si $R_j = \sqrt{\frac{4}{5}}$, $R_j^2 = \frac{4}{5}$ y $VIF = C_{jj}^* = \frac{1}{1 - R_j^2} = \frac{1}{1 - \frac{4}{5}}$

$$= \frac{1}{\frac{1}{5}} = 5$$

∴ No hay problemas de multicolinealidad pues para ser al menos moderada $5 \leq VIF \leq 10$

2. Se genera un modelo de regresión lineal múltiple $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ con vector de parámetros $\boldsymbol{\Theta}' = (\beta_0 = -3, \beta_1 = 2, \beta_2 = -4, \sigma^2 = 4)$. Cree dos bases de datos usando las siguientes instrucciones.

```
gen_dat <- function(n) {
  x1 <- runif(n=n, min=0, max=10)
  x2 <- x1 * 2 + rnorm(n=n, sd=0.01) # x2 es el doble de x1 + ruido
  y <- rnorm(n=n, mean= - 3 + 2 * x1 - 4 * x2, sd=2)
  data.frame(y, x1, x2)
}
set.seed(12345)
datos <- gen_dat(n=40)
datos1 <- datos[1:20, ]
datos2 <- datos[21:40, ]
```

Luego de ajustar el modelo, obtenga los coeficientes estimados y comparelos con los reales, ¿qué sucede? Además, calcule los VIF y haga análisis del espectro de la matriz $\mathbf{X}'\mathbf{X}$.

```
> coef(mod1)
(Intercept)          x1          x2
  -1.559905  100.993906  -53.573619
> coef(mod2)
(Intercept)          x1          x2
  -2.401559  -36.607824   15.322999
```

Es claro que las estimaciones se inflan y muestran signos contrarios a los esperados, pudiendo dar sospechas de multicolinealidad.

Para análisis de VIF:

```
> car::vif(mod1)
      x1       x2
294997.4 294997.4
> car::vif(mod2)
      x1       x2
172810.7 172810.7
```

$VIF > 10$ hay problemas graves de colinealidad, ← Muy graves de hecho

Análisis de espectro de X^TX

Eigen_Value	Condition_Index	x1	x2
2.0000e+00	1.000	0.000001	0.000001
2.3099e-06	930.504	0.999999	0.999999

Val propios — Índices de condición — ϕ_{ij}

$\sqrt{\frac{\lambda_{max}}{\lambda_j}} = \sqrt{K}$

- # condición $\sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = 930,504 \geq 31,62$ ∴ Problemas graves!

- Valor propio pequeño y para este valor propio, ϕ_{ij} para x_1 y x_2 es $> 0,5$, por lo que existe colinealidad entre x_1 y x_2

3. Considere la base de datos **earthquake** del paquete **MPV**, seleccione el mejor modelo usando como criterios el MSE_p o equivalentemente R^2_{adj} y el C_p de Mallows al emplear el método de selección de todas las regresiones posibles.

	k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model
1	1	0.017	0.016	29042245	0.101	latitude
2	1	0.002	0.001	29475559	32.538	longitude
3	1	0.000	0.000	29527416	36.420	magnitude
4	2	0.017	0.016	29040930	2.002	latitude longitude
5	2	0.017	0.016	29042210	2.098	latitude magnitude
6	2	0.002	0.001	29472811	34.333	longitude magnitude
7	3	0.017	0.015	29040902	4.000	latitude longitude magnitude

$p-1$

R^2 : max = 0,017, modelo 1 por parsimonioso

R^2_{adj} : max = 0,016, modelo 1 por parsimonioso

C_p : $\min(C_p)$ y $\min(|C_p - p|)$, ojo, $p = K+1$

Se prefiere el mod 5 por parsimonia y